Tilman Schalmey

Computerlinguistische Datierung schriftsprachlicher chinesischer Texte

HASP
HEIDELBERG ASIAN STUDIES PUBLISHING

Computerlinguistische Datierung schriftsprachlicher chinesischer Texte

Tilman Schalmey

Computerlinguistische Datierung schriftsprachlicher chinesischer Texte

HASP
HEIDELBERG ASIAN STUDIES PUBLISHING

ORCID®
Tilman Schalmey https://orcid.org/0000-0002-1894-0696

Dissertation an der Fakultät für Sprach-, Literatur- und Medienwissenschaften der Universität Trier, eingereicht im Dezember 2021.

Die elektronische Erstveröffentlichung ist im CrossAsia Open Access Repository des FID Asien an der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz online zugänglich.
DOI: https://doi.org/10.48796/20221128-000

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://dnb.dnb.de abrufbar.

HASP
HEIDELBERG ASIAN STUDIES PUBLISHING

Publiziert bei Heidelberg Asian Studies Publishing (HASP), 2023

Universität Heidelberg / Universitätsbibliothek
Heidelberg Asian Studies Publishing (HASP),
Grabengasse 1, 69117 Heidelberg
https://hasp.ub.uni-heidelberg.de

Die elektronische Open-Access-Version dieses Buches ist auf der Webseite von Heidelberg Asian Studies Publishing dauerhaft frei verfügbar: https://hasp.ub.uni-heidelberg.de
URN: urn:nbn:de:16-hasp-1153-4
DOI: https://doi.org/10.11588/hasp.1153

Gesetzt vom Autor in LaTeX aus der Vollkorn von Friedrich Althausen und der *FangSong* 仿宋 von Changzhou SinoType 常州華文印刷新技術.

Titelfoto: »kunterbuntrunterrauf« (Treppenhaus in Frankfurt am Main).
https://flic.kr/p/2nv64gn. © Tilman Schalmey, 2022

Eine aktuelle Version der im Rahmen dieser Arbeit entstandenen Software zur Datierung schriftsprachlicher chinesischer Texte kann online unter https://visualtime.schalmey.de/ getestet werden.

Der Quellcode zu den in Kapitel 6 durchgeführten Experimenten ist unter https://github.com/dadiolli/dating-literary-chinese abrufbar.

ISBN 978-3-948791-67-4 (Hardcover)
ISBN 978-3-948791-66-7 (PDF)

Für Peter
1949–2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Vorbemerkungen

Chinesische Begriffe und Namen von Personen, Werken und Orten werden in dieser Arbeit in der *Hanyu Pinyin* 漢語拼音 Transkription wiedergegeben, gefolgt von traditionellen Schriftzeichen (*fantizi* 繁體字). Ausnahmen bilden Namen von Autor:innen, die selbst eine abweichende westliche Schreibweise gewählt haben, z. B. Huang Chu-ren [Juren] 黃居仁. Um eine gute Lesbarkeit der einzelnen Unterkapitel zu gewährleisten, werden die Schriftzeichen stets bei der ersten Nennung im jeweiligen Abschnitt mit angegeben. Dasselbe gilt für Zeitangaben zu chinesischen Dynastien. Als Referenz sind die wichtigsten Dynastien zusätzlich auf der folgenden Seite aufgeführt.
Wörtliche Zitate aus anderen Texten werden originalgetreu übernommen, auch wenn sie in Kurzzeichen (*jiantizi* 简体字) gesetzt sind, sowie andere Umschriften oder Rechtschreibfehler enthalten.
Familiennamen, sowie Namen von Organisationen werden in Kapitälchen gesetzt, *Transkriptionen* und *fremdsprachige Fachbegriffe kursiv*. `Quellcode`, sowie Namen von `Variablen`, `Funktionen` und `Datenbankspalten` werden mit `fester Zeichenbreite` gesetzt.

Dynastien

Shang 商	ca. 1600–1045 v. u. Z. [1]
Zhou 周	ca. 1045–256 v. u. Z.
Westliche Zhou (*Xi Zhou*) 西周	11. Jh.–771 v. u. Z.
Chunqiu 春秋	770–476 v. u. Z.
Streitende Reiche (*Zhanguo*) 戰國	475–221 v. u. Z.
Qin 秦	221–206 v. u. Z.
Han 漢	202 v. u. Z.–220 u. Z.
Frühere Han (*Qian Han*) 前漢	202 v. u. Z.–9 u. Z.
Xin 新	9–23
Östliche Han (*Dong Han*) 東漢	25–220
Drei Reiche (*Sanguo*) 三國	220–280
Wei 魏	220–265
Shu 蜀	221–263
Wu 吳	222–280
Jin 晉	265–420
Südliche und Nördliche Dynastien (*Nanbei chao*) 南北朝	420–589
Nördliche Wei (*Bei Wei*) 北魏	386–534
Nördliche Liang [16 Reiche] (*Bei Liang*) 北涼	401–439
Südliche Song (*Nanchao Song* / Liu Song) 南朝宋 / 劉宋	420–479
Südliche Qi (*Nan Qi*) 南齊	479–502
Südliche Liang (*Nanchao Liang*) 南朝梁	502–557
Östliche Wei (*Dong Wei*) 東魏	534–550
Nördliche Qi (*Bei Qi*) 北齊	550–577
Nördliche Zhou (*Bei Zhou*) 北周	557–581
Chen [Südliche Dynastien] (*Nanchao Chen*) 南朝陳	557–589
Sui 隋	581–618
Tang 唐	618–907
Fünf Dynastien (*Wudai*) 五代	907–960
Liao 遼	916–1125
Song 宋	960–1279
Jin 金	1115–1234
Yuan 元	1279–1368
Ming 明	1368–1644
Qing 清	1644–1912
Republik China (*Zhonghua Minguo*) 中華民國	1912–
VR China (*Zhonghua Renmin Gongheguo*) 中華人民共和國	1949–

1 Angaben aus Endymion Wilkinson 2000: *Chinese History. A Manual.* Revised and Enlarged. Cambridge, MA & London: Harvard University Asia Center, Harvard University Press, S. 10–12.

Abkürzungen

Seitenangaben zur Verwendung der gelisteten Fachbegriffe sind im Stichwortverzeichnis angegeben, die verwendeten Ausgaben der genannten Werke im Literaturverzeichnis.

ACL	*Association for Computational Linguistics*
API	*Application Programming Interface*
AYL	*Average Year of Lexicalization*
BERT	*Bidirectional Encoder Representations from Transformers*
BoW	*Bag of Words*
CBDB	*China Biographical Database*
CKIP	*Chinese Knowledge and Information Processing*
CS	*Cosine Similarity* (Kosinus-Ähnlichkeit)
CTB	*Chinese Treebank*
DDBC	*Dharma Drum Buddhist College Time Authority Database*
DFZ	*Difangzhi* 地方誌
DH	*Digital Humanities*
DHYDCD	Digitale Ausgabe des » ***HYDCD***
GB(K)	*Guojia Biaozhun* (*Kuozhan*) 國家標準 (擴展)
HHS	*Hou Han shu* 後漢書
HMM	Hidden-Markov-Modell
HYDCD	*Hanyu da cidian* 漢語大詞典
idf	*inverse document frequency*
KLD	Kullback-Leibler Divergenz
KPCh	Kommunistische Partei Chinas
LCM	*Largest chronon minimum*
MAE	*Mean average error*
MXBT	*Meng xi bi tan* 夢溪筆談
NER	*Named Entity Recognition*
NLLR	*Normalized Log-Likelihood-Ratio*
NLP	*Natural Language Processing*
OCR	*Optical Character Recognition*
OED	*Oxford English Dictionary*
PCA	*Principal Component Analysis*
PoS	*Part-of-speech*
RegEx	*Regular Expression* (regulärer Ausdruck)
RMRB	*Renmin ribao* 人民日報
SAYL	*Standardized Average Year of Lexicalization*
SKQS	*Si ku quan shu* 四庫全書
SLM	*Statistical Language Model*
SQL	*Stuctured Query Language*
TE	Temporale Entropie
TEI	*Text Encoding Initiative*
tf-idf	*term frequency inverse document frequency*
UTF-8	*8 Bit Unicode Transformation Format*
u. Z.	unserer Zeitrechnung
v. u. Z.	vor unserer Zeitrechnung
FWAYL	*Frequency Weighted* » ***AYL***
XML	*Extensible Markup Language*
XXSKQS	*Xu xiu si ku quan shu* 續修四庫全書

1 Einleitung

„Curiously enough, lexicalization has never been used by historical linguistics for the purpose of dating, although its study is extremely rewarding."[1]

Mario Alinei

Die Datierung von Texten hat in den vergangenen Jahrhunderten Forscher:innen unterschiedlicher Wissenschaften und Kulturkomplexe beschäftigt. Im Vordergrund stehen dabei oft die Exegese und Authentizitätsforschung. Die linguistische Textdatierung stützt sich in aller Regel auf die Beobachtung von und das Vorwissen über Sprachwandel, traditionell durch die Betrachtung bestimmter – einzelner – sprachlicher Phänomene, Wörter, Wortformen, Zeichen oder grammatikalischer Strukturen. Vor allem der Wortschatz jeder aktiv genutzten Sprache, sei es Schrift- oder Umgangssprache, befindet sich in permanentem Wandel. Neue Wörter kommen hinzu, andere fallen nach und nach aus dem Sprachgebrauch heraus.[2]

Von der Autorschaft unabhängige Textdatierung als Aufgabenfeld der Computerlinguistik ist ein sehr junges Forschungsgebiet, das durch einen Aufsatz von de Jong, Rode und Hiemstra (2005) ins Leben gerufen wurde.[3] Die darin vorgestellte Methodik nutzt *Bag of Words* (*BoW*)-Sprachmodelle und statistische Ähnlichkeitsmaße. Texte können so auf ihre Ähnlichkeit zu anderen Texten oder zu Teilen diachroner Vergleichskorpora geprüft und entsprechend zugeordnet werden. Ein wichtiges Ziel der Datierung ist dabei – im Unterschied zur traditionellen Forschung – die Einordnung von Dokumenten nach Relevanz, z. B. für die Sortierung der Ergebnisse von Suchmaschinen. Vergleichbare Ansätze wurden inzwischen für verschiedene europäische Sprachen mit großem Erfolg angewandt.[4] Für das Chinesische liegen bisher aber kaum vergleichbare Veröffentlichungen vor.[5]

1 Mario Alinei 2004: „The Problem of Dating in Linguistics". In: *Quaderni di semantica* 25.2, S. 211–232, S. 225.

2 *Zhi* 之 sei an dieser Stelle als für Sinolog:innen anschauliches Beispiel für solche diachronen Unterschiede genannt. In klassischen bzw. schriftsprachlichen Texten (s. u.) wird *zhi* u. a. als Subordinationspartikel eingesetzt. In einer vergleichbaren Funktion wird in der zeitgenössischen Umgangssprache *de* 的 verwendet.

3 Franciska M. G. de Jong, Henning Rode und Djoerd Hiemstra 2005: „Temporal Language Models for the Disclosure of Historical Text". In: *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*. Amsterdam: Koninklijke Nederlandse Academie van Wetenschappen, S. 161–168.

4 Siehe v. a. Kapitel 3.1, ab S. 40. Vgl. u. a. David Bamman et al. 2017: „Estimating the Date of First Publication in a Large-Scale Digital Library". In: *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, Toronto, Canada, June 2017 (JCDL '17)*, S. 1–10. doi: 10.475/1234; Filip Graliński et al. 2017: „The RetroC Challenge: How to Guess the Publication Year of a Text?" In: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. DATeCH2017. Göttingen: ACM, S. 29–34. doi: 10.1145/3078081.3078095.

5 Ausnahmen bilden ein Aufsatz von Yamada Takahito 山田崇仁, dessen Methodik eher der Stilometrie zugerechnet werden kann. Siehe Yamada Takahito 山田崇仁 2004: „N-gram moderu o riyōshite senshin bunken no seisho jiki o saguru: ‚Sonshi' jūsanhen o jirei toshite N-gram モデルを利用して先秦文献の成書時期を探る —『孫子』十三篇を事例として —, *n*-Gramm Ansatz zur Datierung von chinesischen-Qin Texten am Beispiel der 13 *Sunzi*-Kapitel)". In: *Tōkyōdaigaku tōyō bunka kenkyūjo fuzoku tōyō-gaku kenkyū jōhō sentā, Ajia kenkyū jōhō* 東京大学東洋文化研究所附属東洋学研究情報センター，アジア研究情報 *(Tokyo University Research & Information Center for Asian Studies, Gateway to Asian Studies in Japan)*, siehe auch Kapitel 3.1, S. 40, sowie ein kurzer Aufsatz von Yu Xuejin und Wei Huangfu 2019:

Auf dem Gebiet der Stilometrie wurden in jüngster Zeit auch für schriftsprachliche chinesische Texte[6] spannende Ergebnisse präsentiert.[7] Darin deutet sich an, dass chinesischsprachige Textstile über einen Zeitraum von vielen Jahrhunderten eine hohe Rigidität aufweisen können, so dass Unterschiede zwischen Gattungen bzw. Sprach*stilen* mit statistischen Methoden viel klarer messbar bzw. differenzierbar sind als eine temporale Dimension.[8]

Darüber hinaus liefern orthographische Veränderungen in Sprachen mit alphabetischen Schriftsystemen – sei es bedingt durch tatsächliche phonologische Veränderungen, oder durch Rechtschreibreformen bzw. -konventionen – wertvolle Hinweise zur chronologischen Einordnung von Texten.[9] Die chinesische Schrift(sprache) hat sich diesbezüglich jedoch über einen langen Zeitraum hinweg verhältnismäßig wenig verändert.[10] Andererseits ermöglicht die fast lückenlose Texttradition, dank der uns Textzeugnisse aus dem Zeitraum von ca. 1000 v. u. Z. bis ins 21. Jh. zur Verfügung stehen, im Falle des Chinesischen eine nahezu unvergleichliche zeitliche Tiefe in diachronen sprachwissenschaftlichen Untersuchungen.

Zentrale Fragestellungen und Ziele

Ziel dieser Arbeit ist es, die inhaltsbasierte zeitliche Einordnung bzw. Datierung schriftsprachlicher chinesischer Texte mit computerlinguistischen Methoden zu ermöglichen. Die erwähnten statistischen Sprachmodelle sollen zu diesem Zweck erstmals für chinesisches Textmaterial adaptiert und angewandt werden.[11] Einschränkungen ergeben sich dabei aus der stilistischen Rigidität einiger schriftsprachlicher Textgattungen und der Beständigkeit des chinesischen Schriftsystems. Die Nutzung statistischer Sprachmodelle erfordert überdies ein diachrones Trainingskorpus, das den gesamten Zeitraum abdeckt, aus dem Texte datiert werden sollen.

Da ein geringer syntaktischer und ein eher wenig in der Schrift manifestierter phonologischer Wandel die Möglichkeiten statistischer Methoden begrenzen, soll hier der lexikalische Wandel stärker in den Fokus rücken. Bereits de Jong, Rode und Hiemstra „foresee a role for parsed entries from historical dictionaries in this context […]",[12] ohne diesen Ansatz

„A Machine Learning Model for the Dating of Ancient Chinese Texts". In: *International Conference on Asian Language Processing, IALP 2019, Shanghai, China, November 15-17, 2019*. Hrsg. von Lan Man et al. IEEE, S. 115–120. doi: 10.1109/IALP48816.2019.9037653; Auch Ryan Nichols, Edward Slingerland und Kristoffer Nielbo scheinen sich im Rahmen ihrer Beschäftigung mit *machine learning* und *topic modelling* indirekt mit temporaler Klassifizierung antiker chinesischer Texte zu befassen, legen aber bislang keine Veröffentlichung dazu vor. Siehe Ryan Nichols et al. 2018: „Modeling the Contested Relationship between *Analects*, *Mencius*, and *Xunzi*: Preliminary Evidence from a Machine-Learning Approach". In: *Journal of Asian Studies* 77.1, S. 19–57, S. 23, siehe auch Kapitel 4.1, ab S. 60.

6 Zum Begriff *schriftsprachliches Chinesisch* siehe S. 5.

7 Siehe dazu z. B. die Arbeit von Paul Vierthaler, der unterschiedliche Gattungen von Geschichtstexten untersucht. Siehe Paul Vierthaler 2016a: „Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature". In: *Journal of Cultural Analytics*, S. 1–32. doi: 10.22148/16.003, ausführlicher dazu siehe Kapitel 2.3, ab S. 20 und Kapitel 4.1, ab S. 60.

8 Vgl. dazu auch Tilman Schalmey 2021: „Thoughts on »Reliable« Learner's Vocabularies for Classical and Literary Chinese". In: *Teaching Classical Chinese/Zum Unterricht des Klassischen Chinesischen/Wenyan wen jiaoxue 文言文教学 (Proceedings of the International Symposium on the Teaching of Classical Chinese, December 14–16, 2018)*. Hrsg. von Li Wen 李文 und Ralph Kauz. Gossenberg: Ostasien Verlag, S. 251–261, S. 254–255.

9 Siehe Anne Garcia-Fernandez et al. 2011: „When Was It Written? Automatically Determining Publication Dates". In: *String Processing and Information Retrieval*. Hrsg. von Roberto Grossi, Fabrizio Sebastiani und Fabrizio Silvestri. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 221–236, S. 7; siehe auch Graliński et al. 2017, S. 32.

10 Vgl. z. B. Chou Ya-Min 周亞民 und Huang Chu-ren 黃居仁 2010: „Hantology: conceptual system discovery based on orthographic convention". In: *Ontology and the Lexicon*. Hrsg. von Huang Chu-ren 黃居仁 et al. Studies in Natural Language Processing. Cambridge & New York: Cambridge University Press, S. 122–143, S. 133. Ausführlicher dazu siehe Kapitel 2.2, ab S. 16.

11 Siehe v. a. Kapitel 6.1, ab S. 156.

12 de Jong, Rode und Hiemstra 2005, S. 2.

aber weiter zu verfolgen. Die Idee, Informationen über das Entstehen und Verschwinden von Wörtern aus Wörterbüchern zur Textdatierung zu nutzen, findet sich erneut in der Arbeit von Garcia-Fernandez et al. (2011) für die Datierung französischsprachiger Texte.[13] Dennoch ist die computerlinguistische Nutzung lexikographischer Quellen zur Textdatierung bisher ausgeblieben. Sie beklagen:

> […] there is no pre-compiled list of words with their year of appearance or disappearance. This type of information is sometimes included in dictionaries, but depends on the availability of these resources.[14]

Für das Chinesische steht mit dem *Hanyu da cidian* 漢語大詞典 *(Großes Wörterbuch der chinesischen Sprache)*[15] (*HYDCD*) jedoch eine sehr umfangreiche Quelle digital zur Verfügung,[16] aus der zumindest Informationen über das Erscheinen von Wörtern bzw. Lexemen extrahiert werden können. Zu diesem Zweck wird die *Plain Text* Fassung des *HYDCD* in eine relationale Datenbank umgeformt, in der Angaben über die frühesten Belege von Lexemen abrufbar sind. Auf dieser Grundlage entwickle ich alternative lexikographische Methoden zur Schätzung der Entstehungszeit schriftsprachlicher Texte.[17] Es ergeben sich zwei zentrale Fragestellungen:

1. Lassen sich die erwähnten statistischen Methoden erfolgreich für die Verwendung mit chinesischem Textmaterial adaptieren?
2. Können die beschriebenen Limitationen mit lexikographischen Vorgehensweisen reduziert werden?

Vorbereitend werden grundsätzliche Fragen zu geeigneten Ressourcen und dem computerlinguistischen Umgang mit schriftsprachlichem Chinesisch erörtert. Welche diachronen Trainings- und Testkorpora sind geeignet? Ist eine verlässliche Segmentierung schriftsprachlicher Texte möglich? Wie eingangs erwähnt besteht überdies eine enge Verbindung zwischen linguistischer Datierung und Sprachwandel. Sekundär sollen daher anhand des verwendeten Materials Beobachtungen zum Sprachwandel, insbesondere dem lexikalischen Wandel des Chinesischen, ermöglicht und diskutiert werden.

Die Verwendung der untersuchten Methoden bleibt auf digitales *Plain Text*-Material beschränkt. Sie eignen sich nicht zur Datierung gescannter Drucke oder gar von Handschriften, die nur als Bilddaten vorliegen, auch wenn darin sichtbare Merkmale von Originaldokumenten wie Materialität, Schriftstile, Drucktypen oder Zeichenvarianten wichtige Aspekte der Textdatierung darstellen können.[18]

13 Siehe Garcia-Fernandez et al. 2011, S. 5.

14 Ebd.

15 Luo Zhufeng 羅竹風, Hrsg. 1986–1994: *Hanyu da cidian* 漢語大詞典 *(Großes Wörterbuch der chinesischen Sprache)*. Bd. 1-13. Shanghai 上海: Cishu chubanshe 辭書出版社 (im Folgenden zit. als *HYDCD*).

16 Luo Zhufeng 羅竹風, Hrsg. 2005: *Hanyu da cidian* 漢語大詞典 *UTF-8 (Großes Wörterbuch der chinesischen Sprache, Unicode-Version)*. Shanghai 上海. URL: http://bbs.gxsd.com.cn/forum.php?mod=viewthread&tid=498015 (besucht am 13. 01. 2013) (im Folgenden zit. als *DHYDCD*).

17 Siehe v. a. Kapitel 6.2, ab S. 179 und 6.3, ab S. 210.

18 Neuere archäologische Befunde sprechen zudem dafür, dass die typischerweise linear dargestellte Entwicklung der chinesischen Schrift von Orakel-, Bronze, Siegel- und Kanzleischrift bis hin zu einer standardisierten Schrift, tatsächlich weniger linear verlief als gemeinhin angenommen. Siehe dazu Imre Galambos 2006: *Orthography of Early Chinese Writing: Evidence from Newly Excavated Manuscripts*. Budapest: Department of East Asian Studies, Eötvs Loránd University.

Digital Humanities

Die vorliegende Arbeit verbindet Sinologie und Computerlinguistik und kann übergreifend den *Digital Humanities* (*DH*, *shuzi renwen* 數字人文) zugerechnet werden.[19] Damit sie für computerlinguistisch nicht vorgebildete Leser:innen verständlich und nachvollziehbar bleibt, werden an geeigneter Stelle Konzepte und Begriffe aus der Computerlinguistik erklärt, bzw. auf die weiterführende Fachliteratur verwiesen.

Die Verbindung von Geisteswissenschaft und Informatik ist keineswegs erst ein Trend der vergangenen Jahre. Für den Versuch, die Stücke William Shakespeares chronologisch zu ordnen, wurden bereits Ende des 19. Jhs. – noch ohne den Einsatz von Computern – statistische Methoden für die Beantwortung geisteswissenschaftlicher Fragestellungen eingesetzt.[20] 1949 gelang es Roberto Busa, Thomas Watson von IBM zu überzeugen, eine elektronische Konkordanz der Texte Thomas von Aquins für dessen Forschung zu erstellen.[21] Anhand dieser untersuchte Busa erfolgreich „nicht bewusste Spracheigentümlichkeiten" mit wahrscheinlichkeitstheoretischen Ansätzen.[22] Eine 1964 veröffentlichte Arbeit über die Verfasserschaft der *Federalist Papers* stützte sich ebenfalls auf die Häufigkeit bestimmter Funktionswörter.[23] Solche stilometrischen Analysen der Autorschaft von Texten gehören also zu den frühesten Forschungsbereichen der *DH*.[24]

Toolstack

Für die Programmierung der im Rahmen dieser Arbeit entwickelten Software kommt hauptsächlich *Python 3* zum Einsatz.[25] *Python* gehört „zu den beliebtesten Programmiersprachen weltweit"[26] und „hat sich in den letzten Jahrzehnten zu einem erstklassigen Tool für wissenschaftliche Berech-

19 Die *DH* haben sich als „neues Arbeitsfeld etabliert, das an der Schnittstelle zwischen den Geisteswissenschaften und der Informatik angesiedelt ist." Fotis Jannidis, Hubertus Kohle und Malte Rehbein, Hrsg. 2017: *Digital Humanities – Eine Einführung*. Stuttgart: Metzler, S. XI. Unter dem Begriff *DH* vereint werden sowohl Arbeiten, die vorhandene Tools wie *Voyant* einsetzen, um einer geisteswissenschaftlichen Fragestellung nachzugehen Stéfan Sinclair und Geoffrey Rockwell 2016: *Voyant Tools*. URL: `https://voyant-tools.org/` (besucht am 13. 04. 2023), (Bsp. unter `https://voyant-tools.org/docs/#!/guide/gallery`); als auch Arbeiten, die neue Werkzeuge zur Beantwortung geisteswissenschaftlicher Fragestellungen entwickeln oder verbessern. Vgl. z. B. Adam Kilgarriff et al. 2004: „The Sketch Engine". In: *Proceedings of the 11th EURALEX International Congress*. Hrsg. von Geoffrey Williams und Sandra Vessier. Lorient, France: Université des lettres et des sciences humaines, S. 105–115.

20 Siehe Andrew Murphy 2003: *Shakespeare in Print: A History and Chronology of Shakespeare Publishing*. Cambridge: Cambridge University Press, S. 209–210. Siehe dazu auch Kapitel 3.1, S. 41.

21 Siehe Benjamin Mangrum 2018: „Aggregation, Public Criticism, and the History of Reading Big Data". In: *PMLA* 133.5, S. 1207–1224. DOI: 10.1632/pmla.2018.133.5.1207, 1207. Obwohl Busa zunächst nicht selbst programmierte, wird er als einer der Gründungsväter der Computerlinguistik gefeiert.

22 Siehe auch Manfred Thaller 2017: „Geschichte der Digital Humanities". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 3–12, S. 3–4.

23 Siehe Frederick Mosteller und David L. Wallace 1984 [1964]: *Applied Bayesian and Classical Inference – The Case of The Federalist Papers*. 2. Aufl. New York: Springer, z. B. S. 10–14.

24 Weitere Beispiele, sowie ein Abriss über unterschiedliche Schwerpunkte und Trends der vergangenen 50 Jahre findet sich z. B. in Thaller 2017, einige aktuellere Arbeiten werden in Kapitel 3.1 und mit Schwerpunkt auf das Chinesische in 4.1 vorgestellt.

25 Den entscheidenden Impuls für die Wahl von *Python* gaben ein Workshop sowie die Arbeit von Paul Vierthaler, siehe auch Kapitel 4.1, ab S. 60. Das vorliegende Projekt wurde anfangs in *Python 2.7* entwickelt, jedoch wegen der deutlich verbesserten nativen Unicode-Unterstützung zur Version 3.8 gewechselt. Das vorher umständliche En- und Decodieren der Zeichenrepräsentanzen (siehe auch Kapitel 4.3, S. 69) entfällt seit Version 3 völlig – bzw. geschieht automatisch. Vgl. auch Paul Vierthaler 2020: „Digital humanities and East Asian studies in 2020". In: *History Compass* e12628. DOI: 10.1111/hic3.12628, S. 11 (EN 7).

26 Fotis Jannidis 2017a: „Grundbegriffe des Programmierens". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 68–95, S. 68.

nungen entwickelt, insbesondere auch für die Analyse und Visualisierung großer Datensätze“,[27] z. B. mit der Bibliothek *pandas*[28], sowie zur Lösung von Problemstellungen der quantitativen Linguistik.[29] Für die Verwendung von *Python* zur Bearbeitung von Fragestellungen der historischen Linguistik des Chinesischen spricht zudem die Verfügbarkeit einer wachsenden Anzahl spezialisierter Bibliotheken.[30]

Zur Berechnung und Bewertung statistischer Zusammenhänge wird *R* verwendet, eine Programmiersprache, die sich im Bereich der Statistik als wissenschaftlicher Standard etabliert hat. Die *R*-Programmbibliothek *ggplot2* wird zudem zur Visualisierung von Daten und Ergebnissen eingesetzt.[31]

Schriftsprachliches Chinesisch

Da die zeitliche Einordnung von Texten aus einem Zeitraum von den Anfängen des klassischen Schrifttums (erstes Jahrtausend v. u. Z.) bis ins 20. Jh. Gegenstand dieser Arbeit ist, wird – in Anlehnung an die ebenfalls weit gefassten Begriffe *wenyan (wen)* 文言 (文) bzw. *Literary Chinese* – hier der Begriff *schriftsprachliches Chinesisch* bzw. *Chinesische Schriftsprache* als Sammelbegriff für alle vormodernen Entwicklungsstufen des Chinesischen verwendet. Dies geschieht in Abgrenzung zur modernen Hoch- bzw. Umgangssprache (*putonghua* 普通話 und *guoyo* 國語 bzw. *kouyu* 口語). Häufig werden – auch in der sinologischen Literatur – die Bezeichnungen *Klassisches Chinesisch* bzw. *Classical Chinese* und *Literary Chinese* synonym verwendet, insbesondere im anglo-amerikanischen und chinesischen Sprachgebrauch. Im Deutschen bezeichnet *Klassisches Chinesisch* im Wesentlichen aber das antike Schrifttum – die aus Textzeugnissen etwa der zweiten Hälfte des ersten Jahrtausends v. u. Z. überlieferte Schriftsprache.[32] Im Englischen wird der Begriff auch von Sprachwissenschaftler:innen etwas weiter gefasst: „Classical Chinese is a conventional way of

27 Jake Vanderplas 2018: *Data Science mit Python – Das Handbuch für den Einsatz von IPython, Jupyter, NumPy, Pandas, Matplotlib, Scikit-Learn*. Übers. von Knut Lorenzen. 1. Aufl., S. 14.

28 Als Bibliotheken bzw. *libraries* werden fertige Programmpakete oder Funktionen bereitgestellt, die beliebig innerhalb eigener Programme eingesetzt werden können. *pandas* etwa vereinfacht die Analyse und Manipulation von Daten in tabellenähnlichen Objekten. Siehe Wes McKinney 2010: „Data Structures for Statistical Computing in Python“. In: *Proceedings of the 9th Python in Science Conference*. Hrsg. von Stéfan van der Walt und Jarrod Millman, S. 56–61. doi: 10.25080/Majora-92bf1922-00a; The pandas development team 2020: *pandas 1.5.0*. doi: 10.5281/zenodo.3509134.

29 Ralf Jüngling und Gabriel Altmann 2003: „Python for linguistics?“ In: *Glottometrics* 6, S. 70–82.

30 Einige wichtige Beispiele sind der Tokenizer *Jieba*, siehe Sun Junyi 2018: *Jieba zhongwen fenci* 结巴中文分词 *(Jieba Chinesisch-Tokenizer)*. GitHub Repository. url: https://github.com/fxsjy/jieba (besucht am 26. 02. 2019), ausführlicher siehe Kapitel 4.5, ab S. 83; *sinopy*, das Funktionen zur Konvertierung von Zeichen in versch. Umschriften bereitstellt, wobei nicht nur *Hanyu Pinyin* 漢語拼音, sondern auch die historisierende Umschrift für das mittelchinesische von Baxter, sowie das *International Phonetic Alphabet* (*IPA*) zur Verfügung stehen. Siehe Johann-Mattis List 2018: *SinoPy: Python Library for quantitative tasks in Chinese historical linguistics*. Jena. url: https://pypi.org/project/sinopy/ (besucht am 26. 04. 2020); sowie *mafan*, das Anwender:innen *mafan* 麻煩 („Unannehmlichkeiten“) ersparen soll. Die Bibliothek enthält Tools zur Umwandlung zwischen und Erkennung von traditionellen und vereinfachten Zeichen, sowie zum Umgang mit Codierungen. Siehe Herman Schaaf 2017: *Mafan - Toolkit for working with Chinese in Python*. Python module. url: https://pypi.org/project/mafan/ (besucht am 26. 04. 2020).

31 Hadley Wickham 2016: *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. url: https://ggplot2.tidyverse.org.

32 Siehe z. B. Ulrich Unger 1985a: *Einführung in das Klassische Chinesisch, Teil I: Allgemeines, Chinesische Texte, Indices*. Wiesbaden: Harrassowitz, S. 1; für eine ausführliche Diskussion der Begrifflichkeit siehe z. B. Tilman Schalmey 2009: „Überlegungen zur Konzeption eines neuen Lehrbuchs für das Klassische Chinesische“. Magisterarbeit. München: Ludwig-Maximilians-Universität, S. 7–11; Mit dem Problem, dass die Definition einer „Klassischen“ Sprache *flexibel* ist, haben nicht nur Sinolog:innen zu kämpfen, ähnlich schwammige Terminologien existieren auch für das Hebräische. Siehe z. B. Ian Young und Robert Rezetko 2014: *Linguistic Dating of Biblical Texts*. Bd. 1. Routledge, S. 8.

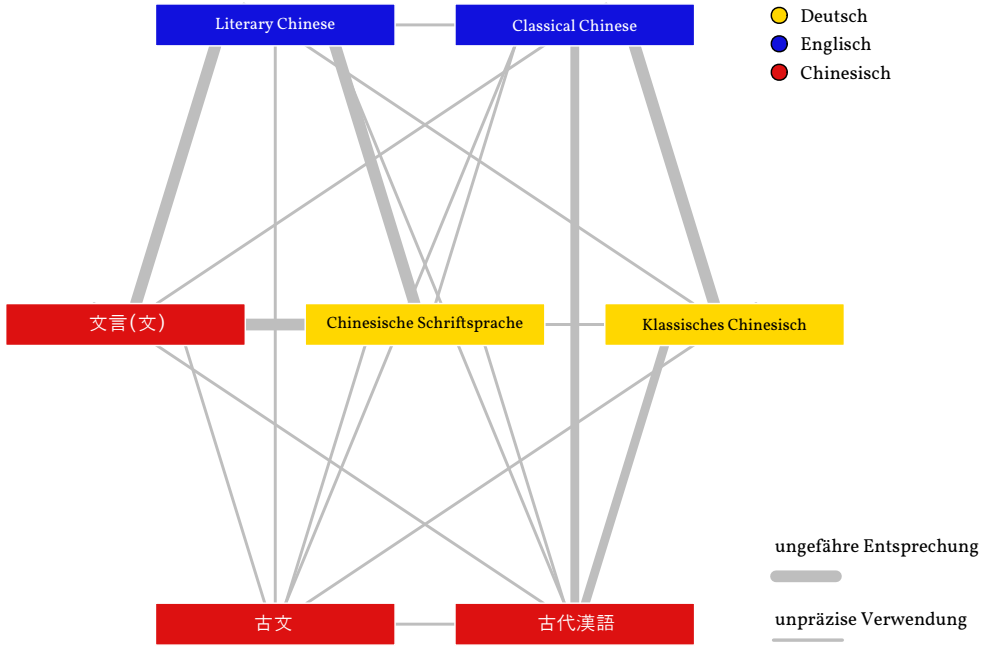


Abbildung 1.1 Klassisches und schriftsprachliches Chinesisch: Übersetz- und Austauschbarkeit der Begriffe

referring to the written form of Old Chinese, the period from the Spring & Autumn period to the end of the Han Dynasty."[33]

Die problematische Austauschbarkeit dieser eigentlich unterschiedlichen Bezeichnungen *Literary* und *Classical Chinese* (Abb. 1.1) hat ihren Ursprung sicherlich in der Dehnbarkeit der chinesischen Begriffe *gudai Hanyu* 古代漢語, *wenyan (wen)* 文言 (文) und *guwen* 古文. Vor allem die beiden letzteren implizieren – wie auch *schriftsprachliches Chinesisch* – häufig keinerlei epochale Einteilung. Stattdessen beziehen sie sich teils eher auf einen schriftsprachlichen *Stil* – im Kontrast zur (gesprochenen) Umgangssprache (*baihua* 白話)[34] der jeweiligen Zeit – oder, wie von der Gabelentz formuliert, Chinesisch „mit Ausschluss des niederen Stiles und der heutigen Umgangssprache".[35] Abgesehen von diesen temporalen und stilistischen Aspekten ist bereits das Konzept des Klassischen Chinesischen an sich eine „kühne Vereinfachung der sprachlichen Vielfalt, die im alten China existierte",[36] denn bereits für die Zeit der Frühlings- und Herbstannalen (*Chunqiu shidai* 春秋時代, ca. 770–476 v. u. Z.)[37] sind dialektale Unterschiede bzw. unterschiedliche Sprachen textlich belegt.[38] Über die Dialekte dieser Zeit ist jedoch zu wenig bekannt, um sie in einer Untersuchung computerlinguistischer Datierungsmethoden separat zu betrachten. Zudem ist es „offensichtlich,

33 Jerry Norman 1988: *Chinese*. Cambridge: Cambridge University Press, S. 83. Normans Definition ist tatsächlicher sinnvoller und präziser, da die uns vorliegenden Fassungen klassischer Texte zumeist durch eine frühestenfalls Han-zeitliche (漢, 202 v. u. Z.–220) Redaktion gegangen sind. Siehe z. B. Martin Kern 2004: „Die Anfänge der chinesischen Literatur". In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 46.

34 Für schriftliche Zeugnisse dieser Umgangssprache sind auch die Begriffe *tongsu* 通俗 bzw. im Englischen (*written*) *vernacular* gebräuchlich.

35 Hans Georg Conon von der Gabelentz 1881: *Chinesische Grammatik: mit Ausschluss des niederen Stiles und der heutigen Umgangssprache*. Leipzig: T. O. Weigel, S. 11.

36 Kai Vogelsang 2021: *Introduction to Classical Chinese*. Oxford: Oxford University Press, S. 363, übersetzt durch den Verfasser.

37 Wilkinson 2000, S. 10.

38 Siehe Norman 1988, S. 208–209; siehe auch Vogelsang 2021, S. 363–364.

dass eine Art vornehmer Standard entstanden war".[39] Geographische Aspekte der Sprachentwicklung sollen hier also außer Acht gelassen werden.

Inhaltsübersicht

Ungeachtet der sprachgeschichtlich unbefriedigenden Ungenauigkeit des Begriffs *schriftsprachliches Chinesisch* unterteilt man die Epochen in der historischen Linguistik selbstverständlich feiner. Kapitel 2 (ab S. 11) ist der Erarbeitung eines Grundverständnisses chinesischer Sprachgeschichte, des Sprachwandels und insbesondere des Wortschatzwandels gewidmet. Es fällt auf, dass in der historischen Sprachwissenschaft – der traditionellen chinesischen wie der westlichen[40] – besonders der Aspekt des phonologischen Wandels für das Chinesische intensiv untersucht wird, während der lexikalische Wandel eher stiefväterlich behandelt wird.[41] Anhand der offiziellen Dynastiegeschichten (*zhengshi* 正史) werden zudem beispielhaft konkrete Erscheinungsformen des syntaktischen und vor allem des lexikalischen Wandels beobachtet (Kapitel 2.3, ab S. 20). Dieses Korpus ermöglicht solche diachronen Betrachtungen innerhalb einer stilistisch gefestigten Textgattung über einen Zeitraum von mehr als 2.000 Jahren.

Kapitel 3 (ab S. 35) gibt einen Überblick über die für westliche Sprachen zur Verfügung stehenden computerlinguistischen Textdatierungsmethoden. Besondere Aufmerksamkeit wird dabei den eingangs erwähnten statistischen Sprachmodellen zuteil. Die bestehenden Ansätze werden mit Blick auf ihre Eignung für die Anwendung auf chinesischsprachiges – vor allem schriftsprachliches bzw. klassisches Textmaterial beleuchtet.

In Kapitel 4 (ab S. 59) wird geprüft, ob und wie gut sich bestehende computerlinguistische Ressourcen, Methoden und Tools für die Verarbeitung schriftsprachlicher Texte nutzen lassen. Bevor geeignet erscheinende Datierungsmethoden angewandt werden können, müssen passende diachrone Test- und Trainingskorpora festgelegt werden (Kapitel 4.2, ab S. 62). Überdies muss der Einfluss besonderer Eigenschaften der chinesischen Schrift und Schriftsprache auf die Verwendung computerlinguistischer Werkzeuge berücksichtigt werden. Während z. B. das fast vollständige Fehlen von Flexion[42] von Vorteil sein kann, erschweren Zeichenvarianten (Kapitel 4.3, ab S. 69) und vor allem das Fehlen von Leerzeichen zur Markierung von Wortgrenzen quantitative Analysen. Ein zentraler Aspekt ist daher die zur Erstellung von Worthäufigkeitslisten notwendige Möglichkeit der Segmentierung bzw. Tokenisierung von Texten (Kapitel 4.4–4.6, ab S. 73). Implikationen ergeben sich auch für die Erkennung von Personennamen (Kapitel 4.7, ab S. 97) und *temporal expressions* (Kapitel 4.8, ab S. 103).

In Kapitel 5 (ab S. 107) werden Entstehungsgeschichte und Datenqualität des *HYDCD* untersucht. Neben den rein lexikographischen Informationen, Lexemen und ihren Bedeutungen,

39 Norman 1988, S. 209, übersetzt durch den Verfasser.

40 Während eine professionelle historische Linguistik im Westen erst eine Erscheinung des späten 18. Jahrhunderts ist (siehe Roger Lass 2014: „Lineage and the Constructive Imagination: The Birth of Historical Linguistics". In: *The Routledge Handbook of Historical Linguistics*. Hrsg. von Claire Bowern und Bethwyn Evans. London & New York: Routledge, S. 45–63, v. a. S. 45–48;), deutet sich mit der in China bereits zur Han-Zeit florierenden Kultur der Kommentarliteratur mit ihren erklärenden Glossen für die Aussprache und Bedeutung von Zeichen in früheren, kanonischen Texte ein frühes Bewusstsein von Sprachwandel an. Siehe z. B. Kern 2004, S. 82–87. Solche Kommentare im Kontext der *xiaoxue* 小學 (Philologie) stellen damit eine frühe Form der historischen Sprachwissenschaft dar.

41 Siehe James H.-Y. Tai und Marjorie K. M. Chan 1999: „Some Reflections on the Periodization of the Chinese Language". In: *Studies in Chinese Historical Syntax and Morphology: Linguistic Essays in Honor of Mei Tsu-lin*. Hrsg. von Alain Peyraube und Sun Chaofen. Paris: École des Hautes Études en Sciences Sociales, S. 223–239, S. 227, siehe auch S. 233.

42 Siehe Axel Schuessler 2015: „Old Chinese Morphology". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill, für eine aktuelle Diskussion über das Vorhandensein von Flexion in früheren Entwicklungsstufen des Chinesischen.

enthält es zahlreiche Textbelegstellen, die – bestenfalls – die früheste Verwendung dieser Wörter dokumentieren. Daraus wird eine diachrone *mySQL*[43] -Lexemdatenbank aufgebaut. Die so strukturierten Daten können bei geringer Redundanz bequem für unterschiedliche Zwecke abgefragt werden. Die spärlichen bibliographischen Angaben im *HYDCD* werden unter anderem mit Daten aus dem *China Biographical Database Project (CBDB)*[44] verknüpft, was eine genauere chronologische Einordnung der lexikalisierten Wörter anhand ihrer Belegstellen ermöglicht. Aus den Textbelegen wird zudem ein provisorisches, diachrones Behelfskorpus für den Zeitraum vom 7. Jh. v. u. Z. bis ins 20. Jh. erzeugt (Kapitel 5.6, S. 137).

Die gewonnenen Daten werfen ein neues Licht auf die Machart dieses wichtigen Standardwerks und ermöglichen darüber hinaus Beobachtungen zum Wortschatzwachstum und -wandel des Chinesischen (Kapitel 5.7, ab S. 138).

Kapitel 6 (ab S. 155) widmet sich der Entwicklung und Anwendung von computerlinguistischen Datierungsmethoden für schriftsprachliches Chinesisch. Die statistischen Methoden aus der westlichen Computerlinguistik werden erstmals für die Datierung chinesischsprachiger Texte implementiert und mit verschiedenen (Behelfs-)korpora getestet (Kapitel 6.1, ab S. 156). Temporale Sprachmodelle erweisen sich dabei als für die chronologische Einordnung schriftsprachlicher Texte grundsätzlich geeignet. Versuche mit dem *HYDCD*-Arbeitskorpus zeigen, dass auch über einen langen Betrachtungszeitraum hinweg eine grobe Einordnung möglich ist. Der Erfolg bleibt aber stark vom Stil der zu datierenden Texte und den verfügbaren Trainingsdaten abhängig. Während das Verständnis von Sprachwandel ein Fundament der linguistischen Textdatierung darstellt, können die verwendeten statistischen Modelle andersherum auch genutzt werden, um Sprachwandel zu erkennen.[45]

Wünschenswert wäre eine von Genre und Trainingskorpus unabhängige Datierung schriftsprachlicher Texte. Eine lexikographische Herangehensweise auf Basis der diachronen Lexemdatenbank könnte dies ermöglichen. In Kapitel 6.2 (ab S. 179) wird eine Methode vorgestellt, anhand derer Texte aufgrund der enthaltenen Zeichenkombinationen chronologisch eingeordnet werden. Auch Personennamen und die Erkennung von *temporal expressions* können hierfür eingesetzt werden. Die so erzeugten temporalen Textprofile eignen sich neben einer automatisierten Datierung auch als Ausgangspunkt für eine qualitative Analyse.

Darüber hinaus wird untersucht, ob sich auch eine stark abstrahierte Messgröße, die durchschnittliche Entstehungszeit der in einem Text nachgewiesenen Lexeme (*Average Year of Lexicalization*, *AYL*) eignet, um Rückschlüsse auf seine Entstehungszeit zu ziehen (Kapitel 6.3, ab S. 210). Experimente offenbaren einen linearen Zusammenhang zwischen *AYL* und Textgenese, der zur Datierung aber nur bedingt herangezogen werden kann.

In Kapitel 6.4 (ab S. 229) wird ein Vergleich der vorgestellten linguistischen Datierungsmethoden angestrebt und auf Vor- und Nachteile der unterschiedlichen Herangehensweisen eingegangen. Zuletzt entwickle ich ein *user interface*, das mit geringen Vorkenntnissen die Verwendung der erarbeiteten Methodik im *Browser* ermöglicht. Benutzer:innen erhalten damit

43 *mySQL* ist eine verbreitete *Structured Query Language* (*SQL*). *SQL*s sind Programmiersprachen zur Formulierung von Datenbankabfragen. Die Daten werden in einer tabellenartigen Struktur gehalten und können so mit allen verbreiteten, modernen Programmiersprachen problemlos gelesen, geschrieben, sortiert, durchsucht, miteinander verknüpft und transformiert werden. Siehe z. B. Edwin Schicker 2017: *Datenbanken und SQL*. 5. Aufl. Wiesbaden: Springer Vieweg, S. 3–7.

44 Michael A. Fuller 2017: *China Biographical Database Project (CBDB)*. URL: `https://projects.iq.harvard.edu/cbdb` (besucht am 24. 04. 2017) (im Folgenden zit. als *CBDB*).

45 Siehe de Jong, Rode und Hiemstra 2005, S. 1. Diese Perspektive wird von den Autor:innen zwar angesprochen, aber nicht weiter beleuchtet und findet auch in der auf dieser Studie aufbauenden Arbeit anderer Forscher:innen wenig Beachtung. In Kapitel 6.1.4 (ab S. 177) wird diese Idee aufgegriffen.

Anhaltspunkte für die zeitliche Einordnung eines unbekannten Texts. Die Arbeit einer Philolog:in kann damit nicht ersetzt, aber unterstützt und erleichtert werden.

In Kapitel 7 (ab S. 243) werden wichtige Erkenntnisse dieser Studie zusammengefasst und Limitationen der (computer-)linguistischen Datierung schriftsprachlicher chinesischer Texte diskutiert. Abschließend werden Ideen für künftige Erweiterungen und Verbesserungen lexikographischer und statistischer Datierungsmethoden skizziert.

2 Sprach- und Wortschatzwandel

„Since 'tis Nature's Law to change,
Constancy alone is strange."[1]
John Wilmot

Alle in dieser Arbeit vorgestellten Methoden der Textdatierung basieren primär auf dem Konzept des Sprachwandels.[2] Sprachwandel kann in unterschiedlichen Formen auftreten. Veränderungen in Wortschatz und -gebrauch, phonologischer Wandel, sowie morphologischer und syntaktischer Wandel.[3] Im Folgenden sollen für die Textdatierung relevante Erkenntnisse zum Sprachwandel zusammengefasst und in einen Kontext mit der historischen Entwicklung der chinesischen Sprache gestellt werden. Eine besondere Bedeutung wird dabei dem Wortschatzwandel bzw. lexikalischen Wandel zuteil, der einen Großteil der (langfristigen) Veränderungen der geschriebenen Sprache ausmacht.[4]

Dass Sprache sich im Laufe der Zeit verändert, kann sicherlich als Allgemeinplatz angesehen werden, denn „language […], like everything else, gradually transforms itself over the centuries. There is nothing surprising in this. […] it would be strange if language alone remained unaltered."[5] Sprachwandel lässt sich zudem durch statistische Beobachtungen empirisch nachweisen, indem einzelne Sprachmerkmale, z. B. die relative Häufigkeit einzelner Wörter oder bestimmter Strukturen auf einer Zeitachse betrachtet werden.[6] Bereits 1953 hat der Linguist Alvar Ellegård so über einen Zeitraum von etwa 300 Jahren Veränderungen in der Verwendung von „do" in der englischen Sprache analysiert. Auch wenn es ihm keineswegs um die Datierung von Texten anhand solcher Beobachtungen ging, sondern lediglich um die Beobachtung des Sprachwandels selbst,[7] kann das Wissen über genau solche Häufigkeitsveränderungen für die Textdatierung hilfreich sein – gerade bei Lexemen, die für andere Aufgaben der Korpuslinguistik als *stop words* aussortiert werden.[8] Während traditionell oft Trends einzelner sprachlicher Phänomene untersucht werden, können aus großen Datensätzen, wie sie durch den *Google nGram Service* bereitgestellt werden, auch Erkenntnisse darüber gewonnen werden, *welche* Phänomene sich überhaupt verändert haben,

1 John Wilmot (1647–1680), zitiert in Jean Aitchison 2001 [1991]: *Language Change – Progress or Decay*. 3. Aufl. Cambridge: Cambridge University Press, S. 3.
2 Siehe auch Kapitel 3.1, ab S. 40.
3 Siehe z. B. Joan Bybee 2015: *Language Change*. Cambridge: Cambridge University Press, S. 1–2.
4 Jean Aitchison bezeichnet diese Art der Betrachtung von Sprachwandel als „armchair method". Siehe Aitchison 2001 [1991], S. 19. Die *armchair method* stellt dabei den Gegenentwurf zur *tape-recorder method* dar, mit der gerade stattfindende Veränderungen in der gesprochenen Sprache *synchron* untersucht werden können.
5 Ebd., S. 4.
6 Vgl. dazu Abschnitt 2.3, ab S. 20.
7 Siehe Alvar Ellegård 1953: *The auxiliary do: the establishment and regulation of its use in English*. Gothenburg studies in English. Göteborg: Almquist & Wiksell. URL: https://books.google.de/books?id=VcRZAAAAMAAJ, *passim*.
8 Als *stop words* bezeichnet man sehr häufige Wörter oder Wortformen wie „der", „dass", die für die inhaltliche Texterschließung, z. B. für die Erstellung von Suchmaschinenindizes, nicht relevant sind. Siehe z. B. Ronen Feldman und James Sanger 2006: *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, S. 68.

Veröffentlicht in: Tilman Schalmey: *Computerlinguistische Datierung schriftsprachlicher chinesischer Texte*.
Heidelberg: HASP, 2023. DOI: https://doi.org/10.11588/hasp.1153

und diese Veränderungen modelliert werden.[9] Maciej Eder stellt dabei fest, dass Sprachwandel und stilistische Änderungen dabei nicht zuverlässig getrennt werden können, da sich – wie auch im Bereich der Stilometrie beobachtet werden konnte – zu bestimmten Zeiten dominante Genres gleichermaßen auf Worthäufigkeiten auswirken können wie der Sprachwandel.[10]

Während die „Universalität des Wandels […] eine empirische Feststellung"[11] zu sein scheint, sind Ursachen für, sowie Geschwindigkeit, Art und Umfang sprachlicher Veränderungen weniger offensichtlich.[12] Verbreitet ist die Theorie der „unsichtbaren Hand" (*Invisible Hand*), wobei Sprache als „Phänomen der dritten Art", das irgendwo zwischen natürlich und künstlich geschaffen liegt, betrachtet wird.[13] Erklärungsansätze nehmen zudem häufig Anleihen bei der Evolutionsbiologie.[14] Viele Ansätze sind dabei allerdings „keine Erklärungen […] oder Pseudoargumente"[15]

Das Auftreten neuer Wörter wird zurückgeführt auf kulturelle, politische, gesellschaftliche oder technologische bzw. wissenschaftliche Veränderungen und Sprachkontakt.[16] Beobachtbar ist zudem die Veränderung ihrer Bedeutung.[17] Auf der anderen Seite „gehen [Wörter] unter, weil […] die Sachen verschwunden sind, die sie bezeichnen."[18] Bezeichnungen für Objekte können verändert werden, um die Beschreibung zu präzisieren, oder Veränderungen in dem Objekt selbst widerzuspiegeln.[19] Allerdings können Veränderungen oft auch ohne tatsächliche oder offensichtliche Notwendigkeit stattfinden.[20] Das Auftreten neuer Wörter oder Formen bedingt zudem noch lange nicht das Verschwinden der vorherigen Form, denn alte und neue Form können, teils über Jahrhunderte, parallel existieren, ohne dass sich eine Form durchsetzt. In anderen Fällen kann die Ersetzung als graduelle Ausbreitung wiederum mit der Vergrößerung eines herabrollenden Schneeballs verglichen werden.[21]

Typische Gründe für das Verschwinden von Wörtern aus dem Wortschatz sind unter anderem die Verdrängung durch amtliche Wörter, unbeliebte Lautformen, die Durchsetzung von Synonymen durch sprachprägende Texte, Verdrängung von Dialektwörtern durch die (politische) Durchsetzung von Hochsprachen, sowie die Verwendung von Euphemismen.[22] Während gemeinhin von der Sprachwissenschaft das „Problem des Wortuntergangs […] sehr stiefmütterlich behandelt"[23] wird, denn „es liegt in der Natur der Sache […], daß […] Neubildungen, Neu-Übernahmen, neue Zusammensetzungen das Interesse stärker auf sich […] ziehen als die schwer zu verfolgenden Vorgänge des Verdrängt-Werdens und Erlöschens",[24] ist die Entste-

9 Vgl. Maciej Eder 2018: „Words that Have Made History, or Modeling the Dynamics of Linguistic Changes". In: *Digital Humanities 2018 Puentes-Bridges: Book of Abstracts*. Hrsg. von Jonathan Girón Palau und Isabel Galina Russell. Mexico City: El Colegio de México, S. 362–365, S. 362.
10 Siehe ebd., S. 362–363.
11 Rudi Keller 2003: *Sprachwandel*. 3. Aufl. Tübingen & Basel: A. Francke, S. 21.
12 Dies wird an anderer Stelle umfassend diskutiert, siehe z. B. Keller 2003; Aitchison 2001 [1991]; August Dauses 1990: *Theorien des Sprachwandels – Eine kritische Übersicht*. Stuttgart: Franz Steiner Verlag.
13 Siehe Keller 2003, S. 87–88, S. 93, siehe auch S. 209–210.
14 Siehe ebd., v. a. S. 193–206.
15 Roger Lass 1980: *On Explaining Language Change*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press, S. XI.
16 Siehe Nabil Osman 1988 [1971]: *Kleines Lexikon untergegangener Wörter – Wortuntergang seit dem Ende des 18. Jahrhunderts*. 5. Aufl. München: C. H. Beck, S. 246; siehe auch Ernst Gamillscheg 1928: *Die Sprachgeographie und ihre Ergebnisse für die allgemeine Sprachwissenschaft*. Bielefeld & Leipzig: Velhagen & Klasing, S. 46.
17 Siehe auch Aitchison 2001 [1991], S. 17.
18 Osman 1988 [1971], S. 245.
19 Siehe Alinei 2004, S. 228–229.
20 Siehe z. B. Keller 2003, S. 21.
21 Siehe Aitchison 2001 [1991], S. 99, S. 107, S. 113.
22 Siehe Osman 1988 [1971], S. 248–249.
23 Ebd., S. 11.
24 Ebd., S. 7–8.

hung neuer Wörter und Wortformen relativ gut erforscht. Auch für das Chinesische existieren einige Veröffentlichungen, die sich mit *xinci* 新詞 („Neologismen“) bestimmer Dynastien oder unterschiedlicher Kategorien beschäftigen.[25]

Neben den Ursachen des Sprachwandels wird auch seine Bewertung kontrovers diskutiert. Häufig wird ihm, auch in der wissenschaftlichen Literatur, ein negativer Charakter im Sinne eines „Verfalls“ beigemessen, denn „Neuerungen kommen uns meist erst einmal barbarisch vor.“[26] Aitchison gibt ihrem Standardwerk *Language Change* daher sogar den Untertitel *Progress or Decay?*.[27] Derartige Bewertungen sind jedoch für uns hier kaum relevant, interessanter sind die Fragen nach *Wie?*, *Wann?* und *Durch wen?*[28]

Das „Aussterben“ von Wörtern kann nur bedingt zum Zweck der Textdatierung eingesetzt werden: zum einen bestehen in der Regel keine zuverlässigen Aufzeichnungen über ihr Verschwinden. Selbst wenn man aus großen, diachronen Korpora das (vorläufig) letzte Auftreten von Wortformen ermitteln würde,[29] kann weder der Zeitpunkt des Aussterbens, noch das Aussterben selbst zuverlässig etabliert werden. Zwar lässt sich „nur in seltenen Fällen […] untergegangenes Sprachgut wieder zum Leben erwecken“[30], doch ist z. B. in der Dichtung auch die Verwendung veralteter Wortformen nicht unüblich.[31] Hinweise, die die chronologische oder stilistische Einordnung von Texten erleichtern, ergeben sich aber aus typischen Häufigkeiten bestimmter Wortformen während bestimmter Zeiträume.[32] Für mehrere westliche Sprachen konnte auf Basis von Daten des *Google n-Gram-Viewers*[33] gezeigt werden, dass die Veränderung des Kernwortschatzes gemeinhin langsamer voranschreitet, als dies bei weniger frequenten Wörtern der Fall ist.[34]

25 Siehe z. B. Zhuo Jing-Schmidt und Shu-Kai Hsieh 2019: „Chinese neologisms“. In: *The Routledge Handbook of Chinese Applied Linguistics*. Hrsg. von Huang Chu-ren 黃居仁, Zhuo Jing-Schmidt und Barbara Meisterernst. Abingdon, Oxon & New York: Routledge, S. 514–534, für einen allgemeinen Überblick; Siehe z. B. auch Zhao Jindan 趙金丹 2007: „《朱子語類》新詞新語初探 (Erste Untersuchungen über neue Wörter und Phrasen im «Zhuzi yulei»)“. Masterarbeit. Shǎnxi 陕西: Shǎnxi shifan daxue 陕西師範大學; z. B. Sun Xiaoxuan 孫曉玄 2011: „基于《漢語大詞典》語料庫的宋代新詞研究 (On the New Vocabulary of Song Dynasty Based on Corpus According to <The Great Chinese Dictionary>)“. Diss. Shandong daxue 山東大學, für eine Arbeit, der das *HYDCD* zugrunde liegt; Zhang Weizhong 張衛中 2016: „新词语与清末民初作家的科幻想象 (Neologism and the Imagination of Science Fiction Writers in the Late Qing Dynasty and the Early Republic of China)“. In: *Journal of Central China Normal University (Humanities and Social Sciences)* 华中师范大学学报 *(*人文社会科学版*)* 55.6, S. 103–109.

26 Keller 2003, S. 19. Keller verdeutlicht die Absurdität solcher Vorbehalte zu Veränderungen im Sprachgebrauch auch direkt, denn „wenn sie gang und gäbe geworden sind, belächeln wir die vorherige Version“. Vgl. auch ebd., S. 23.

27 Aitchison 2001 [1991], S. Ui.

28 Siehe ebd., S. 6.

29 Siehe z. B. Costin-Gabriel Chiru und Traian Rebedea 2014: „Archaisms and Neologisms Identification in Texts“. In: *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference*. Chișinău: IEEE. doi: 10.1109/RoEduNet-RENAM.2014.6955312, Durch Beobachtung von Worthäufigkeiten über den in *Google n-Grams* aufgezeichneten Zeitraum zwischen 1800 und 2000 sollen Wörter als Archaismen und Neologismen kategorisiert werden.

30 Osman 1988 [1971], S. 247.

31 Siehe z. B. ebd.

32 Siehe dazu Kapitel 3, ab S. 35, sowie auch Kapitel 6.1, ab S. 156.

33 Jean-Baptiste Michel et al. 2011: „Quantitative Analysis of Culture Using Millions of Digitized Books“. In: *Science* 331.6014, S. 176–182. doi: 10.1126/science.1199644.

34 Siehe Vladimir Bochkarev, Valery Solovyev und Søren Wichmann 2014: „Universals versus historical contingencies in lexical evolution“. In: *Journal of The Royal Society: Interface* 11.101. doi: 10.1098/rsif.2014.0841, S. 6–7; vgl. auch George Kingsley Zipf 1947: „Prehistoric ‘Cultural Strata’ in the Evolution of Germanic: The Case of Gothic“. In: *Modern Language Notes* 62.8, S. 522–530, S. 523. Zipf findet für den „core“ einen „conservatism that seems to preserve it relatively intact through the generations, […]. This core […] consists almost exclusively of the relatively most frequently used words […].“

Das Entstehen neuer Wörter lässt sich im Gegensatz zu ihrem Verschwinden in einzelnen Fällen relativ genau zeitlich einordnen, da es eine „der Lexikalisierung innewohnende Chronologie“ gibt, die Alinei als „lexical self-dating“ bezeichnet.[35] Er schreibt:

> Lexicalization of datable referents tends to have the same date as the referents themselves. In more general terms, the date of a word tends to coincide with the date of the event or the concept it denotes.[36]

Als Beispiele für Wörter, für welche diese Art der intrinsischen Datierung perfekt funktioniert, dienen Alinei zum einen Bezeichnungen für Erfindungen der Moderne wie *fox trot* (1915) oder *telegraph* (1805), aber auch Namen für Pflanzen, die während der frühen Neuzeit durch die Seefahrt nach Europa kamen, z. B. *tabacco* (Mitte 16. Jh.) oder *chocolate* (1580).[37] Die Beispiele aus beiden Epochen bleiben problematisch, da sich die ursprünglichen Bezeichnungen datierbarer Bezugsobjekte nach ihrer ersten Lexikalisierung geändert haben können.[38] Ein „celebrated example“ für eine solche Änderung im Chinesischen ist die Bezeichnung für Telefon, da sich anstatt des zunächst phonetisch transkribierten *delüfeng* 德律風 („te-le-phone“) später die Bezeichnung *dianhua* 電話 („elektrische Worte“) durchsetzte.[39] Relevant für die Datierung des *Begriffes* ist also doch eher die älteste schriftliche Überlieferung als die Erfindung oder Entdeckung des Referenten, was insbesondere auch für Bezugsobjekte gelten muss, die „dem Menschen selbst präexistent sind, wie die elementareren Aspekte der Natur: ‚Wasser‘, ‚Sonne‘, ‚Wind‘, sowie die Namen von Tieren, Pflanzen und dergleichen.“[40]

2.1 Das Piotrowski-Gesetz

In Untersuchungen von Sprachwandel wird regelmäßig auf das Piotrowski-Gesetz bzw. Piotrowski-Altmann-Gesetz Bezug genommen. Ein solches „Gesetz des logistischen Wachstums“[41] wurde ursprünglich bereits im 19. Jahrhundert von Pierre François Verhulst zur Beschreibung natürlicher Veränderung vorgeschlagen[42] und „gehört zu den grundlegenden Gesetzen der Entwicklung von selbstorganisierten Systemen.“[43] Anna Piotrowskaja und Rajmund Piotrowski griffen für ihre Überlegungen auf Erkenntnisse aus der Epidemiologie zurück: wie bei der Verbreitung eines Virus würde eine neue sprachliche Form andere Sprecher „infizieren“ und so eine zunehmende Verbreitung finden, bis nahezu die gesamte Population infiziert ist – modellierbar mit einer *s*-förmigen Kurve.[44] Bereits zuvor wurde die Modellierung der Veränderung

35 Alinei 2004, S. 211.
36 Ebd., S. 226.
37 Siehe ebd.
38 Siehe ebd., S. 228.
39 Siehe z. B. Hsieh Feng-fan 謝豐帆 2015: „Transcribing Foreign Names“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Leiden: Brill.
40 Alinei 2004, S. 228, übersetzt durch den Verfasser, siehe auch S. 229–230.
41 Karl-Heinz Best 2003: „Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes“. In: *Glottometrics* 6, S. 9–34, S. 31.
42 Ebd.
43 Juhan Tuldava 1998: *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Übers. von Gabriel Altmann und Reinhard Köhler. Trier: Wissenschaftlicher Verlag Trier, S. 140.
44 Siehe Edda Leopold 2005: „Das Piotrowski-Gesetz“. In: *Quantitative Linguistik – Quantitative Linguistics – An International Handbook / Ein internationales Handbuch*. Hrsg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. Berlin & New York: Walter de Gruyter, S. 627–633, S. 627–628; siehe z. B. auch Karl-Heinz Best und Zhu Jinyang 2006: „Sprachwandel im Chinesischen“. In: *Archív Orientální* 74.2, S. 203–214, S. 203.

einzelner sprachlicher Phänomene mit einer *s*-Kurve von Sprachwissenschaftlern diskutiert.[45] Eine solche Gesetzmäßigkeit ließ sich anhand unterschiedlicher Untersuchungen immer wieder empirisch belegen.[46] Formalisieren lässt sich eine solche *s*-Kurve mit einer Funktion wie:[47]

$$p(t) = \frac{1}{1 + a \times e^{-rt}}$$

Abbildung 2.1 zeigt diese Funktion für $a = 1$ und verschiedene Werte von r.

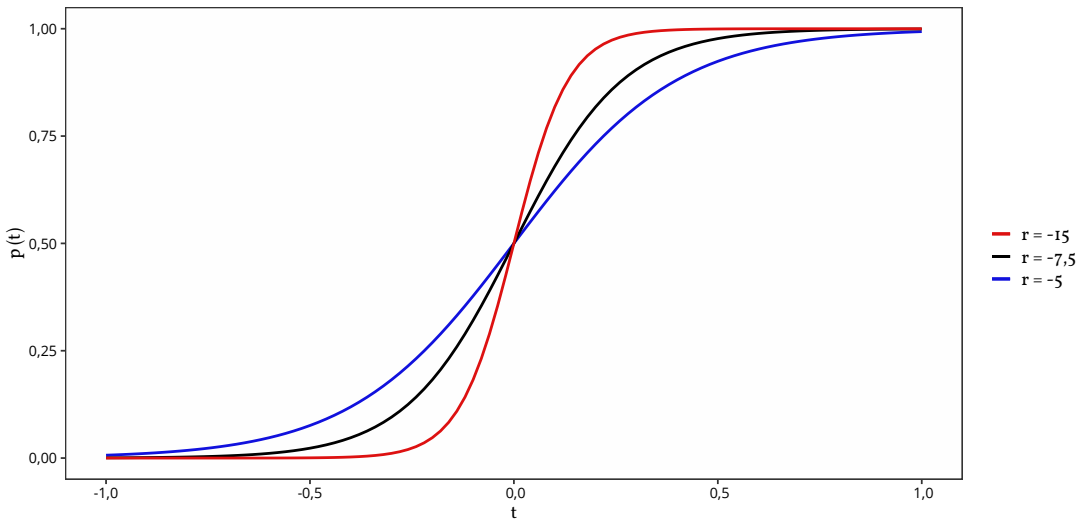


Abbildung 2.1 *p*(*t*) für unterschiedliche Werte von *r*, angelehnt an die Darstellung von Leopold.

Die Formalisierung von Sprachwandel mittels einer *s*-Kurve versetzt Sprachwissenschaftler:innen teilweise in die Lage, Vorhersagen oder Prognosen über „predictable grooves of change“[48] abzugeben. Der oben gezeigte Fall modelliert die vollständige Ersetzung eines sprachli-

45 Siehe Gabriel Altmann et al. 1983: „A law of change in language“. In: *Historical Linguistics*. Hrsg. von Barron Brainerd. Quantitative Linguistics 18. Bochum: Dr. N. Brockmeyer, S. 104–115, S. 204; siehe auch Charles E. Osgood und Thomas A. Sebeok 1954: *Psycholinguistics: A Survey of Theory and Research Problems*. Baltimore: Waverly Press, S. 155: „The process of change in the community would most probably be represented by an S-curve.“ siehe auch Uriel Weinreich, William Labov und Marvin I. Herzog 1968: „Empirical Foundations for a Theory of Language Change“. In: *Directions for Historical Linguistics*. Hrsg. von Winfried P. Lehmann und Yakov Malkiel. Austin: University of Texas Press, S. 95–195, S. 113: „the progress of a language change through a community follows a lawful course, an S-curve“.

46 Siehe z. B. Gabriel Altmann 1983: „Das Piotrowski-Gesetz und seine Verallgemeinerungen“. In: *Exakte Sprachwandelforschung*. Hrsg. von Karl-Heinz Best und Jörg Kohlhaase. Göttingen: Herodot, S. 59–90; zitiert in Leopold 2005, S. 628; siehe auch Best 2003, *passim* für eine Vielzahl von Anwendungsbeispielen; für eine aktuellere Zusammenfassung der möglichen Anwendungsgebiete siehe auch Kamil Stachowski 2020: „Piotrowski-Altmann law: State of the art“. In: *Glottotheory* 11.1, S. 1–12. doi: 10.1515/glot-2020-2002.

47 Siehe Leopold 2005, S. 628; Kroch schlägt vor, durch Logarithmierung den Zusammenhang als lineare Funktion zu notieren. In seiner Schreibweise: $ln\frac{p}{1-p} = k + s \times t$, wobei *s* als Steigung die *replacement rate* und *k* als Achsenabschnitt die Häufigkeit *p* einer neuen sprachlichen Form zum Zeitpunkt $t = 0$ angeben. Siehe Anthony S. Kroch 1989: „Reflexes of grammar in patterns of language change“. In: *Language Variation and Change* 1, S. 199–244. doi: 10.1017/s0954394500000168, S. 204.

48 Aitchison 2001 [1991], S. 114.

chen Phänomens nach einem „slow – quick – quick – slow pattern“.[49] Es lassen sich aber auch Fälle modellieren, in denen die sprachliche Neuerung unvollständig bleibt oder wieder zurückgeht.[50]

Karl-Heinz Best und Zhu Jinyang haben im Kontext des Wortschatzwandels für das Chinesische auf diese Weise sowohl die Zunahme von Schriftzeichen[51] als auch die Entwicklung der Wortlänge[52] modelliert, „obwohl die Datenlage nicht besonders reichhaltig ist“[53] und befinden auch für das Chinesische „das logistische Gesetz […] als gut geeignet […] um den beobachteten Verlauf von Sprachwandelprozessen zu modellieren.“[54] Mit geeigneten Daten lässt sich auch das Wortschatzwachstum im Chinesischen so modellieren.[55]

2.2 Sprachwandel mit „chinesischen Besonderheiten“

> *„Through the lens of neologisms, we can see history – lexical, social and cultural.“*[56]
>
> Zhuo Jing-Schmidt und Hsieh Shu-kai

Auch für das Chinesische ist der Sprachwandel sehr gut erforscht, wobei ein starker Fokus auf den phonologischen Wandel besteht. Schon in der Ming-Zeit (明, 1368–1644) führt der Gelehrte Chen Di 陳第 (1541–1617) fehlende Reime in alten Liedern auf einen phonologischen Wandel zurück.[57] Während der phonologische Wandel und Lautrekonstruktionen also schon in der frühen Neuzeit Gegenstand sprachwissenschaftlicher bzw. philologischer Arbeit waren, waren sprachliche Entwicklungen allgemein „vielleicht aufgrund der scheinbaren Unveränderlichkeit der Sprache hinter dem über die Jahrtausende stabilen Vorhang eines nicht-alphabetischen Schriftsystems“[58] in China weniger von Interesse. Der für die Textdatierung wichtige Aspekt des lexikalischen Wandels ist dabei am wenigsten systematisch erforscht. Implizite Hinweise auf die Etymologie meist einzelner Zeichen, die auch Rückschlüsse auf einen Bedeutungswandel zulassen, findet man aber oft in Kommentaren zu klassischen Texten. Sie können dem Bereich der *xunguxue* 訓詁學 genannten lexikalischen Exegese zugeordnet werden.[59] Wang Li 王力 geht in seinem Standardwerk *Hanyu shi*

49 Aitchison 2001 [1991], S. 91.
50 Siehe Best und Zhu Jinyang 2006, S. 205–206.
51 Siehe Best und Zhu Jinyang 2006, S. 208–209. Die „Zunahme der Schriftzeichen“ wird dabei allerdings mit nur sechs Datenpunkten vom Han-zeitlichen *Shuo wen jie zi* 說文解字 mit 9.553 Zeichen bis zum *Quan Hanzi shu* 全漢字書 von 1995 mit 70.000 Zeichen unter Ignoranz Ming- und Qing-zeitlicher Quellen modelliert. Eine vollständigere Datenlage dazu findet sich in Zhao Shouhui und Zhang Dongbo 2008: „The Totality of Chinese Characters–A Digital Perspective“. In: *Journal of Chinese Language and Computing* 17.2, S. 107–125, S. 108.
52 Siehe Best und Zhu Jinyang 2006, S. 209–211. Untersucht wurde hier die durchschnittliche Wortlänge in kurzen, diachronisch aber eklektisch ausgewählten Texten, wobei „ein eindeutiger Trend zu längeren Wörtern nachgewiesen“ wird.
53 Ebd., S. 206.
54 Ebd., S. 211.
55 Siehe dazu Kapitel 5.7.2, ab S. 142.
56 Jing-Schmidt und Hsieh 2019, S. 515.
57 Ausführlicher dazu siehe Johann-Mattis List 2013: „Theoretische und praktische Aspekte der quantitativen historischen Linguistik“. Seminarskript, Universität Marburg, S. 20.
58 Wolfgang Behr 2019: „‚Urheimat‘ der Chinesen. Die Sprachwissenschaft und die Suche nach ‚Wurzeln‘“. In: *Geschichte der Gegenwart*. URL: https://geschichtedergegenwart.ch/urheimat-der-chinesen-die-sprachwissenschaft-und-die-suche-nach-wurzeln/ (besucht am 24. 04. 2021).
59 Siehe z. B. William G. Boltz 2015: „Etymology“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill; Als Teil der *xiaoxue* 小學, „study of language and characters“, die eine Art frühe (historische) Sprachwissenschaft darstellt, steht dabei der hermeneutische Aspekt der Auslegung oft im Vordergrund. Siehe z. B. Wilkinson 2000, S. 60.

gao 漢語史稿 (*Entwurf einer Geschichte des Chinesischen*)[60] zwar darauf ein, die einzige Monographie über die Entwicklung des Wortschatzes bleibt aber *Hanyu cihui shi gaiyao* 汉语词汇史概要 (*Zusammenfassende Geschichte des Wortschatzes des Chinesischen*) von Pan Yunzhong 潘云中.[61] Hinzu kommen in jüngerer Zeit Aufsätze von Tai und M. K. M. Chan,[62] sowie Jing-Schmidt und Hsieh.[63]

Für die chinesische Sprache existieren unterschiedliche Periodisierungen, die ihre Entwicklung in Stufen beschreiben sollen. Für die syntaktische und phonologische Entwicklung werden dabei häufig voneinander abweichende Einteilungen vorgenommen.[64] Nicht nur dadurch gibt es für die Periodisierung aber viele ungeklärte Fragen.[65] Obwohl Sprachwandel gemeinhin als kontinuierlicher Prozess verstanden und modelliert wird, neigen auch Sprachwissenschaftler:innen zu Einteilungen, die stufenhafte Umbrüche zumindest suggerieren und sich zudem vor allem an historischen Epochen zu orientieren scheinen. Umgangssprachlich ist eine Dichotomie zwischen „modernem“ (*xiandai* 現代) und antikem (*gudai* 古代) Chinesisch gängig, wobei letzteres den Zeitraum von den frühesten schriftlichen Überlieferungen, ca. 1400 v. u. Z., bis 1911 beschreiben kann – ein Zeitraum, der sprachgeschichtlich kaum als konsistent gelten kann.[66] Ebenfalls verbreitet ist eine nicht besonders feingliedrige sprachgeschichtliche Unterteilung der Entwicklung der chinesischen Sprache in vier Epochen:[67]

1. Die klassische, antike, oder archaische (*shanggu* 上古) Periode, bis zum 3. Jh. v. u. Z.,
2. die mittlere oder mittelalterliche (*zhonggu* 中古) Periode, ab dem 3. Jh. v. u. Z. bis in die Mitte des 11. bzw. 13. Jh. (Mitte oder Ende der Song-Dynastie) und
3. die neuere (*jindai* 今代) Periode, bis zum Ende des 19. Jahrhunderts., bzw. Ende der Qing-Dynastie, sowie
4. heutiges (*xiandai* 現代) Chinesisch bzw. die Periode nach der 4. Mai-Bewegung (*wusi yundong* 五四運動) ab 1919, oder – je nach Auffassung – bereits nach dem 1. Opiumkrieg (1839–1842).[68]

Trotz Wang Lis Bemühungen um eine Berücksichtigung von Phonologie und Wortschatzwandel[69] handelt es sich hierbei vor allem um syntaktische Epochen, wobei *jindai* eher phonologisch und nur

60 Wang Li 王力 2011 [1958]: *Hanyu shi gao* 漢語史稿 (*Entwurf einer Geschichte des Chinesischen*). Beijing 北京: Zhonghua shuju 中華書局.

61 Pan Yunzhong 潘云中 1989: *Hanyu cihui shi gaiyao* 汉语词汇史概要 (*Zusammenfassende Geschichte des Wortschatzes des Chinesischen*). Shanghai 上海: Guji chubanshe 古籍出版社.

62 Tai und M. K. M. Chan 1999.

63 Jing-Schmidt und Hsieh 2019.

64 Ein präziser, ausführlicher Überblick über den Forschungsstand zur syntaktischen und phonologischen Periodisierung des Chinesischen ist Alain Peyraube 2015: „Periodization“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Leiden: Brill; für eine kurze Zusammenfassung siehe z. B. T. Schalmey 2009, S. 8; oder Tai und M. K. M. Chan 1999, S. 223; ausführlicher auch in Wilkinson 2000, S. 17–28.

65 Siehe Tai und M. K. M. Chan 1999, S. 223.

66 Siehe z. B. Alain Peyraube 2004: „Ancient Chinese“. In: *The Cambridge Encyclopedia of the World's Ancient Languages*. Hrsg. von Roger D. Woodard. Cambridge University Press, S. 988–1014, S. 988. Siehe auch Kapitel 1.

67 Eine genauere, vor allem syntaktisch begründete Unterteilung wird von Peyraube vorgeschlagen. Siehe Peyraube 2015.

68 Diese Unterteilung geht im Wesentlichen auf Wang Li 王力 2011 [1958], S. 43–44 zurück. Siehe aber z. B. auch Peyraube 2004, S. 989; oder Peter Kupfer 2009: „Language“. In: *Brill's Encyclopedia of China*. Hrsg. von Daniel Leese. Leiden & Boston: Brill, S. 544–549, S. 546; Während Wang Li die 4. Mai-Bewegung als Beginn der *xiandai*-Epoche sieht, argumentiert Masini für einen Beginn ab dem ersten Opiumkrieg. Siehe Federico Masini 2015: „Modern Lexicon, Formation“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill; eine ähnliche, ebenfalls lexikographisch begründete Auffassung findet sich auch in Pan Yunzhong 潘云中 1989, S. 146–163.

69 Vgl. Tai und M. K. M. Chan 1999, S. 226.

xiandai hauptsächlich durch Veränderungen im Wortschatz geprägt ist.[70] Eine abweichende Epochisierung der phonologischen Entwicklung richtet sich primär nach der zeitlichen Einordnung von Lautrekonstruktionen.[71] In einer quantitativen diachronen Untersuchung von Wortbildungsmustern auf Basis des „Sheffield Corpus of Chinese for Diachronic Linguistic Study"[72], ordnet Ji Meng die „morpho-syntactic patterns underpinning the evolution of Chinese lexis"[73] drei wesentlichen Epochen zu.[74] Während die historische Morphologie, Syntax und vor allem Phonologie sehr gut beforscht sind, ist es um die Erforschung des Wortschatzwandels weniger gut bestellt, denn „die Geschichte lexikalischer Veränderungen in der chinesischen Sprache ist sehr komplex und noch nicht gut erforscht."[75] Pan Yunzhong 潘云中 schlägt in seiner Studie eine Unterteilung in vier Hauptperioden des lexikalischen Wandels vor.[76] Diese sind in wesentlichen Zügen identisch mit der allgemeinen Unterteilung von Wang Li 王力 (siehe oben) – mit dem einzigen Unterschied, dass die *zhonggu* Epoche bereits in der Tang-Zeit (唐, 618–907) endet.[77] Unterschieden und sprachgeschichtlich teilweise getrennt wird dabei „native" und „borrowed vocabulary".[78]

Obwohl langfristig auch für das Chinesische insgesamt ein kontinuierliches, *s*-förmiges Wortschatzwachstum modelliert werden kann,[79] wird angenommen, dass Neologismen nicht unbedingt zu jeder Zeit gleichmäßig auftreten, sondern dass – gerade im Fall des *borrowed vocabulary* – in bestimmten Wellen vermehrt z. B. fremdsprachige Ausdrücke in die Sprache gelangen, denn „language contact during periods of disunity propelled language borrowing and language change"[80] und „in the history of Chinese, neologisms emerged in many waves."[81] Eine der bekanntesten und einflussreichsten dieser Wellen, die sich über einen längeren Zeitraum ab der östlichen Han-Zeit (25–220) bis in die Song-Zeit (960–1279) mit einem Höhepunkt etwa im 5. Jahrhundert erstreckt, entstand durch die Verbreitung von Übersetzungen buddhistischer Sutren aus Sanskrit संस्कृत und Prakrit प्राकृत ins Chinesische, wodurch sich Ausdrücke wie *Emituofo* 阿彌陀佛 (*Amitābha* अमिताभ), *niepan* 涅槃 (*nirvāna* निर्वाण) und *sēng* 僧 (*saṃgha* संघ, „Mönch") etablierten.[82] Weitere solche Wellen fanden ab Ende des 17. bis v. a. Ende des 19. / Anfang des 20. Jhs. durch

70 Siehe Tai und M. K. M. Chan 1999, S. 227; Wang Li selbst sieht als Hauptmerkmale seiner *xiandai*-Epoche eine „Absorption westlicher Grammatik und eine starke Zunahme mehrsilbiger Wörter" („適當地吸收西洋語法；大量地增加複音詞" Wang Li 王力 2011 [1958], S. 44.

71 Siehe T. Schalmey 2009, S. 8–9 für eine kompakte Zusammenfassung; ein ausführlicherer Überblick mit Erklärung der Methodik findet sich z. B. in Norman 1988, S. 39–76; rezentere Forschungsergebnisse sind dargestellt in: William H. Baxter und Laurent Sagart 2014: *Old Chinese: A New Reconstruction*. Oxford & New York: Oxford University Press.

72 Hu Xiaoling, Nigel Williamson und Jamie McLaughlin 2005: „Sheffield Corpus of Chinese for Diachronic Linguistic Study". In: *Literary and Linguistic Computing* 20.3, S. 281–293. DOI: 10.1093/llc/fqi034.

73 Ji Meng 2010: „A corpus-based study of lexical periodization in historical Chinese". In: *Literary and Linguistic Computing* 25.2, S. 199–213. DOI: 10.1093/llc/fqq002, S. 199.

74 Siehe ebd., Die Unterteilung in *Modern* (Ming, Qing), *Late Medieval* (Song und Yuan) und *Archaic* (Prä-Qin–Han) folgt dabei prinzipiell den Abteilungen des *Sheffield Corpus*.

75 Tai und M. K. M. Chan 1999, S. 233, übersetzt durch den Verfasse, siehe auch S. 226.

76 Siehe ebd., S. 233.

77 Vgl. Pan Yunzhong 潘云中 1989, S. 2–11; siehe auch Tai und M. K. M. Chan 1999, S. 233.

78 Siehe Tai und M. K. M. Chan 1999, S. 233–234; vgl. auch Wang Li 王力 2011 [1958], er unterscheidet zwischen dem Basiswortschatz (*jiben cihui* 基本詞彙) auf der einen und phonetischen Übertragungen (*jieci* 借詞) bzw. übersetzten Begriffen (*yici* 譯詞) auf der anderen Seite. Siehe S. 561, S. 587.

79 Siehe dazu Kapitel 5.7.2, ab S. 142.

80 Tai und M. K. M. Chan 1999, S. 225.

81 Jing-Schmidt und Hsieh 2019, S. 514; Vergleichbare Entlehnungswellen werden auch für andere Sprachen beschrieben, z. B. für das Deutsche aus dem Lateinischen, Französischen und zuletzt aus dem Englischen, siehe z. B. Elisabeth Knipf-Komlósi, Roberta V. Rada und Csilla Bernáth 2006: *Aspekte des Deutschen Wortschatzes*. Budapest: Bölcsész Konzorcium, S. 111–113.

82 Jing-Schmidt und Hsieh 2019, S. 515–516.

Wissenstransfer aus dem Westen, teilweise über Japan, durch die 4. Mai-Bewegung,[83] sowie im 20. Jh. durch politische Importe aus der Sowjetunion statt.[84]

Mit Blick auf Entwicklungen im Bereich der sog. Internetsprache (*wangluo yuyan* 網絡語言),[85] die nicht nur eine rapide Erweiterung des Wortschatzes, sondern auch Veränderungen in Morphologie und Syntax mit sich bringen, könnte man die vergangenen beiden Jahrzehnte sprachgeschichtlich bereits einer neuen, fünften Epoche zurechnen. Tatsächlich hat sich um diesen Themenkomplex in den vergangenen Jahren ein eigenes, sehr produktives Forschungsfeld etabliert.[86]

Während Sprachhistoriker:innen sich vornehmlich dem Aspekt des Wandels zuwenden, „wohl in der stillschweigenden Annahme ‚Wo sich nichts geändert hat, gibt es auch nichts zu erklären‘“,[87] umfasst die Evolution der Sprache nicht nur den Sprach*wandel* – Sprachen sind auch Traditionen: „Languages have an existence in some sense independent of that of their speakers: that is, they have traditions; perhaps, more accurately, they are traditions.“[88] Gerade für die Datierung schriftsprachlicher chinesischer Texte muss auch das inhärente Gegenteil des Sprachwandels, die Unveränderlichkeit über einen längeren Zeitraum, verstanden werden.

Eine wichtige Rolle für ein hohes Traditionsbewusstsein in der chinesischen Textkultur und damit für die Entwicklung bzw. Beständigkeit der Schriftsprache mag – neben der Schrift selbst – das chinesische Bildungs- bzw. Beamtenprüfungssystem (*keju* 科舉) gespielt haben. Es basiert weitestgehend auf konfuzianischer Ideologie; als Prüfungsaufgabe mussten oft Aufsätze über oder nach bestimmten antiken, stilistischen Vorbildern verfasst werden, vor allem zu kanonischen, historischen und philosophischen klassischen Texten und der zugehörigen Kommentartradition, sowie offiziellen Dokumenten. Ab dem 2. Jh. kanonisiert blieben gewisse Texte so beinah während der gesamten Kaiserzeit zentraler Bestandteil klassischer Bildung.[89] Während im Mittelalter auch zeitgenössische Poesie Teil der Prüfungen war,[90] und auch der Daoismus zum offiziellen Curriculum gehörte,[91] ging der Fokus während der Ming 明-Zeit (1368–1644) ab 1370 zurück zu Inhalten aus den sogenannten fünf Klassikern (*wujing* 五經) und vier Büchern (*sishu* 四書).[92] Dazu gehörte unter anderem die Vermeidung von modernem,

83 Eine ausführliche Darstellung findet sich bei Pan Yunzhong 潘云中 1989, S. 146–163; siehe z. B. auch Michael Lackner, Iwo Amelung und Joachim Kurtz 2001: „Introduction“. In: *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*. Hrsg. von Michael Lackner, Iwo Amelung und Joachim Kurtz. Leiden: Brill; Masini 2015; siehe auch Jing-Schmidt und Hsieh 2019, S. 516.

84 Siehe z. B. Jing-Schmidt und Hsieh 2019, S. 517.

85 Der Begriff Internetsprache definiert sich primär über das Internet als Kommunikationsmedium und umfasst eine Vielzahl an (teilweise neuartigen) Textgattungen wie Textnachrichten oder Blogs. Besonderheiten können z. B. die Verwendung von Emojis oder die Integration englischsprachiger Abkürzungen und Begriffe sein – als wesentliches Merkmal wird aber vor allem die schnelle Veränderlichkeit der Sprache ausgemacht. Siehe z. B. Eleni Andrist 2015: „Internet Language“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

86 Siehe z. B. Jing-Schmidt und Hsieh 2019, S. 517–519.

87 Keller 2003, S. 191.

88 Roger Lass 1984: *Language and Time: A Historian's View*. Inaugural lectures, University of Cape Town 90. Cape Town: University of Cape Town, S. 5; zitiert in Keller 2003, S. 186.

89 Siehe Thomas H. C. Lee 2000: *Education in Traditional China: A History*. Hrsg. von Erik Zürcher, Stephen F. Teiser und Martin Kern. Handbook of Oriental Studies, Section Four: China 13. Leiden: Brill, v. a. S. 21–22, S. 29, S. 256–257.

90 Bereits im 12. Jahrhundert sprach sich der neokonfuzianische Gelehrte Zhu Xi 朱熹 (1130–1200) für eine Abschaffung dieses Prüfungsteils aus. Siehe Hilde deWeerdt 2006: „Changing Minds through Examinations: Examination Critics in Late Imperial China“. In: *Journal of the American Oriental Society* 126.3, S. 367–377. doi: 10.2307/20064514, S. 368.

91 Siehe T. H. C. Lee 2000, S. 33.

92 Siehe Benjamin Elman 2013: „The Civil Examination System in Late Imperial China, 1400-1900“. In: *Frontiers of History in China* 8.1, S. 32–50. doi: 10.3868/s020-002-013-0003-9, S. 36.

buddhistischem Vokabular, umgangssprachlicher Phrasen und der „stylistic anarchy of popular novels“,[93] während die klassische Sprache als Idealbild propagiert wurde. Das Training für solche Prüfungen ermöglichte ein „kulturelles Repertoire linguistischer Zeichen und begrifflicher Kategorien“[94] von hoher Traditionalität. Auch namhafte Fürsprecher wie Han Yu 韓愈 (768–824) oder Ouyang Xiu 歐陽修 (1007–1072) beförderten immer wieder das Erstarken eines antiken Schreibstils (*guwen* 古文).[95] Watson spricht sogar von einem „starken Traditionalismus, der in allen chinesischen Kunstformen vorhanden ist“.[96] Die klassische Prägung und die Notwendigkeit, die Klassiker zu beherrschen, konnten die so gebildeten Gelehrten aber natürlich auch nicht von kreativem Schreiben in zeitgenössischer Sprache abhalten.[97] Vor diesem Hintergrund scheint die Entwicklung von *wenyan* 文言 und *baihua* 白話 als zweier quasi paralleler Sprachsysteme ab dem 3. Jahrhundert mit einer teils zunehmenden Spaltung zwischen gesprochener und geschriebener Sprache[98] und dem ausgeprägten Traditionalismus einer verhältnismäßig beständigen Schriftsprache wenig verwunderlich.

Eine weitere Ursache semantischer Rigidität bestimmter Wörter kann in ihrer Verwendung in Zeremonien und ihren Beschreibungen in vor-*Qin*-zeitlichen Ritenbüchern, z. B. im Rahmen von Dialogen der beteiligten Personen sowie honoriger Anreden gesehen werden.[99]

Ungeachtet stilistischer, syntaktischer und semantischer Unveränderlichkeit muss der Wortschatz einer Sprache stets erweitert werden, um „sich auf neue Dinge zu beziehen, neue Ideen auszudrücken, neue Identitäten zu konstruieren“[100], damit sie ein erfolgreiches Kommunikationsmittel bleibt.[101] Selbst bei Texten, die in einem „altertümelnden Stil“[102] verfasst sind, können die enthaltenen Lexeme oder Bezeichnungen also gegebenenfalls Hinweise auf die Entstehungszeit sein.[103]

2.3 Sprachwandel im Chinesischen am Beispiel der *zhengshi* 正史

Bei den *zhengshi* 正史 handelt es sich um eine Textgattung „offizieller“ chinesischer Geschichtsschreibung.[104] In der westlichen sinologischen Literatur wird der Begriff häufig als *Standard*

93 Elman 2013, S. 36.
94 Ebd., S. 41, übersetzt durch den Verfasser.
95 Siehe T. H. C. Lee 2000, S. 256–257; siehe auch Richard L. Davis 2004: *Historical Records of the Five Dynasties*. New York: Columbia University Press, S. xliv; siehe z. B. auch William Watson 1973: „On Some Categories of Archaism in Chinese Bronze“. In: *Ars Orientalis* 9, S. 1–13, S. 1–3.
96 Watson 1973, S. 1, übersetzt durch den Verfasser.
97 Siehe Elman 2013, S. 34–40.
98 Siehe z. B. Masini 2015; siehe dazu auch Norman 1988, S. 105, der über das Klassische Chinesisch schreibt: „In the Postclassical period, writers continued to model their prose on this early literary language, and the written languages thus began to take on an archaic aspect as the spoken language underwent a very different and by and large independent development.“
99 Siehe z. B. Ye Guoliang 葉國良 2008: *„Xian Qin lishu zhong baocun de guyu ji qi yiyi* 先秦礼书中保存的古语及其意义 (Archaismen aus vor-Qinzeitlichen Ritenbüchern und ihre Bedeutungen)“. In: *Journal of Northwest University (Philosophy and Social Sciences Edition)* 西北大学学报 *(*哲学社会科学版*)* 38.1, S. 86–90, S. 89.
100 Jing-Schmidt und Hsieh 2019, S. 514, übersetzt durch den Verfasser.
101 Siehe Jing-Schmidt und Hsieh 2019, S. 514; eine gegensätzliche Auffassung findet sich z. B. bei Keller 2003, S. 20: „Neuerungen in unserer Welt sind weder notwendig noch hinreichend für Veränderungen in unserer Sprache.“
102 Günther Debon 1989: *Chinesische Dichtung: Geschichte, Struktur, Theorie*. Handbook of Oriental Studies, Section Four: China 2. Leiden: Brill, S. 90.
103 Siehe dazu auch Kapitel 6.1, ab S. 156 und 6.2, ab S. 179.
104 Während die ältesten Texte dieses Genres von Privatgelehrten geschrieben wurden, war spätestens ab der Tang-Zeit meist eine offizielle staatliche Behörde involviert.

Histories, *Official Histories* oder *Dynastic Histories* übersetzt.[105] Er steht in der Regel für 24 Dynastiegeschichten (*ershisi shi* 二十四史), die anhand von Annalen, Tabellen, Biographien über Herzöge und Fürsten (*shijia* 世家) und andere beispielhafte Persönlichkeiten (*liezhuan* 列傳), sowie Monographien (*shu* 書) zumeist über die vorangegangene Dynastie und deren Persönlichkeiten, Ereignisse und Politik berichten. Den Anfang dieser Tradition bildet das hauptsächlich von Sima Tan 司馬談 (gest. 110 v. u. Z.) und dessen Sohn Sima Qian 司馬遷 (ca. 145–86 v. u. Z.) verfasste *Shiji* 史記.[106] Während im *Shiji* die gesamte Frühgeschichte bis zum mythischen *Huangdi* 黃帝 („Gelber Kaiser") aufgearbeitet wird, befassen sich die meisten späteren Standardgeschichten mit je exakt einer – zum Veröffentlichungszeitpunkt bereits beendeten – Herrscherdynastie.[107] Üblicherweise wurde jeweils die Vorgängerdynastie aufgearbeitet, so dass heute ein stilistisch recht homogenes Textkorpus vorliegt, dessen – wenn man die 1928 veröffentlichte *Qingshi gao* 清史稿 (*Draft History of the Qing*) hinzunimmt – 25 sehr unterschiedlich lange Texte über einen Zeitraum von insgesamt 2.019 Jahren hinweg verfasst wurden.[108] Abb. 2.2 und Tabelle 2.1 geben einen Überblick über die historische Verteilung von Veröffentlichung bzw. Entstehung und von den Texten inhaltlich abgedeckten Zeitperioden.[109]

Vierthaler hat die – im Hinblick auf den langen Betrachtungszeitraum verblüffende – stilistische Homogenität dieses Textkorpus in einer stilometrischen Studie mittels einer *Principal Component Analysis* (*PCA*) eindrücklich gezeigt, wobei die Geschichtstexte ein gemeinsames *Cluster* bilden, von denen sich literarische Texte klar abgrenzen lassen.[110] Trotzdem können in diesem recht rigiden, schriftsprachlichen Korpus sowohl Veränderungen an der Häufigkeit wichtiger Partikel, sowie vor allem auch das Aufkommen neuer Begriffe gezeigt werden.

Untersucht man anhand der *zhengshi* sprachliche Veränderungen aus einer diachronen Perspektive, darf nicht vergessen werden, dass ein Korpus mit so wenigen Einzeltexten und nur eines Genres keine allgemeingültigen Schlüsse über die Sprache der jeweiligen Zeit oder sprachliche Entwicklungen zulässt. Die Sprache der *zhengshi* kann auch von Stil und Vorlieben der jeweiligen Autor:innen, Kompilator:innen und Herausgeber:innen geprägt sein, wobei die Aspekte des Schreibstils und der sprachgeschichtlichen Entwicklung kaum trennbar sind.[111] Auch kann zum Zeitpunkt der Fertigstellung einer *zhengshi* als kompiliertes Werk ein Teil des darin verwendeten Textmaterials schon mehrere hundert Jahre alt sein – aus der Zeit, über die geschrieben wird.[113] Eine weitere Schwäche dieses Textkorpus besteht in den ab dem 7. Jh. ungleich großen zeitlichen Abständen zwischen der Entstehung neuer Dynastiegeschichten.

105 Siehe Wilkinson 2000, S. 501.
106 Siehe z. B. ebd.
107 Siehe z. B. ebd., S. 502.
108 Hier von *ershiwu shi* 二十五史 (25 Dynastiegeschichten) zu sprechen wäre irreführend, da die erst 1920 veröffentlichte *Xin Yuan shi* 新元史 hier oft den 25. Text darstellt. Da ein Zeitraum von 523–714 Jahren zwischen Kompilation (1890–1920) und der darin beschriebenen Zeitperiode (1206–1367) liegt, wird sie für das hier zusammengestellte Textkorpus jedoch nicht verwendet. Siehe dazu auch ebd., S. 505.
109 Zu Datenformaten und -quellen der hier verwendeten Ausgaben siehe Kapitel 4.2, S. 65.
110 Vgl. Vierthaler 2016a, *passim*, v. a. S. 18–19. Zur *PCA* siehe auch Kapitel 3.1, S. 41.
111 Vgl. auch Eder 2018, S. 362–363.
112 Beschriebener Zeitraum und Jahr der Veröffentlichung bzw. Präsentation der Texte aus Wilkinson 2000, S. 503–505; Wilkinson gibt die Veröffentlichung des *Han shu* mit dem Jahr 92 an, einige Teile wurden aber erst nach Ban Gus Tod fertiggestellt. Siehe z. B. Anthony François Paulus Hulsewé 1993a: „Han shu 漢書". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 129–136, S. 129.
113 Als Beispiel sei hier das *Hou Han shu* 後漢書 (*HHS*) genannt, für das Fan Ye 范曄 mehr als zweihundert Jahre *nach* Ende der Han-Dynastie Han-zeitliche Dokumente verwertete. Zur Entstehungsgeschichte des *Hou Han shu* 後漢書 siehe Hans Bielenstein 1954: „The Restoration of the Han Dynasty. With Prolegomena on the Historiography of the Hou Han Shu". In: *BMFEA [Bulletin of the Museum of Far Eastern Antiquities]* 26, S. 1–209, S. 9–17. Siehe dazu auch Kapitel 5.7.4, ab S. 150 u. 6.3, ab S. 210.

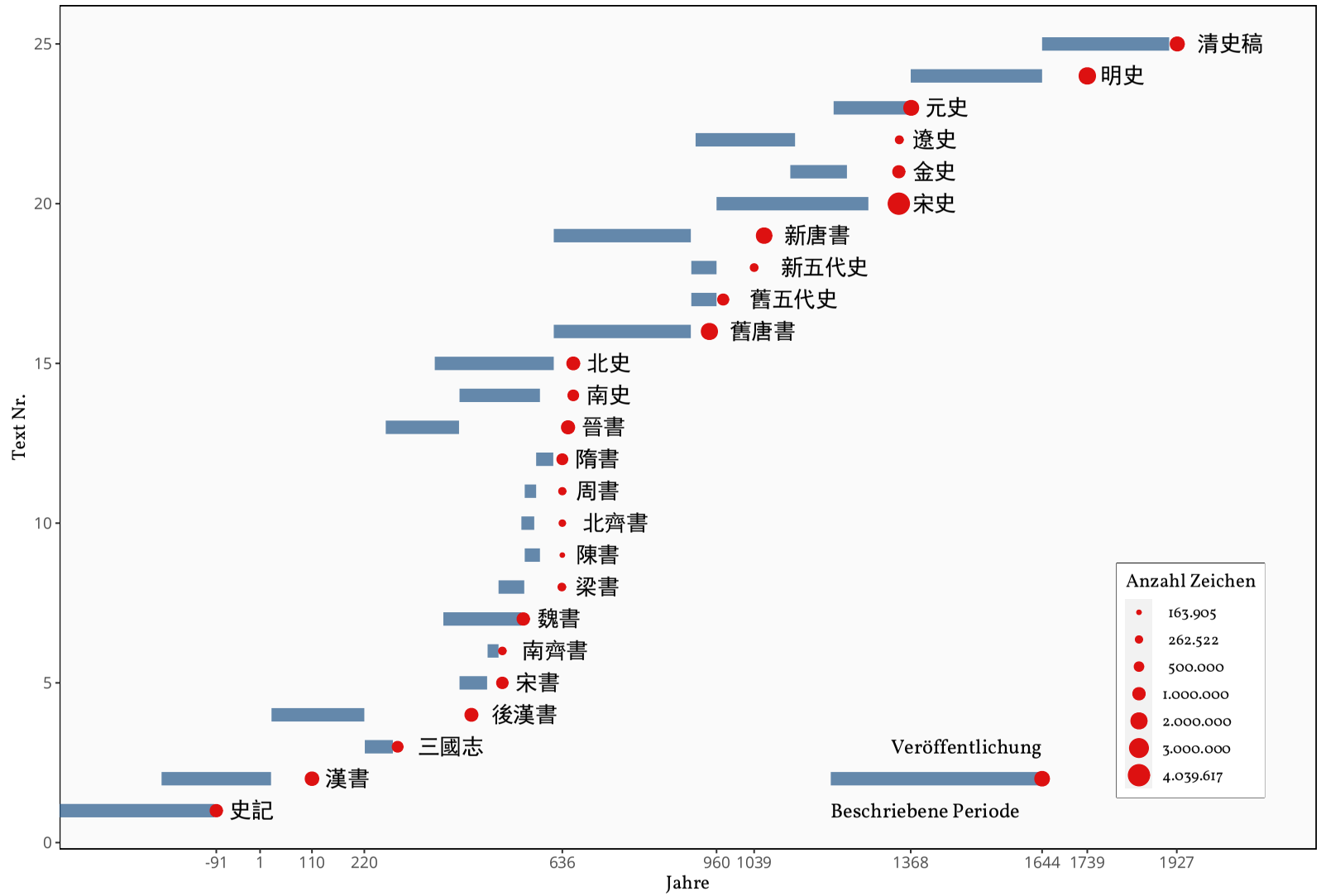


Abbildung 2.2 *zhengshi* nach Jahr der Veröffentlichung. Die Größe der Datenpunkte spiegelt den Umfang des Textes (Anzahl Schriftzeichen) wider. Die vom jeweiligen Text abgedeckte historische Zeitperiode wird durch einen graublauen Balken dargestellt.

Tabelle 2.1 *Ershisi shi* 二十四史 und *Qing shi gao* 清史稿

Nr.	Text	Zeitraum[112]	朝代	Autor(en) bzw. Hrsg.	# *juan*	Veröffentlichung[112]
1	*Shiji* 史記	bis -95	西漢	Sima Qian 司馬遷	130	-91
2	*Han shu* 漢書	-206–24	東漢	Ban Gu 班固	100	ca. 110
4	*Sanguo zhi* 三國志	221–280	西晉	Chen Shou 陳壽	65	297
3	*Hou Han shu* 後漢書	25–220	劉宋	Fan Ye 范曄	120	ca. 445
6	*Song shu* 宋書	420–478	梁	Shen Yue 沈約	100	492–493
7	*Nan Qi shu* 南齊書	479–502	梁	Xiao Zixian 蕭子顯	59	537
10	*Wei shu* 魏書	386–550	北齊	Wei Shou 魏收	114	554
8	*Liang shu* 梁書	502–556	唐	Yao Silian 姚思廉	56	636
9	*Chen shu* 陳書	557–589	唐	Yao Silian 姚思廉	36	636
11	*Bei Qi shu* 北齊書	550–577	唐	Li Baiyao 李百藥	50	636
12	*Zhou shu* 周書	557–581	唐	Linghu Defen et. al. 令狐德棻等	50	636
13	*Sui shu* 隋書	581–617	唐	Wei Zheng et. al. 魏徵等	85	636
5	*Jin shu* 晉書	265–419	唐	Fang Xuanling et. al. 房玄齡等	130	646
14	*Nan shi* 南史	420–589	唐	Li Yanshou 李延壽	80	659
15	*Bei shi* 北史	368–618	唐	Li Yanshou 李延壽	100	659
16	*Jiu Tang shu* 舊唐書	618–906	后晉	Liu Xu et. al. 劉昫等	200	945
18	*Jiu Wu Dai shi* 舊五代史	907–960	北宋	Xue Juzheng et. al. 薛居正等	150	974
19	*Xin Wu Dai shi* 新五代史	907–960	北宋	Ouyang Xiu 歐陽脩	74	1072
17	*Xin Tang shu* 新唐書	618–906	北宋	Ouyang Xiu, Song Qi 歐陽脩、宋祁	225	1060
20	*Song shi* 宋史	960–1279	元	Toqto'a et. al. 脱脱等	496	1343
22	*Jin shi* 金史	1115–1234	元	Toqto'a et. al. 脱脱等	135	1343
21	*Liao shi* 遼史	916–1125	元	Toqto'a et. al. 脱脱等	116	1344
23	*Yuan shi* 元史	1206–1369	明	Song Lian et. al. 宋濂等	210	1369
24	*Ming shi* 明史	1368–1644	清	Zhang Tingyu et. al. 張廷玉等	332	1739
25	* *Qing shi gao* 清史稿	1644–1911	民國	Zhao Erxun et. al. 趙爾巽等	529	1928

Dass das *zhengshi*-Korpus trotz der dargelegten Einschränkungen für die Beobachtung von Sprachwandel in der chinesischen Schriftsprache von der Han-Dynastie bis ins 20. Jh. interessant sein kann und dass auch in einem stilistisch sehr homogenen Korpus trotz der schriftsprachlichen Rigidität unterschiedliche Arten von Veränderung beobachtbar sind, soll hier an einigen Beispielen gezeigt werden.

Im Folgenden werden hierzu die Texte gemäß Tabelle 2.1 auf einer Zeitachse (x) und die Häufigkeit von Zeichen bzw. Lexemen in Vorkommen pro 100.000 *tokens* dargestellt.[114]

— 1. Betrachtet man die Vorkommen einiger in schriftsprachlichen Texten sehr häufiger **Funktionswörter** bzw. Partikel, *zhi* 之, *zhe* 者, *ye* 也, *yue* 曰, *wei* 爲/為 und *bu* 不 in allen *zhengshi* (Abb. 2.3),[115] lässt sich bei den meisten dieser Zeichen ein leichter Abwärtstrend erkennen. Vor allem bei *zhi* nimmt die Nutzung deutlich ab, aber auch bei *ye* und *yue* ist eine entsprechende Tendenz erkennbar. Sogar das in der modernen Sprache noch weit verbreitete *bu* scheint im

114 Um auch die Häufigkeit mehrsilbiger *tokens* zu erfassen, wird hier eine *n*-Gramm Zerlegung vorgenommen und die gefundenen *types* anschließend auf Basis der Einträge im *DHYDCD* auf bekannte Lexeme reduziert. Siehe dazu Kapitel 4.5.2, S. 91 und 4.5.3, S. 94.

115 Die genannten Zeichen gehören zu den 30 häufigsten im gesamten Korpus und sind zugleich die frequentesten – je nach Deutung – Grammatikpartikel bzw. Funktionswörter. Einige davon haben selbstverständlich weitere Bedeutungen, wie *wéi* 爲 als Verb („tun, machen, herstellen, sein" oder *zhi* 之 („gehen nach"), die aber ebenfalls eher schriftsprachlich verwendet werden.

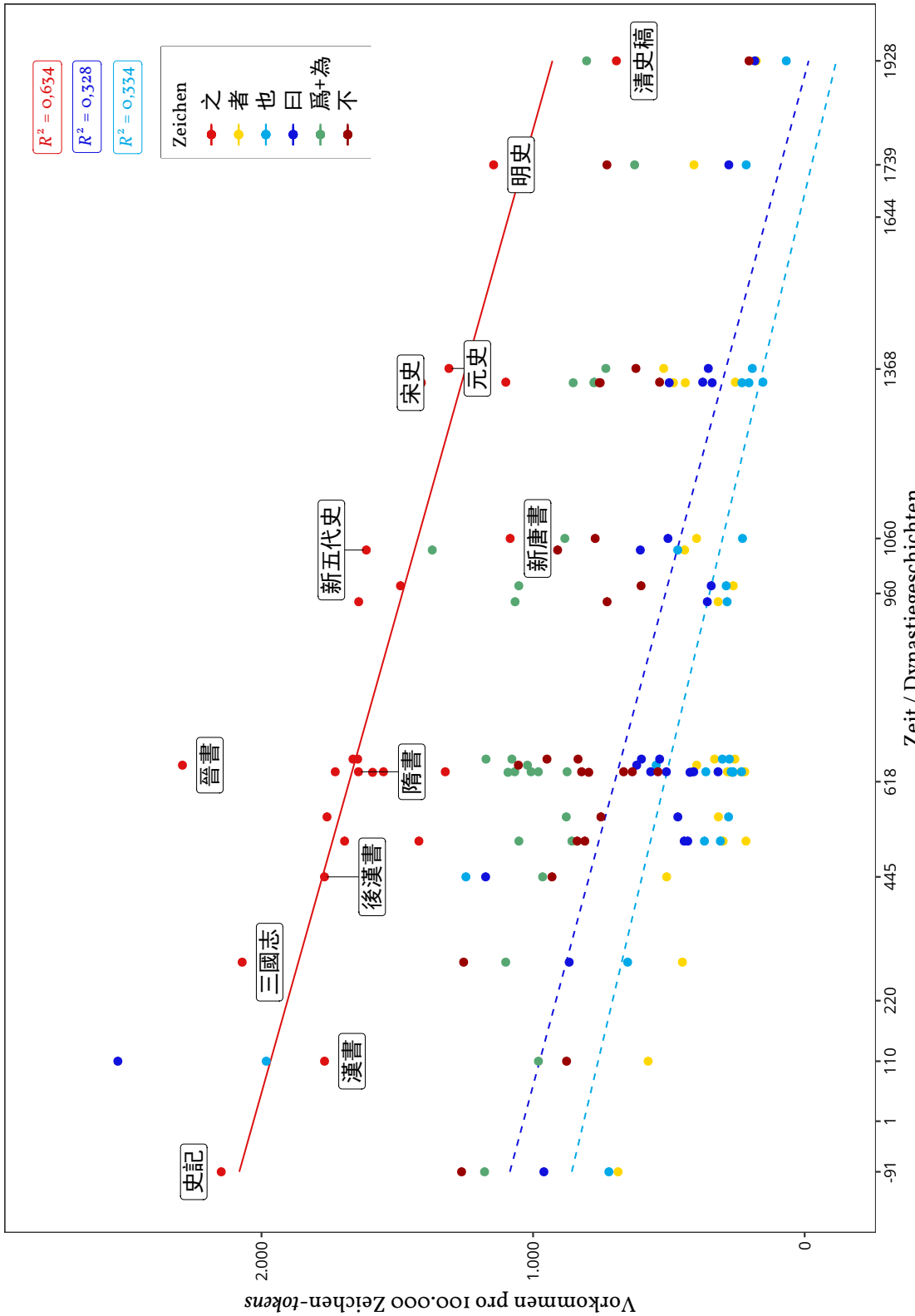


Abbildung 2.3 Abnehmende Nutzung typischer schriftsprachlicher Partikel in den *zhengshi*-Texten

Gebrauch seltener zu werden.[116] Lediglich bei der nominalisierenden Partikel *zhe* lässt sich anhand des *zhengshi*-Korpus keine Tendenz erkennen. Gerade *zhi* 之 und *wei* 為 können zudem – je nach Kontext – sehr unterschiedliche Funktionen und Bedeutungen haben, deren sprachliche Verwendung sicherlich separat betrachtet werden sollte.[117] Um die Häufigkeit bestimmter Verwendungen bzw. Bedeutungen von Zeichen eigenständig zu analysieren wäre zunächst wenigstens ein *Part-of-speech* (*PoS*) Tagging erforderlich, was für schriftsprachliche Texte aber nicht in befriedigender Qualität möglich ist.[118]

Der *CKIP Tagger*[119] erkennt z. B. beim *Tagging* der *zhengshi* drei unterschiedliche Wortarten von *zhi*: Subordinationspartikel („DE", 232.718 Vorkommen), Pronomen („Nh", 168.414 Vorkommen) und als „Hilfswort" („T", 2 vorkommen).[120] Eine diachrone Betrachtung der beiden als signifikant erkannten Verwendungen (Abb. 2.4) legt nahe, dass der Gebrauch als Subordinationspartikel tatsächlich deutlich abnimmt, während sich für das Pronomen kein eindeutiger Trend zeigt.[121] Wie schon die Einteilung fast aller Vorkommen von *zhi* in nur zwei Wortklassen verdeutlicht, müssen die verwendeten Daten als sehr ungenau eingestuft werden.[122] Klar ersichtlich wird aber das Potenzial, welches ein zuverlässiges *PoS-Tagging* für diachrone korpuslinguistische Untersuchungen des Chinesischen hätte. Neben der Möglichkeit der differenzierteren Analyse von Worthäufigkeiten müsste die Untersuchung auf weitere Texte und Genres ausgeweitet werden, um verlässliche Aussagen über sprachhistorische Entwicklungen machen zu können.

Auch mit kruden Mitteln können aber in Veränderungen von Häufigkeiten, die sich in den obigen Beispielen innerhalb von zwei Jahrtausenden vollziehen, teils deutliche Tendenzen abgelesen werden. Sie bestätigen zugleich – trotz der offensichtlichen Veränderung – auch eine hohe stilistische (oder sprachgeschichtliche) Rigidität der chinesischen Schriftsprache, oder zumindest der vorliegenden Textgattung.

— 2. **Ämter bzw. Amtstitel**. Mehr von historischem Interesse ist die Entwicklung der Verwendung einiger in den Dynastiegeschichten häufig genannter Titel bzw. Amtsbezeichnungen von Würdenträgern.[123] *Jiangjun* 將軍, *cishi* 刺史, *taishou* 太守, *yushi* 御史 und *jiedushi* 節度使 sind nur

116 Ähnliche Tendenzen lassen sich z. B auch bei der klassischen Präposition *yu* 於 („an, auf" usw.), *qi* 其 („ihr, sein, dessen"), *er* 而 („und, als, da, weil, obwohl" bei der Verbindung zweier Verbalphrasen) sehen.

117 Die wichtigsten Verwendungen von *zhi* sind als subordinierende Partikel, ähnlich dem modernen *de* 的, als Objektspronomen („sie, ihn, es"), sowie als transitives Verb „gehen nach". *Wei* tritt vor allem als Verb (*wéi* – „sein, machen") und als Präposition (*wèi* - „für") auf.

118 Siehe dazu Kapitel 4, ab S. 59.

119 Li Peng-Hsuan 李朋軒 und Ma Wei-Yun 馬偉雲 2019–: *CKIP Tagger*. GitHub Repository. URL: https://github.com/ckiplab/ckiptagger (besucht am 30. 05. 2021), In einer Prüfung der verfügbarer Tokenizer auf ihre Eignung für die Segmentierung schriftsprachlicher Texte, schneidet der *CKIP Tagger* von den Bibliotheken, die auch *PoS-Tagging* ermöglichen, mit am besten ab. Ausführlicher dazu siehe Kapitel 4.5, ab S. 81.

120 Die *CKIP*-Dokumentation erläutert das *Tag* „DE" als zusammenfassende Kategorie für *de* 的, *zhi* 之, *de* 得 und *de* 地. „Nh" sind Pronomen (*daimingci* 代名詞), „T" Partikel oder Hilfswörter (*yuzhuci* 語助詞) Siehe ebd., Wiki/POS Tags.

121 Eine lineare Regression auf die Häufigkeit von *zhi* als Subordinationspartikel ergibt ein R^2 von 0,652, eine Modellierung als *s*-Kurve hat hier hingegen einen schlechteren Erklärungsgehalt.

122 Die Bedeutung „gehen nach", also die Verwendung als transitives Verb, ist ebenfalls im *zhengshi* Korpus belegt. Das *HYDCD* gibt eine Belegstelle aus dem *Han shu* 漢書 an. Siehe *DHYDCD*, Bd. 1, S. 676, *zhi* 之.

123 Zur Auswahl geeigneter Beispiele wurde mit `pandas` die Varianz aller *HYDCD*-Lexeme im *zhengshi*-Korpus berechnet und Wörter mit hoher Varianz ausgesucht. Siehe The pandas development team 2020; Die Idee, in einem diachronen Korpus die statistische Varianz zum Aufspüren von Wörtern zu nutzen, deren Häufigkeit sich während des betrachteten Zeitraums verhältnismäßig stark ändert, ist inspiriert durch die Arbeiten von Eder 2018; sowie Theodoros Lappas et al. 2009: „On Burstiness-Aware Search for Document Sequences". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. Paris, France: Association for Computing Machinery, S. 477–486. DOI: 10.1145/1557019.1557075.

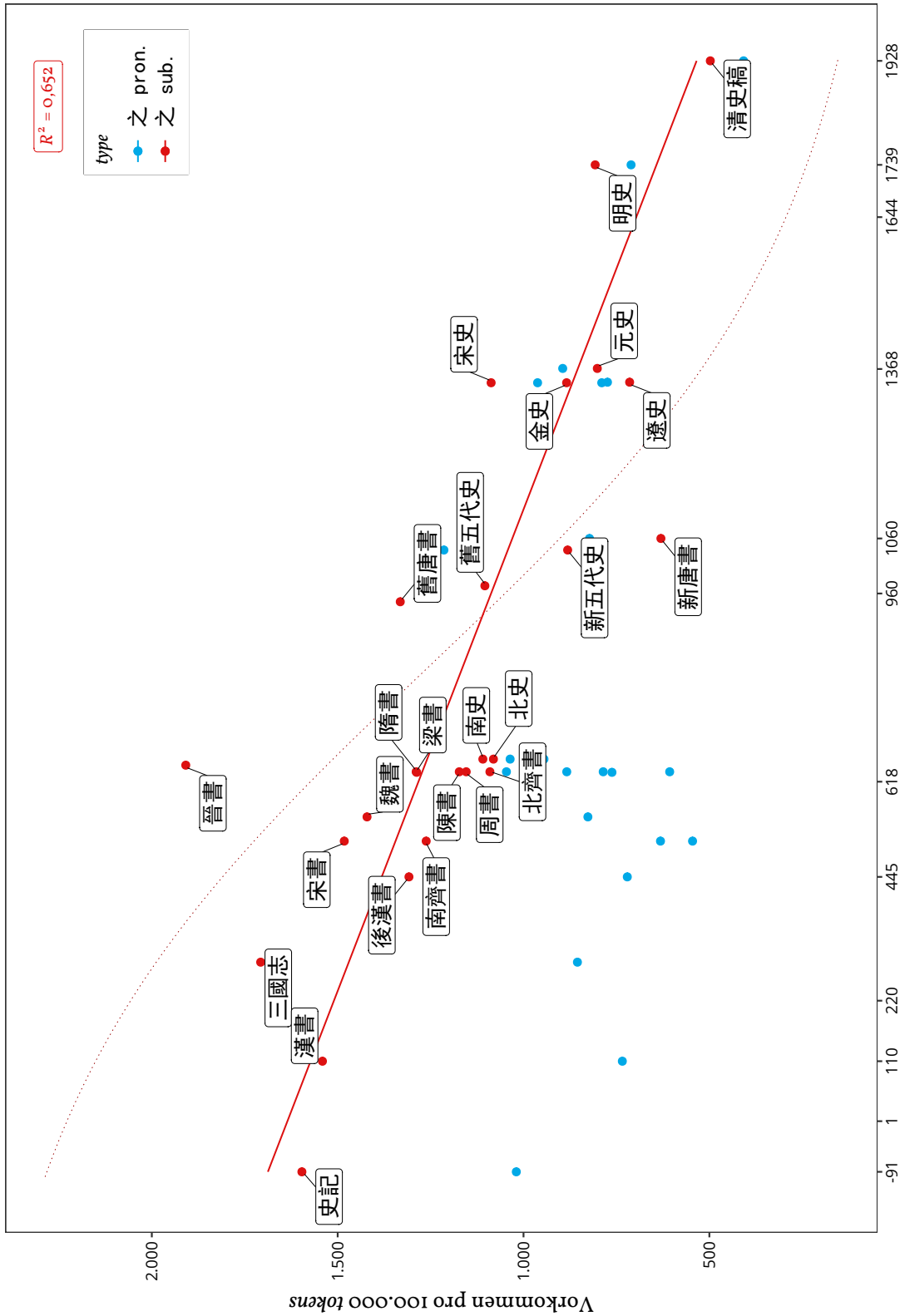


Abbildung 2.4 *Zhi* 之 als Pronomen und subordinierende Partikel in den *zhengshi* (ungefähre Daten!)

fünf von zahllosen Beispielen.[124] Abb. 2.5 zeigt die relative Häufigkeit der genannten Bezeichnungen in allen 25 Texten auf einer Zeitachse. Für jeden der Amtstitel ist eine Verbindungslinie zwischen den einzelnen Texten eingezeichnet. Diese soll hier keine sprachlichen Trends symbolisieren, sondern lediglich Unterbrechungen und Schwankungen in der Wortverwendung im Korpus besser sichtbar machen. Dass ein überwiegender Teil dieser Titel überhaupt in einem so langen Zeitraum teils durchgängig Verwendung findet, spiegelt sicherlich die Traditionsbewusstheit des kaiserlichen Verwaltungsapparates wider. Ungeachtet der Einführung immer neuer, spezifischer Beamtentitel, wurden wesentliche Strukturen und Ämter in großen Teilen auch von Fremddynastien wie den mongolischen Yuan 元 übernommen.[125]
Die Verwendung von *jiangjun* („General"), „throughout history the most common term for the commander of a substantial body of troops",[126] zieht sich durch das gesamte Textkorpus. Nach der Han- bis zur Tang-Zeit (唐, 618–906) scheint die Verwendung deutlich zuzunehmen, was auch für die Bezeichnung *cishi* („Regional Inspector/Chief, Prefect") gilt. Bei letzterem spiegelt die in Abb. 2.5 sichtbare Veränderung in der Häufigkeit den Bedeutungswandel und die damit einhergehende Wichtigkeit und Häufigkeit der Ernennung wider.[127] Ebenfalls während der gesamten Kaiserzeit wird der Titel *yushi*, „Royal Scribe, Censor" genutzt, „the standard generic designation of central government officials […] maintaining […] surveillance over the officialdom […]."[128]

Taishou war von der Qin- 秦 (221–206 v. u. Z.) bis Sui 隨-Zeit (581–617) die Bezeichnung für den Gouverneur einer territorialen Einheit – ein Titel, der während der Tang-Zeit abgeschafft wurde.[129] Entsprechend ist für diesen Zeitraum eine Abnahme der Nennungen zu beobachten, in der *Xin Wudai shi* 新五代史 wird kein *taishou* erwähnt (Abb. 2.5). Ab der Song 宋-Zeit (960–1279) wird er verwendet als „common quasiofficial or unofficial reference to a Prefect"[130] mit offensichtlich weiterhin wenigen Erwähnungen.

Auch wenn Hucker bereits für die Zeit der Drei Reiche (*Sanguo* 三國, 221–280) den Titel *jiedushi* als „Supply Commissioner" angibt, bekommt er erst ab der Tang-Zeit die historisch wichtige Bedeutung „Military Commissioner".[131] Die dreisilbige Kombination *jiedushi* ist tatsächlich erst in Tang-zeitlichen Texten nachgewiesen,[132] während *jiedu* 節度 – wörtlich etwa „kontrollieren und messen" – tatsächlich bereits in den Chroniken der drei Reiche (*Sanguo zhi* 三國志) vorkommt. Deutlich in Abb. 2.5 erkennbar ist auch die gestiegene Bedeutung während der Zeit der Fünf Dynastien (*Wudai* 五代, 907–960), während der „Military Commissioners continued as virtually autonomous satraps in their regions."[133] Während der Song-Zeit wurden Amt und Titel dann allmählich abgeschafft, wobei unter den Liao 遼 (916–1125) im Süden zahlreiche *jiedushi* eingesetzt wurden.[134]

124 Charles O. Hucker verzeichnet in seinem *Dictionary of Official Titles in Imperial China* über 8.000 solcher Amtstitel. Trotzdem schreibt er: „Do not expect comprehensive inclusiveness." Charles O. Hucker 1987 [1985]: *A Dictionary of Official Titles in Imperial China*. Taipei 台北 [Stanford]: Nantian shuju 南天書局 [Stanford University Press], S. 100.
125 Vgl. z. B. ebd., S. 58–69. In diesem Kontext wird häufig der Begriff „Sinisierung" genannt.
126 Ebd., S. 140.
127 „HAN–SUI: **Regional Inspector**, […] N[orthern and]-S[outhern] DIV[ision] […]–SUNG: **Regional Chief**, a title commonly awarded important heads of aboriginal tribes […] SUI–CHIN: **Prefect**; […] in Sung and Chin uncommon […]." Siehe ebd., S. 558–559. Hervorhebungen im Original.
128 Ebd., S. 592.
129 Siehe ebd., S. 482–483.
130 Ebd.
131 Siehe ebd., S. 144.
132 Vgl. *DHYDCD*, 節度使, vgl. auch Abb. 2.5.
133 Hucker 1987 [1985], S. 144.
134 Siehe ebd., S. 144, vgl. auch Abb. 2.5.

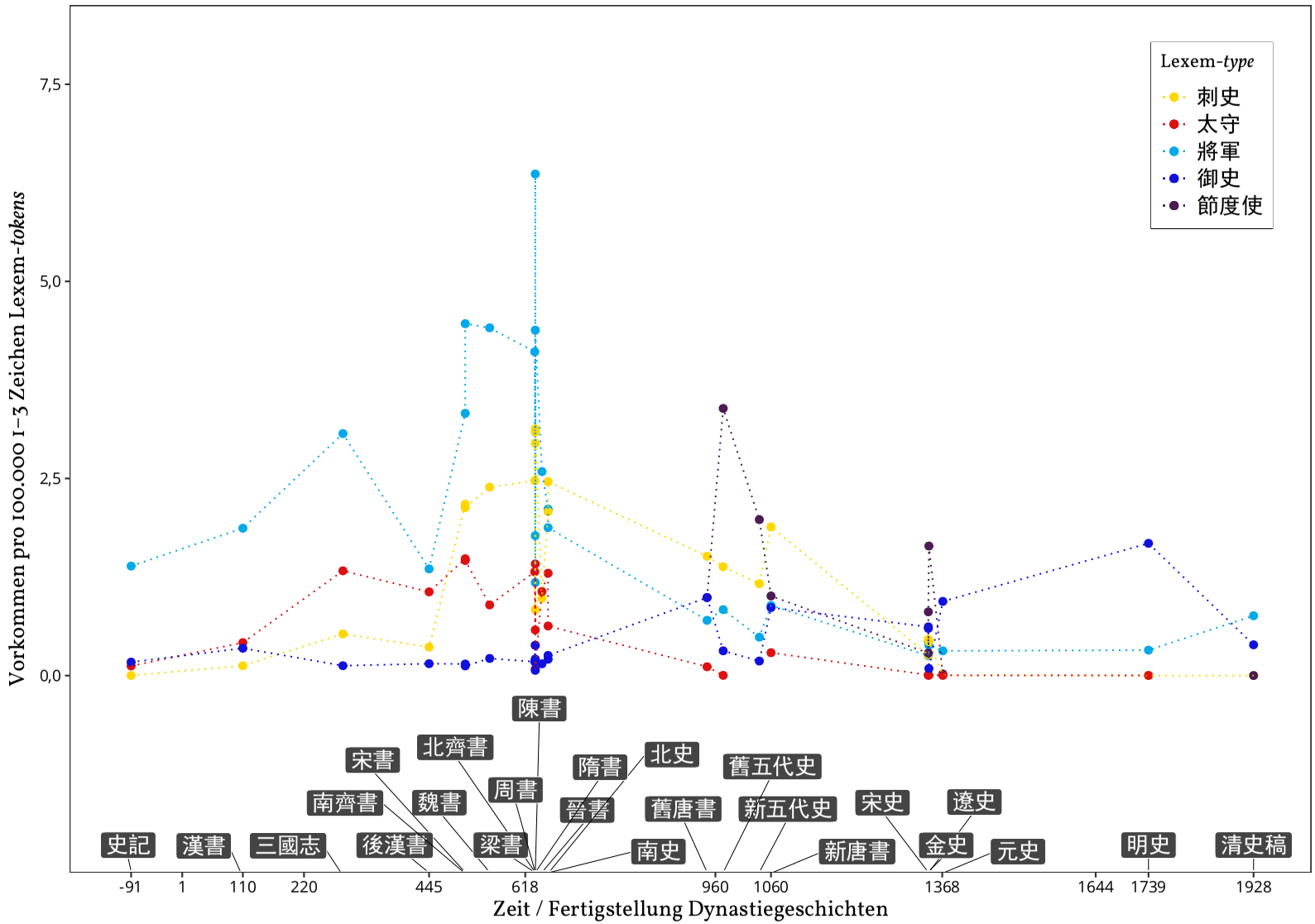


Abbildung 2.5 Verwendung einiger Amtstitel in den *zhengshi*

Das Beispiel zeigt die Relevanz dieser Art der Lexemnachweise für die Textdatierung. Ein Text, der z. B. den Begriff *jiedushi* enthält, ist vermutlich nicht deutlich früher als im 7. Jh. entstanden. Dass das *Jiu Tang shu*, in welchem der Begriff in Abb. 2.5 erstmalig erscheint, erst Mitte des 10. Jhs. kompiliert wurde, erinnert gleichzeitig an die Problematik der hier verwendeten Textreihe: Die beiden Aspekte der Zeit, zu der und über die geschrieben wird, sind verschoben ineinander verwoben und können nur durch sorgfältige, qualitative Arbeit voneinander gelöst werden.

— 3. **Neologismen**. Als Beispiele für Lexeme, die im Rahmen von in Kapitel 2.2 bereits angesprochenen Entlehnungswellen bzw. für Begriffe, die als Bezeichnung neuer Objekte Eingang in die chinesische Sprache finden, werden einige buddhistisch geprägte Lexeme (Abb. 2.6) betrachtet.[135]

Frühe Formen des Buddhismus, der ab dem 1. Jh. in China bekannt wurde, waren noch nicht klar vom Daoismus abgegrenzt. Seine eigentliche Verbreitung begann erst mit der Übersetzung buddhistischer Schriften ins Chinesische ab dem 2. Jh.[136]

Das Zeichen 佛, mit dem gewöhnlich das Morphem *fo* 佛 („Buddha" / „buddhistisch") geschrieben wird, ist in den Lesungen *fu* und *bi* bereits sehr früh belegt.[137] Im *Han shu* 漢書 wird 佛 v. a. in der Kombination *fangfu* 仿佛 („so scheinen wie", auch 髣髴 geschrieben) verwendet. Die heute in Japan übliche Schreibung von *fo* 仏 (jap. *butsu* ブツ, *hotoke* ほとけ usw.) wird im *HYDCD* als „alte Form von *fo*" („*fo de gu zi* , 佛' 的古字") bezeichnet, ist aber ebenfalls erst in der Zeit der Südlichen und Nördlichen Dynastien (*Nanbeichao* 南北朝, 420–581) nachgewiesen.[138] Im überwiegend aus Han-zeitlichem Material zusammengestellten *HHS* wird noch die ältere Transkription, *futu* 浮屠 bzw. 浮圖, („Buddha") verwendet,[139] im Ende des 3. Jh. fertiggestellten *Sanguo zhi* 三國志 findet sich sowohl *futu* 浮圖 als auch bereits *fo* in der in der Wortbildung *fojing* 佛經 („Sutren").[140] Spätestens in der Zeit der Südlichen und Nördlichen Dynastien (*Nanbeichao* 南北朝, 420–581) setzt sich für Buddha *fo* 佛 endgültig durch, was sich sowohl in einem deutlichen Anstieg der Zeichenhäufigkeit (Abb. 2.6) als auch in der Bildung weiterer Komposita wie *foxiang* 佛像 („Buddhastatue"), *fosi* 佛寺 („buddhistische(r) Tempel") und *fotu* 佛徒 („Buddhist") zeigt.[141]

---

135 In beiden Abbildungen sind Datenpunkte der frühesten Erwähnung im Korpus, sowie auffallend hohe Werte und durch Überlappung verdeckte Datenpunkte zur Verbesserung der Lesbarkeit mit dem Titel der jeweiligen *zhengshi* gekennzeichnet.

136 Siehe Rong Xinjiang 榮新江 2004: „Land Route or Sea Route? Commentary on the Study of the Paths of Transmission and Areas in which Buddhism was Disseminated during the Han Period". In: *Sino-Platonic Papers* 144, S. 2, S. 12, „Buddhism appeared to be only an exotic variant of the esoteric Daoism", siehe auch S. 27. Vgl. auch Henri Maspero 1981 [1950, 1971]: *Taoism and Chinese Religion [Le Taoïsme et les religions chinoises]*. Übers. von Frank A. Kierman jr. Amherst [Paris]: University of Massachusetts Press [Gallimard], S. 401–403.

137 Im *Lunyu* 論語 und im *Shiji* 史記 wird der Name 佛肹 Bi Xi (ca. 5. Jh. v. u. Z.) genannt: „Bi Xi war Gouverneur von Zhongmou." („Bi *Xi wei Zhong-mou zai* 佛肹為中牟宰") *HYDCD*, Bd. 1, S. 1288, 佛 5 肹. siehe auch Sima Qian 司馬遷 2008 [91 v. u. Z.] *Shiji* 史記 *(Records of the Grand Historian)*. Project Gutenberg eBook. URL: https://www.gutenberg.org/ebooks/24226 (besucht am 16. 05. 2021), Kapitel 47, *Kongzi shijia* 孔子世家; 佛 als alternative Schreibung zu *fu* 拂 ist z. B. mit dem *Liji* 禮記 belegt *HYDCD*, Bd. 1 , S. 1285, 佛 3. Wann das Zeichen in welchen Ausgaben wie geschrieben wurde geht daraus aber nicht hervor.

138 Siehe Hans-Jörg Bibiko 2006–: *Japanisch-Deutsches Kanji-Lexikon*. URL: https://mpi-lingweb.shh.mpg.de/kanji/ (besucht am 16. 05. 2021), 仏; vgl. auch Morohashi Tetsuji 諸橋轍次, Hrsg. 1955–1960: *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)*. Bd. 1–12. Tokyo 東京: Taishukan shoten 大修館書店, Bd. 1, S. 596, 仏; und *HYDCD*, Bd. 1, S. 1117, 仏.

139 Siehe Fan Ye 范曄 1965 [445]: *Hou Han shu* 後漢書. *12 Bde*. Beijing 北京: Zhonghua shuju 中華書局 (im Folgenden zit. als *HHS*), S. 1428; vgl. auch Erik Zürcher 2007 [1959]: *The Buddhist Conquest of China*. 3. Aufl. Sinica Leidensia XI. Leiden: Brill, S. 26; vgl. auch Rong Xinjiang 榮新江 2004, S. 17.

140 Chen Shou 陳壽 1971 [297]: *Sanguo zhi* 三國志. *5 Bde*. Beijing 北京: Zhonghua shuju 中華書局, S. 1185.

141 *Foxiang* 佛像 werden ab dem *Nan Qi shu* 南齊書, *fosi* 佛寺 ab dem *Song shu* 宋書 in fast allen *zhengshi* erwähnt (siehe Abb. 2.7), die Wortbildung *fotu* findet sich lediglich im *Sui shu* 隋書. Siehe Wei Zheng 魏徵 1973 [636]: *Sui shu* 隋書. *6 Bde*.

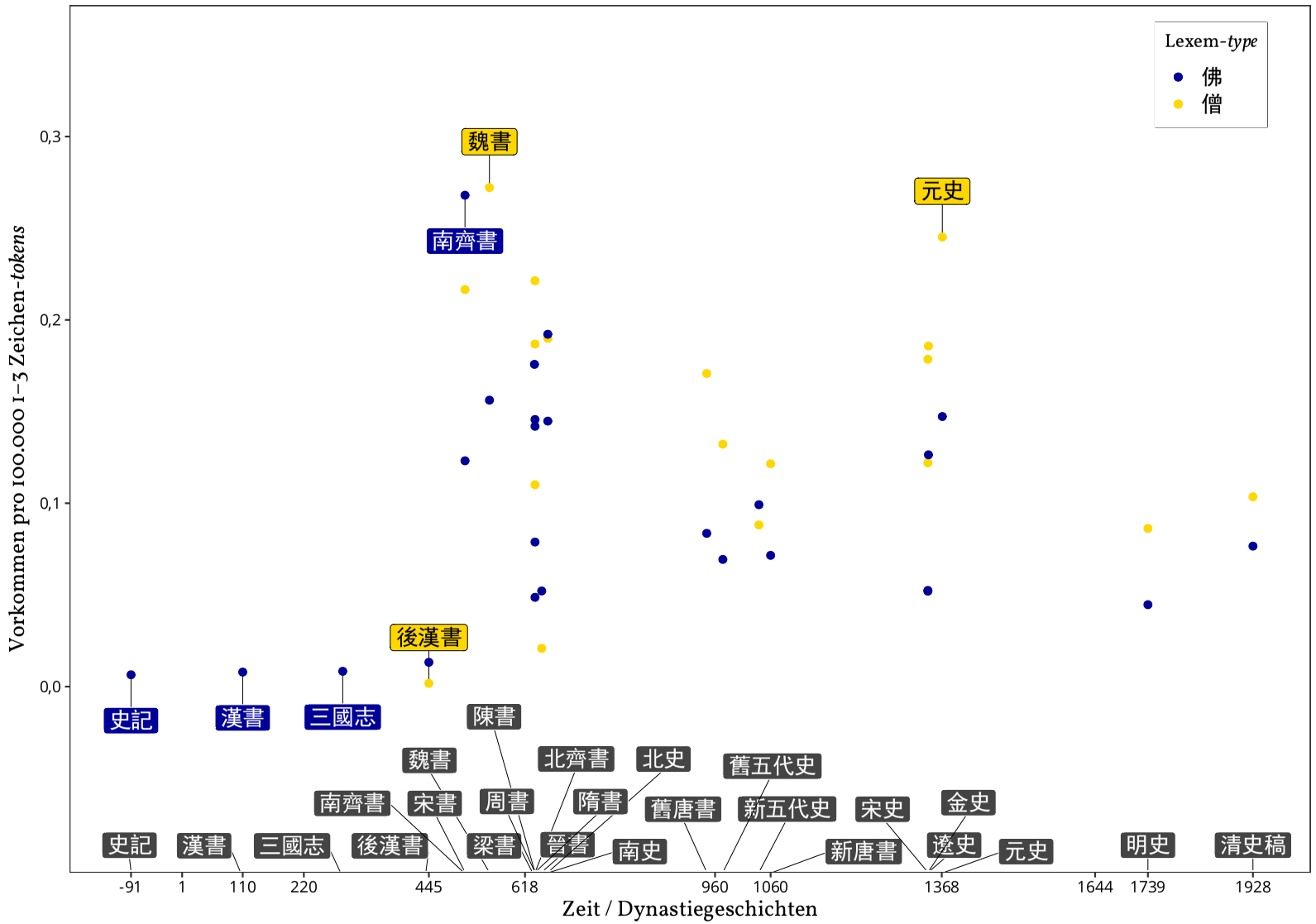


Abbildung 2.6 Vorkommen von *fo/bi/fu* 佛 und *seng* 僧 in den *zhengshi*-Texten

Die heute noch gebräuchliche, rein phonetische Entlehnung *niepan* 涅槃/涅槃 (*nirvāna*) ist im *Wei shu* 魏書 nachgewiesen.[142] Eine alternative Transkription, *miedu* 滅度, die seltener in den *zhengshi* Verwendung findet, wird im *Wei shu* ebenfalls erwähnt.[143]

Das Zeichen *seng* 僧 („Mönch", „Priester") scheint – im Gegensatz zu 佛 – tatsächlich erst etwa zeitgleich mit dem Auftreten des frühen Buddhismus in China entstanden zu sein, wobei ebenfalls eine phonetische Entlehnung aus dem Sanskrit stattgefunden hat.[144] Die älteste im *HYDCD* angegebene Belegstelle stammt aus dem *Wei shu* 魏書.[145] Das von Xu Shen 許慎 (ca. 55–ca. 149) im Jahr 100 fertiggestellte und 121 veröffentlichte *Shuo wen jie zi* 說文解字 enthält bereits eine Glosse für *seng*, die einen klaren Bezug zum Buddhismus herstellt: *„seng futu daoren ye* 僧浮屠道人也".[146] Aufgrund der Überlieferungsgeschichte des *Shuo wen jie zi* kann allerdings nicht ausgeschlossen werden, dass der Eintrag zu *seng* erst in einer späteren Bearbeitung hinzugefügt wurde.[147] Ebenfalls ab dem *Wei shu* findet man auch vermehrt Wortbildungen wie *sengni* 僧尼 („Mönche und Nonnen").[148]

Abb. 2.6 und 2.7 lassen anhand der genannten Beispiele das Aufkommen des Buddhismus während der Han-Zeit, sowie eine zunehmende Verbreitung zwischen der Han- und der Tang-Zeit erahnen. Die genannten Beispiele erinnern dabei an einige Herausforderungen und Limitationen, die bei sprachgeschichtlichen Untersuchungen des Chinesischen zu berücksichtigen sind. Eine quantitative Betrachtung des Textmaterials verschleiert den für die Datierung essentiellen Unterschied zwischen Haupttext und später verfassten Kommentaren.[149] Dazu gehören die Authentizität bzw. Glaubwürdigkeit der verwendeten Quellen, alternative Schreibungen von Zeichen bzw. Varianten (*yiti zi* 異體字) wie 仏/佛, sowie Zeichen wie 佛 mit unterschiedlichen Lesungen und Bedeutungen (Heteronyme, *duoyinzi* 多音字). Die Belege aus dem 1.–6. Jh. lassen einige derselben Wortbildungsmuster, und -strategien erkennen, die so für die moderne chinesische Hochsprache immer noch typisch sind – etwa phonetische

Beijing 北京: Zhonghua shuju 中華書局, S. 1853; das *HYDCD* belegt *foxiang* 佛像 mit dem 17 Jahre später vorgelegten *Wei shu* 魏書 *HYDCD*, Bd. 1, S. 1291; *fosi* mit dem nur 12 Jahre später fertiggestellten *Jin shu* 晉書 *HYDCD*, Bd. 1, S. 1286.

142 Diese Belegstelle ist auch im *HYDCD* angegeben, sowie der noch etwas frühere Text *Niepan wu ming lun* 涅槃無名論 (*Das Nirvana hat keinen Namen*) des buddhistischen Gelehrten Sengzhao 僧肇 (ca. 384–414). Siehe *HYDCD*, Bd. 5, S. 1210; zu Sengzhao siehe Chan Wing-Tsit 陳榮捷 1963: *A Source Book in Chinese Philosophy*. Princeton, New Jersey: Princeton University Press, S. 344; Die Autorschaft des Textes *Niepan wu ming lun* ist umstritten. Siehe Xu Wenming 徐文明 1999: „《*Niepan wu ming lun*》*zhen wei bian* 《涅盤無名論》真偽辨 (Die Authentizität des *Niepan wu ming lun*)". In: *Yuanguang foxue xuebao* 圓光佛學學報 *(Yuanguang Buddhist Journal)* 7.

143 Siehe Wei Shou 魏收 1974 [554]: *Wei shu* 魏書. *8 Bde.* Beijing 北京: Zhonghua shuju 中華書局, S. 3027; im *HYDCD* wird auch noch *nigen* 泥亘 als Bezeichnung für *nirvāna* genannt, die im *zhengshi*-Korpus aber nicht enthalten ist. Siehe *HYDCD*, Bd. 5, S. 1210.

144 Im *HYDCD* ist für *seng[jia]* 僧 [伽] das Sanskrit-Wort *saṃgha* संघ, „Versammlung" als Ursprung angegeben. Siehe *HYDCD*, Bd. 1, S. 1682–1683.

145 Siehe *HYDCD*, Bd. 1, S. 1682–1683. siehe auch Wei Shou 魏收 1974 [554], S. 3026 uvwm. Das *zhengshi*-Korpus enthält mit dem *HHS* eine noch frühere Belegstelle für *seng* 僧. Der betreffende Abschnitt ist allerdings mit *Yuanyi lede geshi* 遠夷樂德歌詩 („Songs of the Distant Barbarians delighted in the Virtue [of Han]") überschrieben. *HHS*, S. 2856; das mit chinesischen Zeichen wiedergegebene […] *yang luo seng lin* 陽雒僧鱗 […] ist eine Transliteration der entsprechenden Fremdsprache. Vgl. Rafe de Crespigny 2007: *A Biographical Dictionary of Later Han to the Three Kingdoms (23-220 AD)*. Handbook of Oriental Studies, Section Four: China 19. Leiden & Boston: Brill, S. 686–787.

146 „*Seng* bedeutet Buddhist oder Daoist." Die Glosse könnte auch als „buddhistischer *daoren*" gelesen werden, also wörtlich jm., der [tugendhaft] dem Weg des Buddhismus folgt. Xu Shen 許慎 1985 [121]: *Shuo wen jie zi* 說文解字. *2 Bde.* Beijing 北京: Zhonghua shuju 中華書局, S. 266, [Eintrag Nr. 5186].

147 Vgl. William G. Boltz 1993b: „Shuo wen chieh tzu 說文解字". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 429–443, S. 434–435.

148 Siehe u. a. Wei Shou 魏收 1974 [554], S. 3039–3044.

149 Im Material des *zhengshi* Korpus teilweise vorhandene Kommentare wurden daher entfernt und für Erstnennungen ein Abgleich mit der *Zhonghua shuju* 中華書局-Ausgabe vorgenommen.

Übertragungen (*yinyi* 音譯) wie *niepan* und *seng*[*jia*] 僧 [伽] und die Bildung neuer Komposita aus vorhandenen, teils entlehnten Wörtern (*banyin banyi* 半音半意) wie bei *foxiang* oder *fosi*.[150]
Als verhältnismäßig homogene Textreihe, die einen Zeitraum von 2.019 Jahren (bzw. einen historischen Zeitraum von etwa 3.000 Jahren) abdeckt, sind die *zhengshi* mit nur 25 Texten in teilweise großen zeitlichen Abständen als diachrones Korpus nur bedingt für die Analyse von Sprachwandel geeignet. Es lässt aber beispielhaft die Beobachtung eines langsamen Sprachwandels des schriftsprachlichen Stils (*guwen* 古文), sowie des Aufkommens neuer Konzepte und Begriffe und lexikalischer Veränderungen zu.

So kann nicht nur die Geschichte „through the lens of neologisms“,[151], sondern auch die Neologismen durch die Brille der Geschichtsschreibung gesehen werden. Wissen über Wortneu- und -umbildungen kann umgekehrt Indizien für die Datierung von Texten liefern, wenn phonologischer Wandel nicht greifbar und stilistische sowie syntaktische Veränderungen gering sind.

150 Vgl. auch Wilkinson 2000, S. 37.
151 Jing-Schmidt und Hsieh 2019, S. 515.

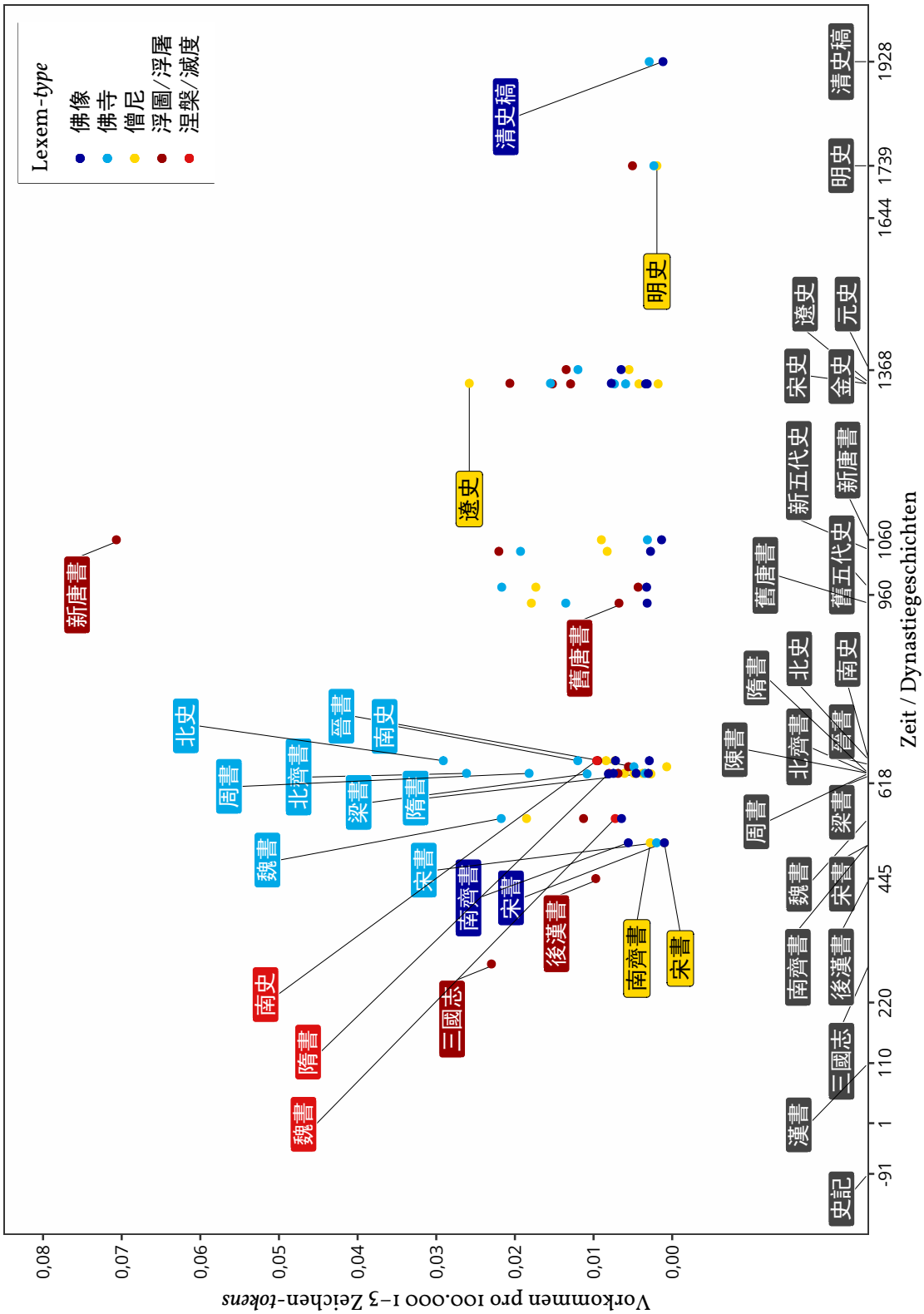


Abbildung 2.7 Vorkommen buddhistischer Lexeme in den *zhengshi*-Texten

3 Linguistische Datierung

„The complexity of linguistic dating is clear
and remains controversial in many disciplines."[1]
Gregory Toner

Die Datierung von Sprachen oder Dialekten steht in der wissenschaftlichen Literatur zur „linguistischen Datierung" als Teilbereich der historischen Linguistik oft im Vordergrund: Wann sind bestimmte Sprachen entstanden oder ausgestorben, wann könnten sich Sprachen von einer Proto-Sprache abgespalten haben?[2] Die Weiterentwicklung und Verbesserung der Datierung von *Sprachen* mit Methoden, die an Überlegungen des Sprachwissenschaftlers Morris Swadesh oder die Evolutionsbiologie angelehnt sind, ist nach wie vor ein wichtiger Schwerpunkt der historischen Sprachwissenschaft.[3] Auch für die sino-tibetische Sprachfamilie wurden unterschiedliche phylogenetische Untersuchungen über Zeit und Ort ihres Ursprungs vorgenommen.[4] Für die vorliegende Arbeit ist dies aber eher peripher von Interesse. Zwar basiert sie ebenfalls auf der Beobachtung des Wortschatzwandels, mit linguistischer Datierung ist hier jedoch primär die chronologische Einordnung einzelner *Texte* gemeint – eine auf vorkommende Lexeme und deren relative Häufigkeit, sowie erwähnte Namen und Ereignisse gestützte Schätzung, (ab) wann ein Text (frühestens) verfasst wurde.

Die Datierung von Texten ist seit langem ein Forschungsbereich von Philolog:innen, Theolog:innen und historischen Linguist:innen. Vor allem die Datierung biblischer Texte beschäftigt Wissenschaftler:innen seit Jahrhunderten. Man erhofft sich davon zusätzliche, objektivere Erkenntnisse für die inhaltliche Deutung der Texte. Avi Hurvitz erklärt, weshalb die Linguistik dafür besonders geeignet ist:

1 Gregory Toner und Han Xiwu 2019: *Language and Chronology – Text Dating by Machine Learning*. Language and Computers, Vol. 84. Leiden & Boston: Brill, S. 38.

2 Siehe dazu z. B. die Arbeit von Morris Swadesh 1955: „Towards Greater Accuracy in Lexicostatistic Dating". In: *International Journal of American Linguistics* 21.2, S. 121–137. URL: http://www.jstor.org/stable/1263939; Trotz seiner innovativen, statistischen Herangehensweise ist die auf Swadesh zurückgehende *Glottochronologie* heute umstritten, da er von der Richtigkeit traditioneller Sprachdatierungen ausgeht und annimmt, dass alle Sprachen eine gleichförmige, naturgegebene Veränderung des Wortschatzes durchlaufen. Vgl. z. B. Alinei 2004, S. 211–212; für eine zusammenfassende Analyse, ob und wie die im Rahmen der Glottochronologie untersuchten Sprachwandelprozesse auch oder stattdessen mithilfe „phylogenetischer", evolutionsbiologischer Methoden analysiert werden können, siehe Jyri Lehtinen 2009: „Language change as an evolutionary process". Masterarbeit. Helsinki: University of Helsinki, *passim*.

3 Vgl. z. B. Eric W. Holman et al. 2011: „Automated Dating of the World's Language Families Based on Lexical Similarity". In: *Current Anthropology* 52.6, S. 1–35. DOI: 10.1086/662127; George Starostin 2013: „Lexicostatistics as a basis for language classification: increasing the pros, reducing the cons". In: *Classification and Evolution in Biology, Linguistics and the History of Science*. Hrsg. von Heiner Fangerau et al. Stuttgart: Franz Steiner Verlag, S. 125–146; Taraka Rama 2014: *Vocabulary lists in computational historical linguistics*. Data linguistica 25. Göteborg: Språkbanken, Department of Swedish; Taraka Rama 2015: *Studies in computational historical linguistics*. Hrsg. von Lars Borin. Data linguistica 27. Göteborg: Språkbanken, Department of Swedish.

4 Siehe Laurent Sagart et al. 2019: „Dated language phylogenies shed light on the ancestry of Sino-Tibetan". In: *Proceedings of the National Academy of Sciences* 116.21, S. 10317–10322. DOI: 10.1073/pnas.1817972116; vgl. auch Zhang Menghan et al. 2019: „Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic". In: *Nature* 569.7754, S. 112–115. DOI: 10.1038/s41586-019-1153-z.

Veröffentlicht in: Tilman Schalmey: *Computerlinguistische Datierung schriftsprachlicher chinesischer Texte*.
Heidelberg: HASP, 2023. DOI: https://doi.org/10.11588/hasp.1153

> The possibility of dating Biblical texts has always held great fascination for scholars. There is the feeling that if we were certain when a particular text has been written, we would have an additional clue to both its meaning and its significance. Unfortunately, the *theological, historical and literary* criteria which have been used for establishing the date of chronologically problematic texts are very often subjective. *Linguistic* studies likewise did not produce satisfactory results, since they were not usually based upon methodologically reliable criteria. However, we believe that it is this linguistic aspect which should be primarily studied in order to gain objective criteria for solving chronological issues. The particular contribution of the linguistic discipline stems from the fact that the solution of exegetical and theological questions involved in Higher Criticism simply does not affect its procedures; hence the considerably objective results it is likely to provide.[5]

Die für die Bibelforschung relevanten Aspekte der Datierung sind für die Sinologie gleichermaßen von Bedeutung, gerade wenn es um die Entstehungszeit klassischer philosophischer und kanonischer Texte bzw. Textabschnitte geht.[6] So wird z. B. für *Shijing* 詩經 und *Shangshu* 尚書 die Datierung einzelner Teile intensiv diskutiert.[7] Das *Shangshu* bzw. *Shujing* 書經, übersetzt als *Buch der Urkunden* bzw. *Buch der Dokumente* ist eine Sammlung vor allem von Aufzeichnungen ritualisierter Reden, entstanden vermeintlich über einen Zeitraum von etwa 2500–500 v. u. Z.[8] Die heute vorliegende Fassung lässt sich in sogenannte Alt- (*guwen* 古文) und Neutext-Kapitel (*jinwen* 今文) unterteilen. Letztere erhielten ihre heutige Form im 2. Jh. v. u. Z., die Alttext-Kapitel wurden mit großer Sicherheit erst während der der Jin-Zeit (晉, 265–420) in der heutigen Fassung verschriftlicht und ergänzt bzw. gefälscht. Die Datierung einzelner Kapitel und Abschnitte, sowie die Authentizität vor allem der *guwen*-Kapitel wurden spätestens ab dem 17. Jh. intensiv diskutiert.[9]

Ähnliches gilt für die Entstehung und Autorschaft der beiden daoistischen Klassiker *Laozi* 老子 und *Zhuangzi* 莊子, die beide für den daoistischen Kanon von großer Bedeutung sind.[10]

Harbsmeier erweitert die Fragestellung der Datierung am Beispiel des *Lunyu* 論語 um zusätzliche Aspekte: den Zeitpunkt der Kompilation, die Frage danach, wann ein kompiliertes Werk seinen Titel erhalten hat und die Frage, wann es unter diesem erstmals zitiert wurde.[11]

5 Avi Hurvitz 1973: „Linguistic Criteria for Dating Problematic Biblical Texts". In: *Hebrew Abstracts* 14, S. 74–79, S. 74; zitiert in Young und Rezetko 2014, S. 16; Während Hurvitz in den 1970er Jahren die Ergebnisse der Linguistik für nicht zufriedenstellend hielt, hat er dennoch selbst etliche überzeugende Arbeiten in diesem Bereich verfasst. Siehe Young und Rezetko 2014, S. 17–23; Hurvitz verwendet zudem Neologismen und sogar Archaismen als Indikator für die Datierung hebräischer Bibeltexte. Vgl. z. B. Young und Rezetko 2014, S. 19–20.

6 Vgl. auch Michael Nylan 2001: *The Five 'Confucian' Classics*. New Haven: Yale University Press, Nylan beschreibt die Stellung des konfuzianischen Kanons in Ostasien als „roughly analogous to that of the Bible in the West" (S. 2).

7 Siehe z. B. ebd., S. 77–88 zur Unterteilung und Entstehung des *Shijing*, S. 127–136 für einen Überblick zum *Shangshu*.

8 Siehe ebd., S. 121.

9 Siehe ebd., v. a. S. 128–135. Sowohl Neu- als auch Alttext-Kapitel enthalten deutlich älteres Textmaterial. Michael Nylan verdeutlicht die Komplexität der Frage nach der Datierung, indem sie Fragen danach aufwirft, was eigentlich tatsächlich datiert werden soll: „What kind of date is most meaningful in the study of a given chapter: the dates when individual passages were composed? the date when most or all of the chapter was compiled, barring later interpolations? the date when the entire chapter was written down as a unit? or the date when the chapter, in part or in whole, was inserted into the *Documents* collection?" (S. 132). Sie schlägt eine erweiterte Klassifizierung einiger Abschnitte in vier Gruppen vor, die unter anderem auf Grammatik und Wortschatz basiert.

10 Siehe Tao Hongyin 2015: „Author Identification and Dating of Texts". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill, „While some consider Lǎozǐ to be a senior contemporary of Confucius (or Kǒng Qiū 孔丘, 551–479 BCE), living in the early to middle Zhànguó 戰國 or Warring States period (trad. 475–221 BCE) and before Zhuāngzǐ 莊子, who is supposed to have lived in the late Zhànguó period but whose dating itself is subject to various speculations, others consider Lǎozǐ to be after Zhuāngzǐ. […]"; Siehe auch Liu Xiaogan 劉笑敢 2005: *Laozi niandai xin kao yu sixiang xin quan* 老子年代新考與思想新詮 *[Neue Untersuchungen über Laozis Zeit und neue Interpretationen seiner Philosophie]*. Taipei 台北: Dongda tushu 東大圖書.

11 Siehe Christoph Harbsmeier 2019: „The Authenticity and Nature of the *Analects* of Confucius". In: *Journal of Chinese Studies* 68, S. 171–233, S. 189.

Er ist überzeugt, dass „dating […] must! rely on well-understood and well-described linguistic changes […].“[12] Intensive Bemühungen um linguistische Argumente und die Debatte, von der traditionellen Einordnung des *Lunyu* als direkte Aufzeichnungen von Konfuzius' (ca. 551–479 v. u. Z.) Schülern zugunsten einer Datierung in die westliche Han-Zeit (*Xi Han* 西漢, 202–9 v. u. Z.) Abstand zu nehmen, haben gezeigt, dass sich mit einer rein linguistischen Perspektive nicht immer unwiderlegbare Fakten schaffen lassen.[13] Dies ist vor allem dann der Fall, wenn ein Text keine sprachlichen Phänomene aufweist, die nachgewiesenermaßen erst zur Zeit der Textentstehung aufgekommen sind, denn „the absence of any linguistic phenomenon in a book as short as the *Analects* by itself proves precious little, except that it demonstrates clearly that *Lunyu* is not given away as a Han dynasty work by the language alone.“[14] Es bleibt zu bedenken, dass die genannten Texte vollkommen ungeachtet unserer heutigen, eurozentrischen Auffassung von Autorschaft über einen Zeitraum von mehreren hundert Jahren kompiliert und bei ihrer Tradierung bis zum Erreichen ihrer heutigen Form immer wieder verändert oder ergänzt wurden.[15] William Boltz argumentiert, dass die so geprägte „composite structure“ in der vorkaiserlichen Zeit eher die Regel war, als die Ausnahme.[16] Die Frage nach der Datierung im Sinne der Verfasserschaft sollte sich dann eher auf einzelne Abschnitte beschränken.

Dass das Erscheinen neuer Begriffe bzw. Wörter entscheidende Hinweise für die Textdatierung liefern kann, wurde bereits in Kapitel 2.3 angedeutet.[17] Dass Anachronismen als Fehler in Fälschungen in der forensischen Linguistik genutzt werden können, wird von Bentham (1825) beschrieben:

> The falsehood of a writing will often be detected, by its making direct mention of […] some fact posterior to the date which it bears. […] The mention of posterior facts; – first indication of forgery. […] In a living language there are always variations in words, in the meaning of words, […], in the manner of spelling, which may detect the age of a writing, and lead to legitimate suspicions of forgery. […] The use of words not used till after the date of the writing; – second indication of forgery.[18]

Ein frühes Beispiel für die Anwendung dieser Methodik ist die Dekonstruktion der *Konstantinischen Schenkung*[19] durch Lorenzo Valla, der neben logischen und historischen Argumenten

12 Ebd., S. 207.

13 Eine ausführliche Darlegung unterschiedlicher aktueller Standpunkte dieser Debatte findet sich in Michael Hunter und Martin Kern, Hrsg. 2018: *Confucius and the Analects Revisited: New Perspectives on Composition, Dating, and Authorship*. Leiden & Boston: Brill; siehe insbesondere auch Martin Kern 2018: „Kongzi as Author in the Han“. In: *Confucius and the Analects Revisited: New Perspectives on Composition, Dating, and Authorship*. Hrsg. von Michael Hunter und Martin Kern. Leiden & Boston: Brill, S. 268–307, S. 270; sowie Wolfgang Behr 2011: *The Analects: A Western Han Text?* Conference Presentation: The Lunyu as a Han Text, Princeton University. doi: https://doi.org/10.5281/zenodo.1405049, Behr wünscht sich allen Widrigkeiten zum Trotz allerdings den Versuch einer computerlinguistischen Herangehensweise.

14 Harbsmeier 2019, S. 207.

15 Toner beschreibt eine vergleichbare Problematik auch für irische Texte, die vor dem 15. Jahrhundert entstanden sind. Siehe Toner und Han Xiwu 2019, S. 11–12. Er spricht in diesem Kontext auch von *linguistic strata*, sprachlichen Schichten. Siehe S. 23.

16 Siehe William G. Boltz 2007: „The Composite Nature of Early Chinese Texts“. In: *Text and Ritual in Early China*. Hrsg. von Martin Kern. Seattle und London: Washington University Press, S. 50–78, S. 52; Zum Verhältnis von Autorschaft, Herausgeberschaft und Kompilation und den zugehörigen Begrifflichkeiten in der klassischen Texttradition siehe auch Li Wai-Yee 李惠儀 2017: „Concepts of Authorship“. In: *Oxford Handbook of Classical Chinese Literature (1000 BCE–900CE)*. Hrsg. von Li Wai-Yee 李惠仪 Wiebke Denecke und Tian Xiaofei 田曉菲. New York: Oxford University Press, S. 360–376, v. a. S. 360–363.

17 Siehe ab S. 20.

18 Jeremy Bentham 1825: *A Treatise on Judicial Evidence*. Hrsg. von Étienne Dumont. London: Baldwin, Cradock und Joy, S. 140.

19 Die „Konstantinische Schenkung“ ist eine gefälschte Urkunde, die belegen sollte, dass Konstantin I. das weströmische Reich Papst Silvester I. per Schenkung übertragen haben soll, um entsprechende Gebietsansprüche geltend

dafür, dass der Text eine Fälschung sein muss,[20] auch linguistische Argumente wie fälschlich gebrauchte Sprachregister, Bezeichnungen und weitere Anachronismen anführt.[21]

Die Fragen nach der Datierung eines Textes und seiner Echtheit sind also eng miteinander verbunden.[22] Fälschungen von Texten können mit den unterschiedlichsten Absichten geschaffen werden, etwa zur Manipulation der Leser:innen durch Verbreitung von Falschinformationen. Durch nachträgliches Verfassen vermeintlich historischer Texte kann Geschichte umgeschrieben werden, z. B. um die Auslegung historischer Ereignisse durch eine herrschende Partei zu stärken, oder einer gegnerischen Partei mittels Desinformationskampagnen zu schaden.[23]

Textfälschungen in Form von Imitationen können aber auch ein Ausdruck von Bewunderung sein. In der chinesischen Textkultur besteht eine lange Tradition von Fälschungen und Imitationen unterschiedlichster Couleur, wobei das Fälschen von Bildern, Kalligraphien oder Texten nicht unbedingt nur negativ konnotiert war. Ihre Existenz konnte für die ursprünglichen Künstler:innen oder Autor:innen als „Beweis für die hohe Qualität der eigenen Werke und für die hohe Wertschätzung, die andere ihnen entgegenbringen"[24] gesehen werden. Zudem Dabei ist teilweise auch eine Bewunderung für die Belesenheit der Fälscher:innen zu spüren:

> Anyone who forges antiquities and passes them off must be fully conversant with antiquity. If someone not conversant with antiquity passes on [a forgery], how could that be the forger's fault.[25]

Eine ähnliche Einstellung schimmert auch bei Lorenzo Valla durch, der sich über den Fälscher der *Konstantinischen Schenkung* echauffiert: „[…] but I am foolish to attack that man's brazenness rather than the madness of those who have believed him."[26]

Fälschungen zogen aber auch die Kritik wichtiger Gelehrter wie Zhu Xi 朱熹 (1130–1200) auf sich,[27] was keineswegs ein langfristiges Umdenken zur Folge hatte. Rusk bezeichnet in diesem Kontext das letzte Drittel der Ming 明-Zeit (1368–1644) als „heyday of textual forgery in the imperial period".[28]

Davon abgesehen war vor der Erfindung moderner Vervielfältigungstechniken das Abschreiben von Texten, ebenso wie das Kopieren von Kunstwerken, sowieso eine Notwendigkeit für die

zu machen. Siehe z. B. Michail A. Bojcov 2015: „Die Konstantinische Schenkung und ähnliche Gaben – im Westen und im Osten Europas". In: *Jahrbücher für Geschichte Osteuropas* 63.1, S. 23–46. URL: http://www.jstor.org/stable/43819721.

20 Siehe Lorenzo Valla 2007 [1440]: *On the Donation of Constantine*. Übers. von Glen W. Bowersock. The I Tatti Renaissance Library. Cambridge & London: Harvard University Press, S. 11–57.

21 Siehe ebd., v. a. S. 65–107. Valla spricht von „linguistic barbarisms" („barbariem sermonis", Übs. von Glen Bowersock).

22 Siehe z. B. auch Dieter Wickmann 1989: „Computergestützte Philologie: Bestimmung der Echtheit und Datierung von Texten / Computer-Aided Philology: Authorship and Chronological Determination". In: *Computational Linguistics, An International Handbook of Computer Oriented Language Research and Applications*. Hrsg. von István S. Bátori, Winfried Lenders und Wolfgang Putschke. Handbücher zur Sprach- und Kommunikationswissenschaft, Band 4. Berlin & New York: De Gruyter, S. 528–534, v. a. S. 528, S. 533.

23 Bekannte Beispiele dafür sind die bereits genannte „Konstantinische Schenkung", sowie die sogenannten „Protokolle der Weisen von Zion", die ab Anfang des 20. Jhs. als antisemitische Propaganda verbreitet wurden. Weiterführend dazu siehe z. B. Wolfgang Benz 2019 [2007]: *Die Protokolle der Weisen von Zion: Die Legende der jüdischen Weltverschwörung*. 4. Aufl. München: C. H. Beck, bzw. Bojcov 2015.

24 Lena Henningsen 2010: *Copyright Matters: Imitation, Creativity and Authenticity in Contemporary Chinese Literature*. Berlin: BWV, S. 91, übersetzt durch den Verfasser. siehe auch William P. Alford 1995: *To Steal a Book Is an Elegant Offense*. Stanford: Stanford University Press, S. 29.

25 Wang Shizhen 王世貞 (1526–1590), zitiert Henningsen 2010, S. 41.

26 Valla 2007 [1440], S. 63.

27 Siehe Bruce Rusk 2006: „Not Written in Stone: Ming Readers of the *Great Learning* and the Impact of Forgery". In: *Harvard Journal of Asiatic Studies* 66.1, S. 189–231. DOI: 10.2307/25066803, S. 196–197.

28 Ebd., S. 189.

Verbreitung solcher Kulturgüter.[29] Dabei besteht ein bedeutender Unterschied zwischen Kopien, Abschriften und später Nachdrucken, die – sogar unabhängig von legalen Fragen – eine Angabe des Urhebers machen, zu Plagiaten, bei denen der ursprüngliche Verfasser nicht genannt wird.[30]

Zusätzlich zur Analyse sprachlicher Eigenschaften eines Texts kann sich die traditionelle Datierung auf andere direkte Indizien im Text stützen. Dazu gehören Erwähnungen von Ereignissen oder Ortsnamen, sowie Referenzen auf historische Persönlichkeiten.[31]

Ein weiterer inhaltlicher Aspekt, der in der Datierung von Texten genutzt werden kann ist derjenige der Intertextualität, der „Präsenz eines Textes in einem anderen"[32] – im Optimalfall für die Textdatierung wörtliche Zitate aus anderen Texten. Dies folgt der logischen Idee, dass ein Text *B*, in welchem Text *A* zitiert wird, neuer sein muss als *A*. Jedoch sollte dabei die Möglichkeit nicht ausgeschlossen werden, dass beide Texte, *A* und *B*, einen noch älteren Text, *C*, zitieren.[33] Mit exakt dieser Methodik konnte bereits im 18. Jh. gezeigt werden, dass das lange fälschlich dem Han-zeitlichen Gelehrten Ma Rong 馬融 (79–166) zugeschriebene *Zhongjing* 忠經 (*Klassiker der Loyalität*) deutlich später entstanden sein muss, da es Zitate aus dem bereits erwähnten Alttext-*Shangshu* enthält.[34]

Ding Yan 丁晏 (1794-1875) kommt in seiner Untersuchung *Shangshu yulun* 尚書餘論 (*Epilog zum Shangshu*) zur gleichen Schlussfolgerung und argumentiert überdies, dass die vermeintliche Tabuisierung vor allem der Zeichen *min* 民 und *zhi* 治 für eine Datierung in die Tang-Zeit (唐, 618–960) spricht.[35] Suwald legt zwar überzeugend dar, weshalb Dings Argumentation hier nicht stichhaltig ist,[36] grundsätzlich stellt die Tabuisierung von Zeichen aber ein wichtiges Standbein für die Datierung chinesischsprachiger Texte dar.[37] Aus Tabus lassen sich jedoch vor allem Schlüsse auf die Entstehungszeit einer bestimmten Ausgabe ziehen, denn Herausgeber:innen können Texte geltenden Tabus anpassen, oder Tabus früherer Ausgaben zugunsten der Lesbarkeit bzw. Verständlichkeit auflösen.

Im Folgenden wird primär auf den Forschungsstand zur *computerlinguistischen Datierung von Texten* eingegangen, deren Ansätze sich jedoch fast ausschließlich auf westliche Sprachen beziehen. Die Datierung wird dabei, ähnlich wie in verwandten Forschungsfeldern der *Digital Humanities*, wie die Zuordnung von Autor:innen (*authorship attribution*), die Identifizierung von Genres oder dem *Topic modelling*, in der Regel als Kategorisierungsproblem betrachtet. Dabei werden für ein

29 Henningsen 2010, S. 35–36.

30 Eine ausführliche Diskussion der Unterschiede zwischen Imitat, Plagiat und Kopie findet sich in ebd., S. 25–33.

31 Siehe Bentham 1825, S. 140; vgl. auch Toner und Han Xiwu 2019, S. 13–15. Auf diese Art der inhaltlichen *temporal cues* wird in Kapitel 4.7, ab S. 97, 4.8, ab S. 103 und 3, ab S. 37 eingegangen.

32 „, […] la présence effective d'un texte dans un autre." Gérard Genette 1982: *Palimpsestes: La littérature au second degré*. Paris: Éditions du Seuil, S. 8; auf Deutsch zitiert nach Wolfgang Hallet 2006: „Intertextualität als methodisches Konzept einer kulturwissenschaftlichen Literaturwissenschaft". In: *Kulturelles Wissen und Intertextualität. Theoriekonzeptionen und Fallstudien zur Kontextualisierung von Literatur*. Hrsg. von Marion Gymnich, Birgit Neumann und Ansgar Nünning. Trier: Wissenschaftlicher Verlag Trier, S. 53–70, S. 55.

33 Siehe Toner und Han Xiwu 2019, S. 18; Toner bezieht sich hier auf die Methodik, die Rudolf Thurneysen 1921 teilweise für die Datierung irischer Helden- und Königssagen anwendet. Siehe dazu Rudolf Thurneysen 1921: *Die irische Helden- und Königssage bis zum siebzehnten Jahrhundert*. 2 Bde. Halle: Max Niemeyer, z. B. S. 45.

34 Der Gelehrte Hui Dong 惠棟 (1697–1758) argumentiert in seinem *Gu jin Shangshu kao zhu* 古今尚書考注 (*Untersuchung und Kommentar von Alt- und Neutext*-Shangshu), dass der han-zeitliche Ma Rong der „falsche Autor" sein muss. Siehe Judith Suwald 2008: „Zhong 忠 und das Zhongjing 忠經". Diss. München: LMU München, S. 71.

35 Siehe ebd., S. 68–69. Die Eigennamen des zweiten (Taizong 太宗, reg. 626–649) und dritten Kaisers der Tang (Gaozong 高宗, reg. 649–683), Li Zhi 李治 und Li Shimin 李世民, enthalten diese Zeichen und sollten daher in der Tang-Zeit tabuisiert werden. Suwald gibt die Regierungszeit von Gaozong mit derjenigen von Kaiser Gaozu 高祖 an: 618–626.

36 Siehe ebd., S. 68–69. Einerseits war *shimin* 世民 nur als Zeichenfolge tabuisiert, andererseits kommt das Zeichen *min* 民 in den Ausgaben des *Zhongjing*, die vermutlich auch Ding vorgelegen haben müssen, ebenfalls vor.

37 In Kapitel 4.3, ab S. 72 wird ausführlicher auf Zeichentabus eingegangen.

passendes Trainingskorpus Kategorien vordefiniert oder ermittelt, z. B. unterschiedliche Textgattungen, verschiedene Autoren oder inhaltliche Themen und beobachtet, wie sich Texte dieses Korpus anhand sprachlicher Merkmale entsprechend zuordnen lassen. Im Kontext der Datierung von Texten unabhängig von ihrer Autorschaft werden dafür oft statistische Sprachmodelle (*Statistical Language Models*, *SLM*) eingesetzt. Sie basieren zumeist auf Worthäufigkeiten und sind auf die Existenz umfangreicher diachroner Trainingskorpora angewiesen. Dasselbe gilt für Arbeiten, die auf Methoden aus dem Bereich des *machine learning* zurückgreifen.

3.1 Computerlinguistische Datierung von Texten

Die Aufgabe, digitale bzw. digitalisierte Texte zu datieren, ist ein Forschungsbereich der Computerlinguistik, zu dem bereits zahlreiche interessante Arbeiten veröffentlicht wurden. Das Spektrum reicht dabei von einer Interpretation vorhandener Metadaten zu den Texten,[38] bis hin zur Konstruktion komplexer statistischer Sprachmodelle und der Entwicklung neuer Methoden und Konzepte.[39] Die teils uneinheitliche Terminologie bisheriger Veröffentlichungen spiegelt sich schon in der Vielzahl der Bezeichnungen wider, die neben dem sowieso mehrdeutigen Begriff *Dating* für teils sehr ähnliche Herangehensweisen zur Textdatierung verwendet werden und so den Zugang zur relevanten Literatur erschweren: „determining time of non-timestamped documents“,[40] „automatically determining publication dates“,[41] „temporal text analysis“,[42] „labeling with timestamps“,[43] „temporal resolution of texts“,[44] „publication date estimation“,[45] „temporal classification of text“,[46] „estimating the date of first publication“, „publication date prediction“, „text-based composition dating“,[47] „guess the publication year of a text“,[48] „predict year of authorship“[49] usw.

Fast die gesamte bestehende computerlinguistische Forschung zur Textdatierung behandelt dabei ausschließlich in modernen, westlichen Sprachen verfasste Texte, die in Alphabetschriften wiedergegeben werden. Relevante Ausnahmen bilden die Arbeiten von Yamada Takahito 山田

38 Siehe z. B. Bamman et al. 2017, S. 4.

39 Siehe z. B. de Jong, Rode und Hiemstra 2005; Nattiya Kanhabua und Kjetil Nørvåg 2008: „Improving Temporal Language Models for Determining Time of Non-timestamped Documents“. In: *Research and Advanced Technology for Digital Libraries: 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Proceedings*. Hrsg. von Birte Christensen-Dalsgaard et al. Berlin & Heidelberg: Springer, S. 358–370. doi: 10.1007/978-3-540-87599-4_37.

40 Kanhabua und Nørvåg 2008.

41 Garcia-Fernandez et al. 2011.

42 Abhimanu Kumar et al. 2012: „Dating Texts without Explicit Temporal Cues“. In: *arXiv [cs. CL]* 1211.2290, S. 1–12, S. 3.

43 Nathanael Chambers 2012: „Labeling Documents with Timestamps: Learning from their Time Expressions: Learning from their Time Expressions“. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju 濟州, Republic of Korea, 8–14 July 2012*, S. 98–106.

44 Abhimanu Kumar 2013: „Supervised Language Models for Temporal Resolution of Text in Absence of Explicit Temporal Cues“. Diss. Austin: University of Texas.

45 Li Yuanpeng et al. 2015: „Publication Date Estimation for Printed Historical Documents Using Convolutional Neural Networks“. In: *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*. HIP '15. Gammarth, Tunisia: ACM, S. 99–106. doi: 10.1145/2809544.2809550.

46 Guo Siyuan et al. 2015: „Toward Enhanced Metadata Quality of Large-Scale Digital Libraries: Estimating Volume Time Range“. In: *iConference 2015 Proceedings*. Urbana: iSchools / University of Illinois. url: http://hdl.handle.net/2142/73656, S. 1; Marcos Zampieri, Shervin Malmasi und Mark Dras 2016: „Modeling Language Change in Historical Corpora: The Case of Portuguese“. In: *ArXiv* abs/1610.00030, S. 1.

47 Bamman et al. 2017.

48 Graliński et al. 2017.

49 Vivek Kulkarni et al. 2018: „Simple Neologism Based Domain Independent Models to Predict Year of Authorship“. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, S. 202–212. url: https://www.aclweb.org/anthology/C18-1017.

崇仁 (2004),[50] sowie Yu Xuejin und Wei Huangfu (2019).[51] Yamada versucht, mit einer der *PCA* mathematisch sehr ähnlichen ***k-means*** Cluster-Analyse,[52] sich dem Problem einer ungefähren Datierung klassischer chinesischer Texte anzunähern. Zu diesem Zweck bildet er zwei Cluster (Gruppen) mit Texten, die vermeintlich jeweils aus dem 3. und 4. Jh. v. u. Z. stammen und untersucht dann, mit welcher der beiden Gruppen der Text *Sunzi* 孫子[53] ***clustert***.[54] Er muss jedoch konstatieren, dass „[…] die Ergebnisse der Cluster-Analyse nicht einfach akzeptiert"[55] werden können. Yu Xuejin und Wei Huangfu verwenden ein *deep learning* Modell, um einige klassische chinesische Texte in drei temporale Kategorien zu klassifizieren. Da Abschnitte derselben Texte in den Test- und Trainingsdaten enthalten sind, kann dabei eine sehr hohe Genauigkeit erzielt werden.[56]

Arbeiten aus dem angrenzenden Bereich der Stilometrie verschreiben sich häufiger Aspekten der Autorschaft oder des Genre, die Altersschätzung von Texten ist hier aber ebenfalls von Bedeutung. Unter anderem das Interesse an der Frage nach der Reihenfolge, in der die Stücke von William Shakespeare (1564–1616) verfasst wurden, motivierte bereits Ende des 19. Jahrhunderts zu quantitativen Analysen mit dem Ziel, seine Stücke so weit wie möglich in die Reihenfolge zu bringen, in der er sie geschrieben hat.[57] Madhukar Yardi (1946) greift die Ergebnisse dieser frühen quantitativen Studien auf und kann die Datierung einiger Stücke mittels einer Regressionsanalyse anhand einzelner stilistischer Merkmale eingrenzen.[58] Für Untersuchungen mit dem Kernanliegen, eine (relative) zeitliche Reihenfolge für das Werk einer Autorin oder eines Autors zu etablieren wurde von Richard Forsyth (1999) der Begriff *Stylochronometry* eingeführt.[59] Dabei wird angenommen, dass „certain aspects of an author's writing style evolve rectilinearly over the course of an author's life time."[60] Es werden stilistische Merkmale herausgearbeitet, anhand derer signifikante Unterschiede vom Früh- zum Spätwerk einer Autor:in festgestellt werden können, wofür z. B. eine *Principal Component Analysis* (*PCA*) einge-

50 Yamada Takahito 山田崇仁 2004.

51 Yu Xuejin und Wei Huangfu 2019.

52 Vgl. Chris Ding und He Xiaofeng 2004: „*K*-means Clustering via Principal Component Analysis". In: *Proceedings Of International Conference of Machine Learning (ICML 2004)*. Hrsg. von Russ Greiner und Dale Schuurmans. New York: ACM Press, S. 225–232.

53 *Sunzi bingfa* 孫子兵法, 13 *juan* 卷, in westlichen Sprachen auch bekannt als „Die Kunst des Krieges".

54 Siehe Yamada Takahito 山田崇仁 2004, *Sunzi clustert* eher mit den Texten aus dem 3. Jh. v. u. Z., deren Zuordnung, wie z. B. beim *Zhuangzi* 莊子, aber selbst strittig ist. Dies wiederum soll als Diskussionsgrundlage gelten, *Sunzi* eher dem 3., als dem 4. Jh. v. u. Z. zuzurechnen.

55 „[…]単純にはクラスター分析の結果を受け入れる事は出来ない。" ebd.

56 Die Autoren verwenden ein *Long short-term memory* (*LSTM*) Netzwerk und berichten von einer *Precision* von etwa 95 %. Da die Testdaten Abschnitte aus denselben Texten sind, die auch den Trainingsdatensatz bilden, ist das allerdings wenig überraschend. Yu und Wei müssen feststellen, dass „if the ancient books are not involved in the training set, the correct rate of the paragraphs of the ancient books will be reduced." Trotzdem sind sie überzeugt, dass „the proposed model offers an effective method on how to date the ancient Chinese texts." Siehe Yu Xuejin und Wei Huangfu 2019, S. 119. Die Aussagekraft dieser Untersuchung ist durch die gewählte Herangehensweise, sowie durch die geringe Anzahl an temporalen Kategorien und untersuchten Texten eingeschränkt.

57 Siehe Frederick J. Furnivall 1874: „Inaugural address to the New Shakspere Society". In: *The New Shakspere Society's Transactions* 1.1–2, S. v–vi, S. vi; zitiert in Murphy 2003, S. 209.

58 Siehe Madhukar R. Yardi 1946: „A Statistical Approach to the Problem of Chronology of Shakespeare's Plays". In: *Sankhyā: The Indian Journal of Statistics* 7.3, S. 263–268, S. 263–264. Yardi greift auf Daten der Shakespeare-Forscher Frederick G. Fleay und Edmund K. Chambers zurück, unter anderem zu Varianz bei Betonungen und Pausen. Eine spätere Studie dazu ist Barron Brainerd 1980: „The Chronology of Shakespeare's Plays: A Statistical Study". In: *Computers and the Humanities* 14, S. 221–230.

59 Constantina Stamou 2007: „Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating". In: *Literary and Linguistic Computing* 23.2, S. 181–199. DOI: 10.1093/llc/fqm029, S. 181; vgl. auch Richard Forsyth 1999: „Stylochronometry with substrings, or: A poet young and old". In: *Literary and Linguistic Computing* 14. DOI: 10.1093/llc/14.4.467.

60 Stamou 2007, S. 181.

setzt werden kann.[61] Dabei werden die betrachteten Dimensionen iterativ so lange auf zwei Hauptkomponenten reduziert, bis die wesentlichsten Unterschiede zwischen *n* Eigenschaften der untersuchten Objekte, hier den Worthäufigkeiten der untersuchten Texte, zweidimensional dargestellt werden können.[62] Die *Stylochronometry* beschäftigt sich vor allem mit der Datierung von Werken einzelner Autor:innen,[63] durch den Vergleich mit einem Hintergrundkorpus aus Werken anderer, zeitgenössischer Autor:innen kann aber sichergestellt werden, dass die gefundenen Merkmale tatsächlich Aussagen über den Stil der jeweiligen Autor:in zulassen und es nicht eher allgemeinere Trends sind, die zu der festgestellten sprachlichen Varianz führen.[64]

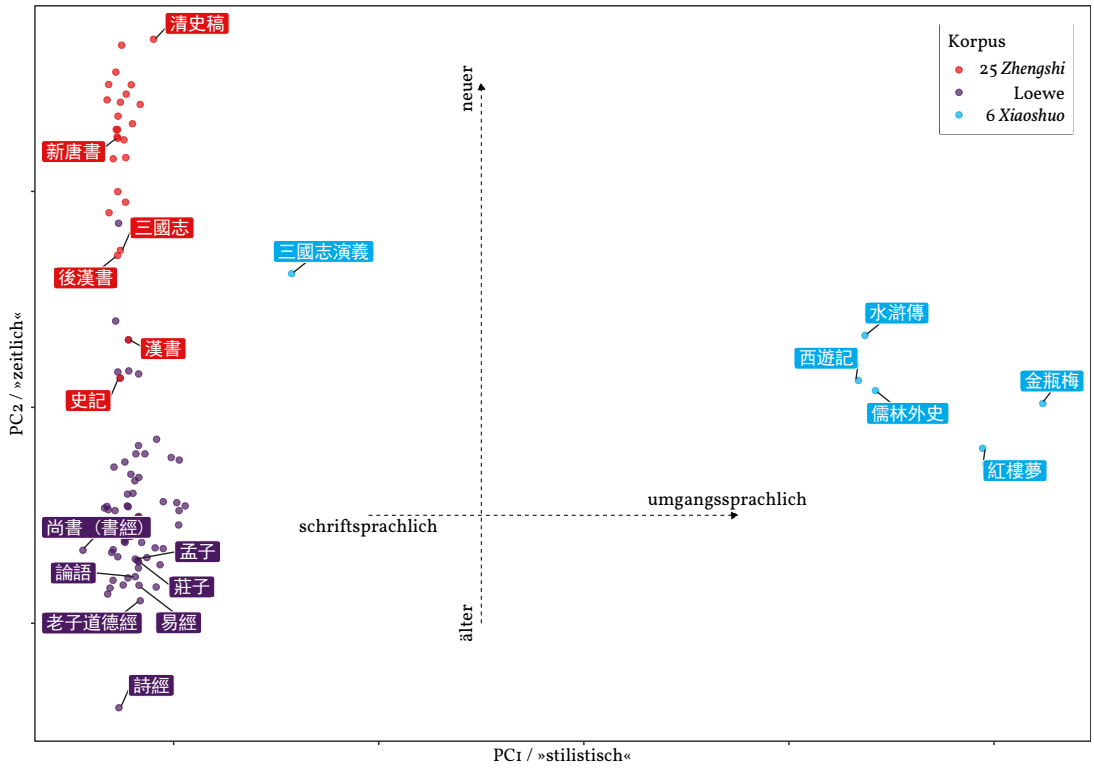


Abbildung 3.1 *PCA*, 1.000 häufigste 1–4 Zeichen Lexeme — *zhengshi* 正史, Loewe und *Xiaoshuo* 小說[65]

61 Vgl. z. B. auch Dries van Hulle und Mike Kestemont 2016: „Periodizing Samuel Beckett's Works: A Stylochronometric Approach". In: *Style* 50.2, S. 172–202, S. 182–186.

62 Siehe z. B. Jose Nilo G. Binongo und M. W. A. Smith 1999: „The Application of Principal Component Analysis to Stylometry". In: *Literary and Linguistic Computing* 14.4, S. 445–465, S. 447.

63 Siehe Stamou 2007, v. a. S. 181–191.

64 Siehe Carmen Klaussner und Carl Vogel 2015: „Stylochronometry: Timeline Prediction in Stylometric Analysis. Proceedings of AI-2015, The Thirty-Fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence". In: *Research and Development in Intelligent Systems XXXII*. Hrsg. von Max Bramer und Miltos Petridis. Cham & Heidelberg: Springer, S. 91–106. DOI: 10.1007/978-3-319-25032-8_6, v. a. S. 102–104; sowie Carmen Klaussner und Carl Vogel 2018: „A Diachronic Corpus for Literary Style Analysis". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7–12, 2018*. European Language Resources Association (ELRA), S. 3496–3503.

Allison et al. (2011) stoßen bei dem Versuch, unterschiedliche Textgenres mittels *PCA* zu unterscheiden, ebenfalls darauf, dass die häufigsten Wörter eines Textes wohl – mehr als über das Genre – Aufschluss über seinen Entstehungszeitraum geben könnten.[66]

Bei einer *PCA* unterschiedlicher diachroner schriftsprachlicher chinesischer Korpora fällt auf, dass sich bei Betrachtung der häufigsten 1.000 1–4-Zeichen-Lexeme stilistische und temporale Unterschiede in den beiden Hauptkomponenten widerspiegeln, da ältere Texte niedrigere *PC2*-Werte und umgangssprachlichere Texte höhere *PC1*-Werte erhalten (Abb. 3.1).[67] Die so entstandene Anordnung der untersuchten Texte ermöglicht zwar keine Datierung der Texte, zeigt aber sehr deutliche Veränderungen in der Wortverwendung über einen Zeitraum von ca. 3.000 Jahren.

In beiden Fällen – *k-means* Cluster-Analyse und *PCA* – sind es Häufigkeitsveränderungen der *n*-Gramme bzw. Wörter, die zur Einstufung der Texte führen. Dabei kann aber keine absolute Datierung angestrebt werden, nur die ungefähre Co-Datierung mit anderen, bereits datierten Texten bzw. die Datierung relativ zu anderen, stilistisch ähnlichen Texten des Korpus.

3.2 Datierung als Kategorisierungsproblem: de Jong, Rode und Hiemstra

de Jong, Rode und Hiemstra haben in ihrem Aufsatz „Temporal Language Models for the Disclosure of Historical Text“ (2005)[68] sicherlich Pionierarbeit in der Datierung von Texten auf Basis statistischer Sprachmodelle geleistet. Sie definieren die Datierung als Kategorisierungsproblem so:

> Given a date-tagged reference corpus, consisting of documents from a certain time span, and a document X with unknown date within the same time span, the system should classify X according to time partitions of predefined granularity.[69]

Als Bezeichnung für solche *time partitions* wird auch der Begriff *chronons* verwendet.[70] Bereits Swan und D. Jensen (2000) erzeugen statistische Sprachmodelle auf Basis eines Korpus aus datierten Dokumenten, untersuchen dabei aber nicht die zeitliche Einordnung der Texte, sondern stellen, ähnlich wie beim *Topic modelling*, die unterschiedlichen Themen der Texte auf einer Zeitleiste dar. Im Vordergrund steht dabei aber erstmalig die Ermittlung temporal diskriminativer Eigenschaften von Texten in Form von Phrasen, Namen und Wörtern.[71]

65 Abb. nach T. Schalmey 2021, S. 255, vgl. auch S. 259–260. Die untersuchten Texte sind die 25 offiziellen Dynastiegeschichten (*zhengshi* 正史, siehe auch Kapitel 2.3, ab S. 20), die 64 in der von Michael Loewe herausgegebenen Bibliographie *Early Chinese Texts* vorgestellten Texte (siehe auch 4.2, S. 66), sowie sechs *Classic Chinese Novels* nach C. T. Hsia. Siehe auch Michael Loewe, Hrsg. 1993: *Early Chinese Texts: A Bibliographical Guide*. Berkeley: The Society for the Study of Early China; The Institute of East Asian Studies; Hsia Chih-tsing 夏志清 1968: *The Classic Chinese Novel: A Critical Introduction*. New York und London: Columbia University Press.

66 Siehe Sarah Allison et al. 2011: „Quantitative Formalism: an Experiment“. In: *Pamphlets of the Stanford Literary Lab* 1, S. 1–24, S. 10; Allison et al. verfolgen diese Erkenntnis nicht weiter. Dass Genres selbst wiederum gewissen Lebenszyklen unterliegen können und die Genrezugehörigkeit damit *per se* umgekehrt Trägerin temporaler Informationen sein kann, zeigt sich auch in Ted E. Underwood 2016: „The Life Cycles of Genres“. In: *Journal of Cultural Analytics* 1.1. DOI: 10.22148/16.005, denn „things we call ‚genres‘ may be entities of different kinds, with different life cycles and degrees of textual coherence.“ (S. 24).

67 Siehe T. Schalmey 2021, S. 254–255.

68 de Jong, Rode und Hiemstra 2005.

69 Ebd., S. 3.

70 Siehe auch S. 50.

71 Siehe Russell Swan und David Jensen 2000: „TimeMines: Constructing Timelines with Statistical Models of Word Usage“. In: *Proceedings of KDD-2000 Workshop on Text Mining*, S. 1, S. 4–5.

de Jong, Rode und Hiemstra stellen fest, dass „diese Modelle […] es uns ermöglichen, einen Text anhand der Zeitspanne zu klassifizieren, aus der er stammt."[72] Die ursprüngliche Absicht der Autor:innen, die zeitliche Einordnung von Texten zu nutzen, um Suchergebnisse nach Relevanz sortiert darzustellen, gerät dabei in den Hintergrund. Die Verbesserung der Relevanz von Suchergebnissen, sowohl bei Internetsuchmaschinen als auch innerhalb von digitalen Bibliotheken, wird dennoch generell als wichtige Motivation für Bemühungen um die Datierung von Texten gesehen.[73]

Aus einem Korpus von niederländischen Zeitungsartikeln aus den Jahren 1999–2005 werden statistische Sprachmodelle erzeugt bzw. trainiert. Dokumente aus einer anderen Zeitung sollen dann anhand dieser Sprachmodelle datiert werden.[74] Hierbei werden zwei unterschiedliche Ansätze verfolgt: die Datierung über den Zeitstempel desjenigen *Dokuments* mit dem ähnlichsten Sprachmodell (d. h. den ähnlichsten Worthäufigkeiten),[75] sowie die Datierung mittels Sprachmodellen für Zeitabschnitte (*temporal language models*). Hierzu werden in unterschiedlicher Granularität (zwei Tage bis zu einem Vierteljahr) Sprachmodelle aus den aggregierten Worthäufigkeiten *aller* Dokumente eines bestimmten Zeitabschnitts (de Jong, Rode und Hiemstra verwenden hier den Begriff *time partitions*) berechnet und der zu datierende Text dann dem Zeitabschnitt mit dem ähnlichsten Sprachmodell chronologisch zugeordnet.[76] Als Ähnlichkeitsmaß wird die von Kraaij (2004) definierte *Normalized Log-Likelihood-Ratio* (*NLLR*, s. u.) verwendet.[77]

Die Datierung über die Zuordnung zum „ähnlichsten" Dokument liefert im gegebenen Kontext insgesamt bessere Ergebnisse als die Datierung über Zeitabschnitte.[78] Dass hier mit Zeitungstexten gearbeitet wurde und die zu datierenden Texte aus anderen Zeitungen stammen als das Trainingskorpus, legt allerdings nahe, dass letzteres häufig Artikel über dieselben Themen oder Ereignisse enthält, die auch im Fokus des zu datierenden Dokuments stehen. Inhaltliche Themen können bei einem solchen *Document Co-Dating* und bei der Verwendung sehr kurzer *time partitions* also mehr für die korrekte Datierung ausschlaggebend gewesen sein als eine sprachliche Veränderung.[79]

Zur Bewertung der *Verlässlichkeit* der mit beiden Methoden vergebenen Zeitstempel wird das „timely scattering" der übereinstimmendsten Zeitabschnitte oder Dokumente verwendet. Dieser Ansatz erscheint intuitiv sinnvoll: Ergeben sich für benachbarte Zeitabschnitte ähnlich hohe Übereinstimmungen, ist die Zuordnung mit hoher Wahrscheinlichkeit richtig. Je weiter hingegen die mit dem zu datierenden Text am stärksten übereinstimmenden Sprachmodelle zeitlich auseinanderliegen, desto geringer die Sicherheit der Zuordnung.[80]

72 de Jong, Rode und Hiemstra 2005, S. 1, übersetzt durch den Verfasser.
73 Vgl. u. a. de Jong, Rode und Hiemstra 2005; Kanhabua und Nørvåg 2008; Bamman et al. 2017.
74 Siehe de Jong, Rode und Hiemstra 2005, S. 6.
75 Siehe ebd., S. 3–5.
76 Siehe ebd., S. 4–5, S. 7.
77 Siehe de Jong, Rode und Hiemstra 2005, S. 3; Wessel Kraaij 2004: *Variations on Language Modeling for Information Retrieval (Diss.)* Enschede: Neslia Paniculata, S. 54.
78 Siehe de Jong, Rode und Hiemstra 2005, S. 8.
79 Vgl. auch ebd., S. 4: „Specific topics usually are discussed during shorter time spans and within a year almost any topic can be mentioned."
80 Siehe ebd., S. 6.

3.3 Systematisierung der bestehenden Ansätze

Die von de Jong, Rode und Hiemstra beschriebene grundlegende Methodik wurde in unterschiedlichen Bereichen weiter entwickelt und experimentell beforscht. Einige grundsätzliche Vorgehensweisen sind fast allen hier vorgestellten Studien gemein:

— 1. ***Pre-processing***. Zumeist kommen computerlinguistische Standardverfahren zum Einsatz, um die verwendeten Korpusdaten aufzubereiten. Dazu zählen unter anderem die Berechnung von Wort- oder Zeichenhäufigkeitslisten, *Part-of-Speech Tagging*,[81] bis hin zu *Collocation extraction*[82] und Unterscheidung von Wortbedeutungen. So kann z. B. bei Vorkommen von „bank" unterschieden werden, ob es sich um eine „money bank" oder eine „river bank" handelt.[83]

— 2. **Trainings- und Testdaten**. Die beschriebenen statistischen Ansätze basieren in der Regel auf der Verfügbarkeit großer diachroner Korpora, die für das Chinesische nur sehr bedingt zur Verfügung stehen.[84] Bei der Arbeit mit Korpora ist es gängige Praxis, einen Großteil der verfügbaren Daten für das Training (in diesem Fall der Datierungssoftware bzw. von Sprachmodellen) zu nutzen, und einen Anteil „beiseite zu legen", mit dem die Performance der Software später getestet bzw. bewertet werden soll.[85]

— 3. **Bewertung**. Der Erfolg einer Methode kann mit einer sogenannten *Baseline* verglichen werden, oft eine vergleichbare, frühere Studie,[86] oder die Wahrscheinlichkeit, mit der ein Zufallsgenerator das richtige Ergebnis liefern würde.[87] Weit verbreitet ist zudem die Angabe der *Accuracy* der zu bewertenden Methode, in diesem Fall also der Anteil korrekt datierter Dokumente, oder der durchschnittliche Fehler (*mean error*) in Jahren (z. B. „*x* % der zu datierenden Dokumente wurden auf *j* Jahre genau datiert.", „die durchschnittliche Abweichung der Datierung vom Zeitstempel beträgt *y* Jahre.") Dennoch können die Ergebnisse unterschiedlicher Studien meist kaum miteinander verglichen werden, da nicht nur Methodik, sondern auch Untersuchungsgegenstände sehr unterschiedlich sein können.

Die folgende Systematisierung der bestehenden Forschung soll einen Überblick über relevante Unterschiede und Gemeinsamkeiten zwischen den inzwischen zahlreichen Studien über Datie-

81 Die Zuweisung von Wortarten (*parts of speech*) zu den einzelnen *tokens* eines Texts ermöglicht es z. B., nur bestimmte Wortarten zu betrachten und andere herauszufiltern. Siehe Kanhabua und Nørvåg 2008, S. 361.

82 Hierbei wird der Wortkontext mit erfasst, etwa um die einzelne Verwendung von „united" oder „states" von dem Ausdruck „United States" zu unterscheiden. Fortschrittliche Segmenter bzw. Tokenizer berücksichtigen solche Ausdrücke, die sich aus mehreren Wörtern zusammensetzen. Siehe ebd., S. 361.

83 Siehe ebd.

84 Siehe dazu Kapitel 4.2, ab S. 62.

85 Siehe z. B. Kanhabua und Nørvåg 2008, S. 366.

86 Siehe z. B. A. Kumar et al. 2012; als *Baseline* verwenden Jannik Strötgen und Michael Gertz 2010: „HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, S. 321–324. URL: http://www.aclweb.org/anthology/S10-1071; vgl. auch Chambers 2012, hier wird sehr stark auf; Kanhabua und Nørvåg 2008, Bezug genommen.

87 Vgl. z. B. de Jong, Rode und Hiemstra 2005, S. 7. Ein Zufallsgenerator, der hier als *Baseline* eingesetzt wird, würde hier nur 4 % der Dokumente dem richtigen Zeitraum zuordnen.

rungsmethoden für westliche Sprachen ermöglichen.[88] Dabei wird im Rahmen der einzelnen Punkte gegebenenfalls auf die Anwendbarkeit und Relevanz für das Chinesische eingegangen.

— 1. **Entstehungszeit vs. erzählte Zeit**. Zwei Zielsetzungen müssen grundsätzlich unterschieden werden: die Schätzung bzw. Ermittlung der „erzählten" Zeit, d. h. der Zeit, über die geschrieben wird,[89] sowie die Ermittlung der Zeit der Entstehung oder Veröffentlichung des Textes.[90] Ein historischer Roman etwa enthält viele Namen und Konzepte aus der Zeit, in der sich die Handlung abspielt, unabhängig davon, wann er verfasst wurde. Auf den Aspekt der Datierung der *Entstehung* von Texten wird hier verstärkt eingegangen – eine strikte Trennung ist aber nicht immer möglich, z. B. wenn mit Nachrichten oder mit Geschichtstexten gearbeitet wird, die kurz nach dem darin beschriebenen Zeitraum entstanden sind.[91] Eine zusätzliche, wichtige Unterscheidung sollte gegebenenfalls zwischen *publication date* und *creation date* getroffen werden. Da große Online-Bibliotheken wie *GoogleBooks* oder *HathiTrust* oft viele Ausgaben desselben Werkes enthalten, einige viele Jahrzehnte nach der Erstveröffentlichung bzw. Entstehung datiert, eichen solche Trainingsdaten eine Datierungssoftware eher auf *Manifestationen* von Werken, weniger auf ihre ursprüngliche Entstehung.[92]

— 2. **Grundlegende Methodik**. In der Methodik für die computerlinguistische Datierung von Texten anhand textueller Merkmale lassen sich zwei Herangehensweisen unterscheiden. Kanhabua und Nørvåg unterteilen diese in „learning based" und „non learning based".[93] Zur Schätzung der Entstehungszeit von Texten werden eher Methoden eingesetzt, die – wie in 3.2 beschrieben – Worthäufigkeiten mit statistischen Ähnlichkeitsmaßen vergleichen und so die Zuordnung des Textes zu einem bestimmten *chronon*,[94] oder eine Co-Datierung[95] mit dem ähnlichsten Text zu ermöglichen. Solche Verfahren erfordern die Verfügbarkeit diachroner Textkorpora, aus denen temporale Sprachmodelle (*temporal language models*) „erlernt" bzw. trainiert werden können. Diese Modelle umfassen im Wesentlichen die Worthäufigkeiten der Texte aus den jeweiligen Trainingsdaten. Diese können über vordefinierte Zeitabschnitte aggregiert, oder einzeln betrachtet werden.[96]

Ergänzend oder alternativ können Methoden aus dem Bereich des *machine learning* für die Textdatierung eingesetzt werden, die in jüngster Zeit zunehmend erforscht wurden.[97]

Zusätzlich können explizite Zeitangaben und ähnliche Hinweise aus Texten extrahiert bzw. markiert werden (*temporal tagging*). Mithilfe von Regeln bzw. Mustern werden Jahreszahlen und andere Zeitausdrücke (*temporal expressions*) wie Daten, Wochentage oder Monate im Text erkannt.

88 Ein knapp gehaltener, aktueller Abriss findet sich in Toner und Han Xiwu 2019, S. 11–66. Darin werden auch rezente Strömungen und Entwicklungen behandelt, die über die hier vorgestellten Methoden hinausgehen, unter anderem aus dem Bereich des *machine learning*. Ein allgemeiner Überblick, der sich auf inhaltsbasierte bzw. statistische Methoden beschränkt, findet sich auch in Kristoffer Berg Gumpen und Øyvind Vik Nygard 2017: „Automatic Document Timestamping". Masterarbeit. Trondheim: Norwegian University od Science und Technology (NTNU), S. 5–9.

89 Siehe z. B. A. Kumar 2013, S. 13–14, S. 15–16. Kumar experimentiert zu diesem Zweck mit Biographien und Artikeln zu einzelnen Jahren aus der englischsprachigen Wikipedia.

90 Vgl. Kanhabua und Nørvåg 2008, S. 359; vgl. auch de Jong, Rode und Hiemstra 2005, S. 6.

91 Siehe dazu Kapitel 6.3, S. 211.

92 Siehe Bamman et al. 2017, S. 8.

93 Siehe Kanhabua und Nørvåg 2008, S. 359.

94 Siehe z. B. ebd., S. 366.

95 Vgl. z. B. de Jong, Rode und Hiemstra 2005, S. 7.

96 Ausführlicher siehe unter 5.., ab S. 50.

97 Siehe Garcia-Fernandez et al. 2011, S. 8–10; oder Bamman et al. 2017, S. 5; siehe v. a. auch Toner und Han Xiwu 2019, S. 3.

Auch die Verarbeitung relativer Zeitangaben („vor zwei Wochen", „heute" usw.) ist möglich.[98] Mit *HeidelTime* steht eine quelloffene, temporale Tagging-Software für zahlreiche Sprachen, auch für modernes Chinesisch, zur Verfügung.[99]

Selbstverständlich eignet sich das Erkennen von konkreten Zeitangaben in Texten vor allem, um zu ermitteln, über welche Zeit geschrieben wird und weniger, wann ein Text verfasst wurde. Dennoch wurden z. B. Nennungen von Jahreszahlen in Kombination mit bestimmten Präpositionen erfolgreich als Ergänzung zu statistischen Sprachmodellen in der Textdatierung eingesetzt.[100] Guo Siyuan et al. haben zudem gezeigt, dass für englischsprachige Texte oft auch die erste im Text erwähnte Jahreszahl („first date in text") gute Hinweise auf die Entstehungszeit eines Textes liefern kann.[101] Besondere Jahreszahlen wie 2000 werden jedoch auch in früheren Texten häufig referenziert und sind daher mit Vorsicht zu genießen.[102] Viele, gerade kurze, literarische Texte enthalten allerdings oft keinerlei solche Angaben. Dass im Text enthaltene Zeitangaben eher Rückschlüsse auf die *erzählte* Zeit zulassen, ist zudem für unterschiedliche Textgattungen mehr oder weniger problematisch. Bei Nachrichten, die kurz nach dem Ereignis geschrieben werden, über das berichtet wird, können Entstehungszeit und erzählte Zeit nahezu identisch sein. Bei anderen, v. a. historiographischen Textgattungen, ist dies aber nicht der Fall.

Die Berechnung von Sprachmodellen lässt sich grundsätzlich auch für schriftsprachliches Chinesisch umsetzen, wobei das Fehlen eines zuverlässigen Tokenizers Einschränkungen mit sich bringt.[103] Methoden zur Erkennung temporaler Ausdrücke müssen stark angepasst bzw. erweitert werden, da für Jahreszahlen in der Regel nicht das Format des gregorianischen Kalenders verwendet wird.[104]

— 3. ***Bag of words*, *n*-Gramme und *shingles*.** In den meisten Studien kommt ein Unigramm-Sprachmodell bzw. *Bag of Words*-Modell (*BoW*) zum Einsatz, bei dem, wie oben beschrieben, die relativen Häufigkeiten der in den zu datierenden Texten vorkommenden Wörter bzw. Wortformen betrachtet werden.[105] Dabei entspricht die Anzahl der unterschiedlichen Wörter (*types*) in allen betrachteten Dokumenten der Dimensionalität des Vektorraums.[106] Diese Dimensionen werden auch als *features* eines Textes bezeichnet. Obwohl weder Kontext, noch Bedeutung oder Reihenfolge der Wörter in einem Text berücksichtigt werden, lassen sich mit solchen Unigramm-Modellen sehr gute Ergebnisse erzielen.[107]

98 Siehe Inderjeet Mani und George Wilson 2000: „Robust Temporal Processing of News". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. ACL '00. Hong Kong: Association for Computational Linguistics, S. 69–76. doi: 10.3115/1075218.1075228. url: https://doi.org/10.3115/1075218.1075228, S. 69–70.

99 Strötgen und Gertz 2010; Für aktuellere Arbeiten aus dem Bereich des *temporal tagging* siehe auch Jannik Strötgen 2015: „Domain-sensitive Temporal Tagging for Event-centric Information Retrieval". Diss. Heidelberg: Universität Heidelberg; Jannik Strötgen und Michael Gertz 2016: *Domain-Sensitive Temporal Tagging*. Hrsg. von Graeme Hirst. Synthesis Lectures on Human Language Technologies 36. San Rafael: Morgan & Claypool.

100 Siehe z. B. Chambers 2012, S. 101–105. Ausdrücke wie z. B. „not […] open until February 2000" werden hier als „temporal constraints" eingesetzt. In dem Beispiel wird davon ausgegangen, dass der Text vor dem Jahr 2000 verfasst wurde.

101 Siehe Guo Siyuan et al. 2015, S. 3; siehe auch Graliński et al. 2017, S. 31.

102 Siehe Graliński et al. 2017, S. 31.

103 Siehe dazu Kapitel 4, insb. 4.4 (ab S. 73) u. 4.5 (S. 77). Ein möglicher Lösungsansatz ist in Kapitel 4.5.3 (S. 94) beschrieben.

104 Siehe dazu Kapitel 4.8, ab S. 103.

105 Siehe z. B. de Jong, Rode und Hiemstra 2005; Kanhabua und Nørvåg 2008; Garcia-Fernandez et al. 2011, S. 5; A. Kumar et al. 2012, S. 5; Zampieri, Malmasi und Dras 2016; Bamman et al. 2017; Vgl. auch Gumpen und Nygard 2017, S. ii: „Due to its simplicity, efficiency, and often surprising accuracy, the bag of words model is one of the most common fixed-length vector representations for text."

106 Siehe auch Feldman und Sanger 2006, S. 68.

107 Siehe z. B. Zampieri, Malmasi und Dras 2016; Kulkarni et al. 2018.

Stehen ausreichend große Korpora zur Verfügung, können die erzeugten Modelle auf Wort-Bi- und Trigramme usw. erweitert werden, um den zusätzlichen Informationsgehalt von Anordnung bzw. Kontext der Wortverwendung in die Analyse einzubeziehen.[108] Solche Wort-*n*-Gramme werden von manchen Autor:innen auch als *k*-Gramme oder *k-shingles* bezeichnet.[109]

Anstelle von Wörtern und Wortfolgen können auch Häufigkeiten von *Zeichen*folgen betrachtet werden. Graliński et al. haben gezeigt, dass die Betrachtung von Zeichenpentagrammen[110] anstelle von vollständigen Wortformen die Performance eines Datierungssystems für das Polnische sogar verbessern kann, da diese robuster gegenüber *OCR*-Rauschen sind und zugleich eine Art implizite Lemmatisierung für viele Wörter vorgenommen wird. Dieser Ansatz lässt sich natürlich auch für andere flektierende Sprachen adaptieren,[111] für isolierende Sprachen wie das Chinesische hat Lemmatisierung aber keine Relevanz.

Dass aussagekräftige *SLM* auch mit Zeichen-*n*-Grammen funktionieren, ist dennoch eine wichtige Erkenntnis, da ohne zuverlässige Tokenisierung für Texte des als *scriptura continua* geschriebenen Chinesischen kein echtes *BoW*-Modell berechnet werden kann.[112] Meng Yuxian et al. argumentieren, dass Zeichen-*n*-Gramm-Repräsentationen von chinesischsprachigen Texten für viele Anwendungsgebiete computerlinguistischer Methoden der Betrachtung von wortsegmentierten Texten sogar überlegen sind.[113] Da fast alle Wörter eine Länge von 1–4 Zeichen aufweisen,[114] verschwimmt für das Chinesische die oben vorgenommenene Unterscheidung zwischen Zeichen- und Wort-*n*-Grammen bzw. *shingles* aber ohnehin. Wie geeignet unterschiedliche Repräsentationen des Chinesischen dabei für die Textdatierung sind, wird in Kapitel 6.1 eingehend untersucht.[115]

— 4. **Sprachwandel**. Allen Methoden, die eine Datierung auf Basis der im Text enthaltenen Wörter bzw. sprachlichen Erscheinungen und deren Häufigkeit anstreben, ist die zugrundeliegende Idee gemein, dass Wörter bzw. Wortformen eine Art temporale Lokalität aufweisen können. Während einige Wörter über einen langen Zeitraum konstant genutzt werden, kommen andere aus der Mode oder es entstehen Neologismen.[116] Dieses Konzept wird im Kontext der computerlinguistischen Textdatierung zuerst von Garcia-Fernandez et al. ausformuliert:

> Both neologisms and archaisms constitute interesting cues for identifying publication dates: given the approximate year of apparition of a word, one can assign a low probability for all preceding years and a high probability to following years (the reverse line of argument can be applied to archaisms). However, there is no pre-compiled list of words with their year of appearance or disappearance.[117]

108 Siehe Guo Siyuan et al. 2015, S. 4.
109 Siehe z. B. Gumpen und Nygard 2017, S. 12; Bamman et al. 2017, S. 4.
110 Die zu betrachtenden Texte werden hierfür in Segmente von jeweils 5 Buchstaben zerlegt.
111 Siehe Graliński et al. 2017, S. 33.
112 Ausführlicher siehe Kapitel 4, insb. 4.4, ab S. 73 u. 4.5, S. 77; vgl. auch 4.5.3, S. 94.
113 Siehe Meng Yuxian et al. 2019: „Is Word Segmentation Necessary for Deep Learning of Chinese Representations?" In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Hrsg. von Anna Korhonen, David R. Traum und Lluís Màrquez. Association for Computational Linguistics, S. 3242–3252. doi: 10.18653/v1/p19-1314, S. 3249.
114 Siehe dazu auch Kapitel 5.7, v. a. S. 149.
115 Siehe ab S. 156.
116 Siehe dazu auch Kapitel 2, ab S. 11.
117 Garcia-Fernandez et al. 2011, S. 5.

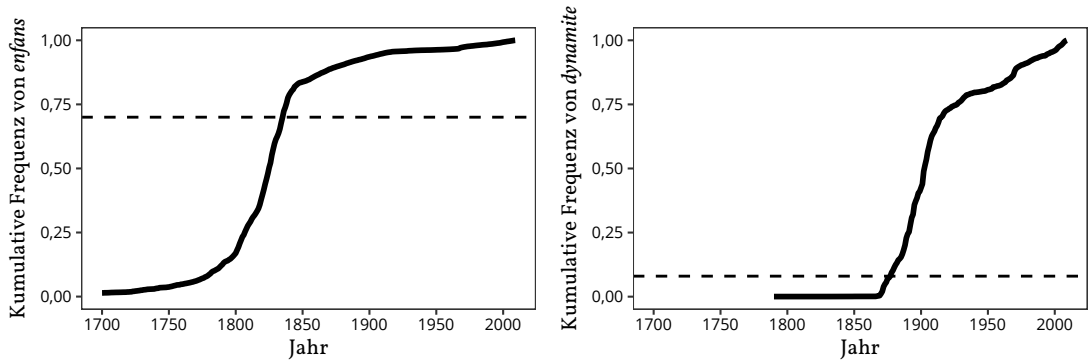


Abbildung 3.2 Kumulative Häufigkeit von „Archaismen“ und „Neologismen“[119]

In Ermangelung einer solchen Wortliste entwickeln sie eine Methodik, um Erkenntnisse über das Entstehen und Verschwinden von Wörtern aus kumulierten *Google n*-Gramm-Daten zu erzeugen.[118] Demnach wird ein Wort, das ab einer bestimmten Zeit mit anschließend steigender kumulativer Häufigkeit auftritt, als Neologismus angesehen. Ein Archaismus wird an einer abflachenden *s*-Kurve erkannt (Abb. 3.2).

Mittels eines auf Basis von Trainingsdaten ermittelten Schwellenwertes[120] können entsprechende Listen erzeugt werden, zu welchem Zeitpunkt welche Neologismen auftreten, bzw. ab wann Begriffe bzw. Schreibweisen außer Mode kommen und als Archaismus eingestuft werden können. Gemäß den Beispielen aus Abb. 3.2 wurde ein französischsprachiger Text mit der Schreibweise *enfans* mit hoher Wahrscheinlichkeit vor 1835 verfasst, ein Text, der den Begriff *dynamite* enthält, nach 1875.

Die Kurve, wie sie v. a. für die kumulative Häufigkeit der so erkannten Archaismen beobachtet werden kann, erinnert dabei an die Visualisierung der Beobachtungen von Piotrowski für Sprachwandel.[121]

Limitationen dieser Herangehensweise ergeben sich daraus, dass die zugrunde liegenden *n*-Gramm-Daten lediglich den Zeitraum nach 1500 abdecken und überdies wenig verlässlich sind, da *Google Books* nicht zwischen Erstausgaben und späteren Manifestationen eines Buches unterscheidet.[122] *Google Books Ngrams* stehen auch für das Chinesische zur Verfügung, wobei nur sehr wenige Texte vor 1850 datiert sind und offensichtlich eine vollständige Normalisierung auf Kurzzeichen vorgenommen wurde. Für die Zeit nach 1900 ließen sich hier aber auch für das Chinesische sicherlich interessante Daten zum Wortschatzwandel gewinnen.

Ein weiterer Aspekt des Sprachwandels ist der phonologische Wandel. Bei Sprachen, die mit einer alphabetischen Schrift repräsentiert werden, deren Zeichen im Wesentlichen die Phoneme

118 Siehe Garcia-Fernandez et al. 2011, S. 5–6; Eine sehr ähnliche Methode auf Basis einer *PCA* wird auch in Chiru und Rebedea 2014, beschrieben. Die englischen 1-Gramm-Daten aus *Google Books* werden graphisch analysiert, um eine automatische Klassifizierung von *types* in Archaismen, Neologismen und „common words“ als Vorstufe für weitere NLP-Anwendungen vorzunehmen.

119 Graphik nach Garcia-Fernandez et al. 2011, S. 6. Daten von *Google Books Ngrams*.

120 Dieser *threshold* wird für Archaismen mit 0,7 und Neologismen mit 0,08 angegeben (gestrichelte Linie in Abb. 3.2). Siehe ebd.

121 Siehe Kapitel 2.1, ab S. 14.

122 Siehe z. B. Bamman et al. 2017, S. 6; siehe auch Geoffrey Nunberg 2009: „Google's Book Search: A Disaster for Scholars“. In: *The Chronicle*. URL: https://www.chronicle.com/article/googles-book-search-a-disaster-for-scholars/ (besucht am 31. 08. 2009).

einer Sprache, Vokale und Konsonanten, grob wiedergeben, ist die Verwendung des Alphabets relativ flexibel und kann diesem Wandel angepasst werden.[123] Diese und andere Veränderungen der Orthographie, z. B. durch Rechtschreibreformen, sind relativ einfach datierbar und können daher die temporale Einordnung von Texten erleichtern.[124] Im Gegensatz dazu ist die chinesische Schrift gegenüber Lautveränderungen in der gesprochenen Sprache ziemlich resistent. Es existieren – bedingt u. a. durch Tabuisierung – überdies zwar zahlreiche zeitlich oder lokal begrenzt verwendete Zeichenvarianten,[125] die in der Regel aber in digitalisierten, normalisierten Textfassungen kaum enthalten sind oder gar nicht wiedergegeben werden können.

— 5. ***Chronons*, kontinuierliche Datierung und Co-Datierung**. Die meisten Studien verwenden sprachliche Modelle von Zeitabschnitten, Intervallen unterschiedlicher Länge, um den zu datierenden Text entsprechend zu klassifizieren. Diese von de Jong, Rode und Hiemstra als *time partitions* eingeführten Zeitabschnitte, auch „buckets"[126] oder *chronons*[127] genannt,[128] werden als ein bestimmter, fixer Zeitraum von z. B. 50 oder 100 Jahren definiert, der für die Berechnung von Sprachmodellen (*language models*) eingesetzt wird.

Ein wesentlicher Nachteil solcher *chronons* sind die arbiträr festgelegten Grenzen zwischen den Zeiträumen. Gerade für Texte am „Rand" zwischen zwei vordefinierten Zeitabschnitten können sich hohe Wahrscheinlichkeiten für beide angrenzenden Zeitabschnitte ergeben. Dieses Problem kann durch die Verwendung überlappender *chronons* minimiert werden.[129] Für jedes *chronon* sollte dabei dieselbe Menge an Trainingsdaten zur Verfügung stehen.[130] Alternativ können *chronons* unterschiedlicher Länge definiert werden, um sie der Verfügbarkeit von Trainingsdaten anzupassen, was allerdings eine ungleiche Granularität der Datierungsergebnisse mit sich bringt.[131] Zudem kann die Granularität der gewünschten Datierung auch gröber als die der trainierten Modelle gewählt werden.[132]

Intuitiv wird Zeit nicht in gleich oder unterschiedlich langen, überlappenden Abschnitten, sondern als kontinuierliche Variable wahrgenommen. Für die Praxis der Textdatierung stellt sich dies jedoch als schwierig bzw. ineffizient heraus. Graliński et al. beklagen, dass „publication time can be viewed as a continuous variable and with given training data any regression algorithm should be able to predict a specific point in time for any input text. However, this approach does not prove to be effective."[133]

Durch Zuweisung des Zeitstempels eines einzelnen Dokuments aus den Trainingsdaten (*Document Co-Dating*) kann zudem ebenfalls die Datierung auf ein bestimmtes Jahr bzw. ein bestimmtes Datum ermöglicht werden. de Jong, Rode und Hiemstra konnten mit Co-Datierung von Texten auf den Text im Trainingskorpus mit dem ähnlichsten Sprachmodell bessere Ergebnisse erzielen

123 Vgl. auch Roger Lass 1997: *Historical Linguistics and Language Change*. Cambridge Studies in Linguistics. Cambridge & New York: Cambridge University Press, S. 79.
124 Siehe z. B. de Jong, Rode und Hiemstra 2005, S. 1.
125 Siehe Kapitel 4.3, ab S. 69, v. a. S. 70.
126 Bamman et al. 2017, S. 8.
127 Siehe z. B. A. Kumar et al. 2012, S. 5; Guo Siyuan et al. 2015, S. 1.
128 Siehe u. a. auch de Jong, Rode und Hiemstra 2005; A. Kumar et al. 2012; Graliński et al. 2017.
129 Siehe de Jong, Rode und Hiemstra 2005, S. 4.
130 Ebd., S. 6–7.
131 Guo Siyuan et al. 2015, S. 3.
132 Siehe de Jong, Rode und Hiemstra 2005, S. 4.
133 Graliński et al. 2017, S. 29; Durch gezieltes Entfernen von *features*, die zu einer Klassifizierung besonders beigetragen haben, kann ein vorhergesagtes Jahr nach vorne oder hinten verschoben werden und so der Betrachtung eine gewisse Linearität zurückgegeben werden. Siehe Bamman et al. 2017, S. 8.

als bei der Verwendung von *chronons*.[134] Bamman et al. bezeichnen eine solche Herangehensweise auch als „Content-based deduplication".[135] Indem die n „ähnlichsten" Texte aus dem Korpus betrachtet und das darin häufigste *chronon* zur Datierung angenommen wird, kann das Konzept der *chronons* oder *time partitions* auch mit *Co-Dating* kombiniert werden.[136]

Neben dem Vergleich von Textinhalt bzw. Worthäufigkeiten können – sofern vorhanden – auch die Metadaten von Texten wie Titel und Autor verglichen werden.[137] Beide Fälle erfordern ein umfassendes Trainingskorpus, das ähnliche Dokumente enthält bzw. im zweiten Fall sogar eine bereits richtig datierte Version des zu datierenden Dokuments.

— 6. **Ähnlichkeitsmaße**. Für den Vergleich von Wort- oder n-Gramm-Häufigkeitslisten bzw. zur Messung der Ähnlichkeit der *Bag of Words* von Texten zu anderen Texten oder berechneten Sprachmodellen werden unterschiedliche Ähnlichkeitsmaße eingesetzt.

Im Folgenden gilt, sofern nicht anders angegeben:
C sei definiert als ein Korpus-Sprachmodell, untergeordnet sind Sprachmodelle für unterschiedliche *chronons* c.
$P(w \mid c)$ sei definiert als die relative Häufigkeit (*term frequency, tf*) eines Wortes w in einem *chronon* c, $P(w \mid d)$ dessen relative Häufigkeit in dem zu datierenden Dokument d, mit f als Anzahl der Vorkommen (*tokens*) von w:

$$P(w|c) = tf_{w,c} = \frac{f_{w,c}}{\sum_{w' \in c} f_{w'}}$$

— 6.1 Die **Kosinus-Ähnlichkeit** (*cosine similiarity*, *CS*) ist ein häufig eingesetztes Maß für die Ähnlichkeit zweier Dokumente.[138] Dabei wird jedes Dokument als ein n-dimensionaler Vektor betrachtet, der die relativen Worthäufigkeiten der *types* des jeweiligen Dokuments beinhaltet.[139]

Die *CS* als Vergleich der Wahrscheinlichkeitsverteilungen zwischen einem *chronon*-Modell c und einem Vergleichsdokument (Query-Dokument) d ist definiert als:[140]

$$CS_{d,c} = \frac{\sum_{w \in d} P(w|d) \times P(w|c)}{\sqrt{\sum_{w \in d} P(w|d)^2} \times \sqrt{\sum_{w \in c} P(w|c)^2}}$$

— 6.2 Die von Kraaij definierte ***Normalized Log-Likelihood-Ratio***[141] (*NLLR*) wird bereits in der Studie von de Jong, Rode und Hiemstra als Ähnlichkeitsmaß für den Vergleich zwischen Dokumenten und *chronon* Sprachmodellen eingesetzt. Die Besonderheit ist dabei,

134 de Jong, Rode und Hiemstra 2005, S. 7–8.
135 Bamman et al. 2017, S. 4.
136 Siehe de Jong, Rode und Hiemstra 2005, S. 5. Bessere Ergebnisse erzielen die Autor:innen aber mit $n = 1$, also einem reinen *Document Co-Dating*. Siehe S. 7.
137 Siehe Bamman et al. 2017, Ausführlicher siehe auch auf S. 56.
138 Siehe Guo Siyuan et al. 2015, S. 5.
139 Siehe z. B. Tan Pang-Ning 陳封能, Michael Steinbach und Vipin Kumar 2013 [2005]: *Introduction to Data Mining*. Essex: Pearson, S. 69–72; Auch für das Clustering von Dokumenten wird *cosine similarity* gerne eingesetzt. Siehe Feldman und Sanger 2006, S. 85.
140 Siehe Guo Siyuan et al. 2015, S. 5.
141 Kraaij 2004, S. 54. Kraaij selbst präferiert für die *NLLR* die Bezeichnung „cross-entropy reduction ranking" (*CER*), da sie als Differenz zwischen der Kreuzentropie von d und c sowie d und C geschrieben werden kann.

dass zusätzlich zur Häufigkeit einer Wortform in dem zu untersuchenden und dem Vergleichsdokument bzw. -*chronon* auch die Häufigkeit im gesamten Korpus (d. h. über alle *chronons*) betrachtet wird.[142]

Die *NLLR* ist definiert als:

$$NLLR_{d,c} = \sum_{w \in d} P(w \mid d) \times \log\left(\frac{P(w \mid c)}{P(w \mid C)}\right)$$

wobei d ein Query-Dokument, c das aggregierte *chronon*-Modell und C das gesamte Korpus als Hintergrundmodell bezeichnen, $P(w \mid d)$ als relative Häufigkeit bzw. Wahrscheinlichkeit des Auftretens von w in d usw.[143]

Für die korrekte Berechnung der *NLLR* muss $P(w \mid c) > 0$ sein, da sonst der Logarithmus nicht definiert ist.[144]

— 6.3 Die **Kullback-Leibler-Divergenz** (*KLD*) ist ein weiteres Maß für die Unterschiedlichkeit von Wahrscheinlichkeitsverteilungen, das von Solomon Kullback und Richard Leibler definiert wurde.[145] Als Maß für den Unterschied zwischen der Wortwahrscheinlichkeitsverteilung eines Texts d und eines *chronon*-Modells c[146] ist sie definiert als:[147]

$$KLD_{d,c} = \sum_{w \in d} P(w \mid d) \times \log \frac{P(w \mid d)}{P(w \mid c)}$$

Wie bei der *NLLR* muss für die Berechnung der *KLD* $P(w \mid c) > 0$ sein, da die Division mit Divisor 0 nicht definiert ist.

— 6.4 Einfach zu berechnen ist der **Jaccard-Koeffizient**, auch Jaccard *similarity* genannt. Dieses Maß für die Ähnlichkeit zweier Mengen geht auf den Schweizer Botaniker Paul Jaccard zurück und gibt an, welcher Anteil der Merkmale zweier Mengen in beiden Mengen auftreten.[148] Übertragen auf ein einfaches *BoW*-Modell: Welcher Anteil der vorkommenden Wortformen tritt – ungeachtet ihrer Häufigkeit – in beiden Texten auf.

Der Jaccard-Koeffizient J als Ähnlichkeitsmaß zwischen zwei Texten, bzw. zwischen einem Dokument d und einem Vergleichs-*chronon* c ist dabei definiert als:

$$J_{d,c} = \frac{|\, c \cap d \,|}{|\, c \cup d \,|}$$

Abgesehen vom Jaccard-Koeffizienten, der lediglich die Schnittmenge der vorkommenden *types* betrachtet, berücksichtigen alle hier aufgeführten Ähnlichkeitsmaße die relativen Worthäufigkeiten des zu datierenden Dokuments im Verhältnis zum Vergleichsmodell, d. h. denen der jeweils aggregierten *chronons* bzw. der Einzeldokumente aus den Trainingsdaten. Lediglich bei der *NLLR* werden auch die aggregierten Häufigkeiten der gesamten Trainingsdaten betrachtet.

142 Siehe Kraaij 2004, v. a. S. 203–204; siehe auch de Jong, Rode und Hiemstra 2005, S. 3.
143 Siehe de Jong, Rode und Hiemstra 2005, S. 3.
144 Siehe auch ebd.
145 Solomon Kullback und Richard A. Leibler 1951: „On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, S. 79–86. doi: 10.1214/aoms/1177729694.
146 Siehe z. B. A. Kumar 2013, S. *vi*.
147 Siehe Guo Siyuan et al. 2015, S. 5.
148 Siehe Paul Jaccard 1902: „Lois de distribution florale dans la zone alpine". In: *Bulletin de la Sociéé Vaudoise des Sciences Naturelles* 38.144, S. 69–130. doi: 10.5169/seals-266762, S. 72.

Über die Verwendung einzelner Ähnlichkeitsmaße hinaus kann die Zuverlässigkeit der Datierung durch die Kombination unterschiedlicher Methoden erhöht und besser messbar gemacht werden. Garcia-Fernandez et al. gelingt es z. B., deutlich mehr Texte korrekt zu datieren, indem *CS* und eine *support vector machine* (*SVM*)[149] miteinander kombiniert werden.[150]

— 7. **Gewichtung und Reduktion von *features***. Da Wörter, deren Häufigkeit über lange Zeit konstant ist, weniger Rückschlüsse über die Entstehungszeit eines Textes zulassen, als solche, die in einem bestimmten Zeitraum oder ab einem bestimmten Zeitpunkt auftreten, können Maßnahmen zur Gewichtung von *features* sinnvoll sein. Auch im Kontext der Datierung kann die im *Information Retrieval*[151] weit verbreitete *term frequency inverse document frequency* (*tf-idf*) eingesetzt werden, welche hier die Verteilung der Worthäufigkeiten über den Betrachtungszeitraum widerspiegelt. Die Häufigkeit jedes Wort-*types* *w* in einem Dokument *d* wird dabei so gewichtet, dass Wörter, die in einem bestimmten *chronon* besonders häufig auftreten mehr, und Wörter, die in besonders vielen Dokumenten bzw. *chronons* auftreten, weniger Gewicht erhalten. Dafür wird $P(w \mid d)$ mit dem logarithmisch skalierten Anteil der Dokumente bzw. *chronons* im Korpus, die *w* enthalten (hier geschrieben als *document frequency*, df_w, d. h. hier die Anzahl der *chronons* *c* in denen *w* vorkommt) multipliziert, wobei *N* die Anzahl der Dokumente bzw. der *chronons* des Korpus ist:[152]

$$tf - idf_{w,c} = tf_{w,c} \times \log_2\left(\frac{N}{df_w}\right)$$

Eine Weiterentwicklung dieses Konzepts ist die von Kanhabua und Nørvåg eingeführte *Temporal Entropy* (temporale Entropie, *TE*), also die Informationsdichte für die zeitliche Zuordnung.[153] Mit ausreichenden Daten für die verwendeten *chronons* kann sie zur Gewichtung eingesetzt werden und die Präzision der Datierung verbessern.[154] Die *TE* berücksichtigt dabei nicht nur, in wie vielen Dokumenten eine Wortform noch auftritt, sondern auch ihre in unterschiedlichen *chronons* unterschiedliche Häufigkeit im Vergleich zum Korpus. Sie ist definiert als:[155]

$$\mathrm{TE}_w = 1 + \frac{1}{\log_2 N_C} \times \left(\sum_{c \in C} tf_{w,c} \times \log_2 \frac{tf_{w,c}}{\sum_{c \in C} tf_{w,c}}\right)$$

Kanhabua und Nørvåg schlagen außerdem vor, alternativ die Nutzungsstatistiken der *Google*-Suche (veröffentlicht unter dem Namen *Google Zeitgeist*, inzwischen *Google Trends*) als externe Datenquelle einzusetzen. Diese Informationen über stark zu- und abnehmende Trends von

149 *SVM* ist ein Verfahren aus dem Bereich des *machine learning*, das die Klassifizierung von Objekten (z. B. Texten) durch das „Erlernen" der Unterschiede zwischen Gruppen von bereits klassifizierten Trainingsobjekten anhand eine Vektorraums (z. B. Worthäufigkeiten) ermöglicht.

150 Siehe Garcia-Fernandez et al. 2011, S. 8–10.

151 *Information Retrieval* bezeichnet „jede Form der Wiedergewinnung von gespeicherten Daten […], relevante Informationen in Datenquellen zu finden und nicht-relevante Informationen zu erkennen und auszuschließen. […]" Siehe Harald Klinke 2017: „Information Retrieval". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 268–278, S. 268.

152 Siehe z. B. Feldman und Sanger 2006, S. 68; zitiert in Gumpen und Nygard 2017, S. 13.

153 Siehe Kanhabua und Nørvåg 2008, S. 361, S. 364.

154 Siehe ebd., S. 368.

155 Siehe Kanhabua und Nørvåg 2008; vgl. auch Guo Siyuan et al. 2015, S. 6; Die ursprüngliche Definition der *Entropy* geht allerdings zurück auf Karen E. Lochbaum und Lynn A. Streeter 1989: „Comparing and Combining the Effectiveness of Latent Semantic Indexing and the Ordinary Vector Space Model for Information Retrieval". In: *Information Processing & Management* 25.6, S. 665–676, S. 672. Das Konzept wird dort auch als *noise measure* bezeichnet und ist rechnerisch identisch.

Suchbegriffen können zur Beurteilung der Wahrscheinlichkeiten für die entsprechenden *chronons* verwendet werden, sofern die zu datierenden Texte entsprechende Begriffe enthalten.[156] Im Vergleich mit dem Konzept der *TE* fällt die Verbesserung der *Accuracy* der durchgeführten Datierungsexperimente allerdings etwas geringer aus.[157]

Ein weiteres, ähnliches Konzept, das ebenfalls erfolgreich in der Textdatierung eingesetzt wird, ist *(term) burstiness*. Begriffe, die im Vergleich zu ihrer Häufigkeit in einem diachronen Gesamtkorpus in einzelnen Zeitabschnitten auffällig häufig sind, werden dabei als *bursty* angesehen, wofür die Diskrepanz der Worthäufigkeiten in den betrachteten Zeitabschnitten untersucht wird.[158] Die *types* in Dokumenten eines diachronen Korpus und deren *bursty intervals*, also Zeitabschnitte mit ungewöhnlichen großen Worthäufigkeiten,[159] werden ermittelt.[160] Kotsakos et al. verwenden *burstiness* in ihrer Textdatierungsmethode, indem zunächst mithilfe des Jaccard-Koeffizienten die Dokumente aus den Trainingsdaten ermittelt werden,[161] deren Wortschatz dem zu datierenden Dokument am ähnlichsten ist. Für jedes Wort-*type*, für das der Zeitstempel des Trainingsdokuments in das Intervall fällt, in dem auch das Wort *bursty* ist, wird ein Punkt vergeben. Diese Wertung wird dann auf die Anzahl der übereinstimmenden *types* normalisiert. Zuletzt wird von den so bewerteten Zeitintervallen das mit der höchsten Wertung, d. h. mit den relativ meisten *types*, die übereinstimmend *bursty* sind ausgewählt und als Zeitstempel vergeben.[162]

Mit Konzepten wie *TE* oder *term burstiness* kann modelliert werden, dass einige Lexeme einer Sprache in ihrer Häufigkeit bzw. Verwendung über viele Jahrhunderte hinweg verhältnismäßig konstant bleiben (*lexical retention*),[163] während sich die Verwendung anderer Wörter ändert oder sie nur in einem bestimmten Zeitraum (gehäuft) auftreten.

Neben der Gewichtung der zur Datierung verwendeten Dimensionen können diese auch reduziert werden, indem weniger relevante oder irrelevante *features* außer Acht gelassen werden. Neben der bereits erwähnten Arbeit von Garcia-Fernandez et. al., die zu diesem Zweck Neologismen und Archaismen als für die Datierung besonders relevante *features* ermitteln, verwenden z. B. auch Kulkarni et al. Neologismus-Informationen zur Gewichtung der Elemente in einer *Bag of Words* (*BoW*) und können so die Anzahl der für eine Datierung benötigten *features* stark reduzieren.[164]

— 8. ***Smoothing* und Interpolation**. Bei der Arbeit mit Sprachmodellen kommen häufig unterschiedliche Methoden zur Glättung (*smoothing*) und Interpolation zum Einsatz.[165] Für die Berechnung von Ähnlichkeitsmaßen wie *KLD* oder *NLLR* ist *smoothing* notwendig, um sogenannte *unseen events*, im Modell fehlende Häufigkeiten, zu schätzen. Für ein Wort *w*, das im zu

156 Siehe Kanhabua und Nørvåg 2008, S. 365.
157 Siehe ebd., S. 368.
158 Siehe Lappas et al. 2009; zitiert in Gumpen und Nygard 2017, S. 15.
159 Siehe Gumpen und Nygard 2017, S. 6.
160 Siehe ebd., S. 25.
161 Es werden Experimente mit mehreren diachronen Datensätzen von Zeitungs- bzw. Online-Nachrichtenartikeln und unterschiedlicher Granularität der Datierung zw. einem Monat und einem Jahr durchgeführt. Dimitrios Kotsakos et al. 2014: „A Burstiness-aware Approach for Document Dating". In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14. Gold Coast, Queensland, Australia: ACM, S. 1003–1006. DOI: 10.1145/2600428.2609495, S. 1005–1006.
162 Siehe Kotsakos et al. 2014, S. 1004; zitiert in Gumpen und Nygard 2017, S. 22.
163 Vgl. Robert B. Lees 1953: „The Basis of Glottochronology". In: *Language* 29.2, S. 113–127. DOI: 10.2307/410164, v. a. S. 124–125; zitiert in Swadesh 1955, S. 122.
164 Siehe Kulkarni et al. 2018, S. 203.
165 Ein ausführlicher Überblick zu diesem Themenkomplex und gängigen Methoden findet sich in Stanley Chen und Joshua Goodman 1998: „An Empirical Study of Smoothing Techniques for Language Modeling". In: *Harvard Computer Science Group Technical Report* 10.

datierenden Dokument *d* und im Korpus *C*, nicht jedoch im Vergleichs-*chronon c* vorkommt, kann zu diesem Zweck eine (geringe) Häufigkeit $P(w \mid c)$ angenommen werden. DE JONG, RODE und HIEMSTRA, die mit *linear interpolation smoothing* (auch JELINEK-MERCER *smoothing*) und DIRICHLET *smoothing*[166] experimentieren, konstatieren, dass „for temporal language models built from large fractions of the reference corpus the smoothing is negligible".[167] Dennoch wird in vielen weiterführenden Studien teils großer Aufwand betrieben, um das *smoothing* zu optimieren bzw. für temporale Modelle gut geeignete Methoden zu finden.[168] Hierbei sollte zwischen unterschiedlichen Herangehensweisen differenziert werden:

— 8.1 Bei häufig eingesetzten Methoden wie *linear interpolation* und DIRICHLET *smoothing* wird die Häufigkeit von *unseen events* anhand der Häufigkeit desselben Wortes im zu datierenden Text und/oder im gesamten Korpus durch Multiplikation mit einem geeigneten Glättungsparameter geschätzt. Dieser muss abhängig von der *Smoothing*-Methode bestimmt bzw. optimiert werden.[169]

— 8.2 Eine sehr einfache Glättungsmethode ist das LAPLACE-*smoothing*.[170] Dabei wird die Häufigkeit aller Vorkommen um eine bestimmte Anzahl, häufig 1 („*add one smoothing*"), erhöht. Dadurch erhält in jedem Vergleichsdokument bzw. *chronon* jedes *type* des Korpus eine Mindesthäufigkeit von 1.

Eine große Gefahr beider Herangehensweisen ist, dass dadurch „important characteristics of a specific time span" „herausgeglättet" werden können.[171]

— 8.3 Andere Methoden ergänzen oder berechnen durch Interpolation Häufigkeiten aus denjenigen desselben Wortes aus anderen, z. B. benachbarten *chronons*.[172] Die Auswirkung eher zufälliger Schwankungen der Worthäufigkeit in den Trainingsdatensätzen, die stärker zufällige Eigenschaften dieser Texte als tatsächliche sprachliche Trends reflektieren, sollen so reduziert werden. BAMMAN et al. nennen als Beispiel hierfür etwa die Häufigkeit von *thee*, die im von ihnen verwendeten Korpus 1717 doppelt so hoch ist wie 1716, obwohl ein klarer Abwärtstrend bei der Häufigkeit dieser Wortform festgestellt werden kann und begegnen dieser Problematik durch die Verwendung eines gleitenden Durchschnitts über 50 Jahre.[173] KANHABUA und NØRVÅG schlagen vor, in einzelnen *chronons* fehlende Häufigkeiten wiederkehrender („recurring") *types*, sofern vorhanden, aus benachbarten *chronons* zu ergänzen, um die zu geringe Größe des Trainingskorpus auszugleichen.[174]

— 9. **Korpora**. Den meisten Studien liegt ein großes diachrones Korpus von bereits datierten Dokumenten[175] zugrunde, das auch Zeitraum, Sprache und Genre der datierbaren Texte vorgibt. Verwendet wurden hierfür etwa die Bücher des *HathiTrust*,[176] Romane aus dem *Project Gutenberg*,[177]

166 Siehe dazu Kapitel 6.1.1, S. 164.
167 DE JONG, RODE und HIEMSTRA 2005, S. 3.
168 Siehe z. B. A. KUMAR 2013, S. 22–23, S. 36–39.
169 Ausführlicher dazu siehe Kapitel 6.1.1, ab S. 164. Siehe z. B. auch A. KUMAR et al. 2012, S. 7–8; u. CHAMBERS 2012, S. 103.
170 Siehe dazu Kapitel 6.1.1, S. 164.
171 DE JONG, RODE und HIEMSTRA 2005, S. 3.
172 Siehe v. a. KANHABUA und NØRVÅG 2008, S. 362–363; BAMMAN et al. 2017, S. 5.
173 Siehe BAMMAN et al. 2017, S. 5.
174 Siehe KANHABUA und NØRVÅG 2008, S. 363.
175 Siehe z. B. DE JONG, RODE und HIEMSTRA 2005.
176 Siehe z. B. BAMMAN et al. 2017; GUO Siyuan et al. 2015.
177 Siehe z. B. A. KUMAR 2013.

Artikel aus der *Wikipedia*[178], Zeitungsartikel[179] oder diachrone Daten über Worthäufigkeiten wie *Google n-Grams*,[180] die bereits implizit vorberechnete Sprachmodelle für einzelne Jahre sind. Lediglich „nicht-lernende" Methoden, z. B. *temporal tagging*, kommen ohne solche Daten aus.[181]

Die verwendete Datenbasis bzw. die zugrunde liegenden Korpora geben dabei stets vor, welche Textsorte(n) und vor allem welcher mögliche Zeitraum in der Datierung abgedeckt werden. Diese **Zeiträume** sind damit sehr unterschiedlich und bewegen sich in Reichweiten zwischen wenigen Jahren[182] und einigen Jahrhunderten.[183] Typische Betrachtungszeiträume für Zeitungsartikel oder Webseiten sind wenige Jahre vor Veröffentlichung der jeweiligen Studie,[184] für Belletristik oder genreübergreifende Studien, die auf Online-Bibliotheken wie *Google Books*[185] oder dem *HathiTrust*[186] basieren, die vergangenen 2–5 Jahrhunderte. Typisch ist entsprechend auch die Spezialisierung der Datierung auf ein bestimmtes Textgenre, z. B. Nachrichten[187] oder Kurzgeschichten.[188] Wie beim abgedeckten Zeitraum richten sich die Möglichkeiten hier nach den verfügbaren Trainingsdaten.[189]

— 10. **Verwendung von Namen**. Eine weitere Idee, die von unterschiedlichen Autor:innen verfolgt wird, ist die Verwendung von *Named Entity Recognition* (*NER*).[190] Wie bei Neologismen ist die Grundidee auch hier, dass die Erwähnung des Namens einer Person *vor* ihrer Geburt sehr unwahrscheinlich ist bzw. ausgeschlossen werden kann, sofern nicht mehrere Personen des gleichen Namens nachgewiesen sind. Als Datenquelle hierfür eignen sich z. B. die Biographien einzelner Personen in der *Wikipedia*,[191] sowie die ebenfalls häufig innerhalb von *Wikipedia* verfügbaren Übersichtsseiten, auf denen Personen nach deren Geburtsjahr gelistet werden (*born in...*, *geboren...*, *naissance en…*, *... nian chusheng* 年出生).[192]

Für das Chinesische steht mit der *CBDB* eine frei nutzbare Datenbank mit biographischen Daten zu mehr als 360.000 Personen der chinesischen Geschichte zur Verfügung.[193] Die Verwendung der darin enthaltenen Namen für die Textdatierung ist aber nicht unproblematisch.[194]

— 11. **Verwendung von Metadaten**. Um in digitalen Bibliotheken wie dem *HathiTrust* oder *Google Books* das *creation date* von Texten zu ermitteln, verwenden Bamman et al. auch „a simple heuristic of metadata-based deduplication"[195]. Dabei werden später datierte Ausgaben desselben Werkes auf die Erstveröffentlichung datiert – auf Basis der Übereinstimmung von Titel und Autor.[196] Mit

178 Siehe z. B. A. Kumar 2013.
179 Siehe z. B. de Jong, Rode und Hiemstra 2005; Kotsakos et al. 2014.
180 Bamman et al. 2017; Kulkarni et al. 2018.
181 Vgl. auch Kanhabua und Nørvåg 2008, S. 359.
182 Vgl. z. B. Kanhabua und Nørvåg 2008; Kotsakos et al. 2014.
183 Vgl. z. B. A. Kumar et al. 2012; Bamman et al. 2017.
184 Vgl. z. B. Kanhabua und Nørvåg 2008, S. 366.
185 Garcia-Fernandez et al. 2011.
186 Guo Siyuan et al. 2015.
187 de Jong, Rode und Hiemstra 2005; Kotsakos et al. 2014; Gumpen und Nygard 2017.
188 A. Kumar et al. 2012.
189 de Jong, Rode und Hiemstra 2005, S. 4.
190 Siehe z. B. Garcia-Fernandez et al. 2011, S. 4–5; Kulkarni et al. 2018, S. 202.
191 Siehe z. B. A. Kumar et al. 2012, S. 3; siehe auch A. Kumar 2013, S. 13.
192 Siehe z. B. Garcia-Fernandez et al. 2011, S. 4.
193 *CBDB*.
194 Ausführlicher dazu siehe Kapitel 4.7, ab S. 97.
195 Bamman et al. 2017, S. 1, siehe auch S. 4.
196 Siehe ebd., S. 4. Für Fälle, bei denen das nicht klappt, da etwa der Titel *The life and adventures of Robinson Crusoe* nicht mit *Robinson Crusoe* übereinstimmt, werden die Vorkommen der 500 häufigsten Trigramme vergleichen und auf dieser Basis eine „content based deduplication" vorgenommen.

dieser „Metadata-Dedup“ genannten Methode können die Autoren mehr *HathiTrust*-Texte korrekt datieren als mit statistischen Sprachmodellen.[197] Während dies für literarische Texte relativ unproblematisch sein mag, kann diese Herangehensweise allerdings bei umfangreichen Revisionen z. B. von Sachliteratur, zu Fehldatierungen führen.

Fast alle bekannten bisher veröffentlichten Arbeiten zur inhaltsbasierten Textdatierung befassen sich mit westlichen Sprachen, u. a. Englisch,[198] Französisch,[199] Niederländisch,[200] Polnisch,[201] Portugiesisch,[202] und Irisch.[203] Lediglich die temporale Tagging-Software *HeidelTime* ist für die Verwendung mit unterschiedlichen, teils auch ostasiatischen Sprachen ausgelegt.[204] Grundsätzlich spricht nichts gegen die Übertragbarkeit der beschriebenen Ansätze und Methoden auf die Verwendung mit außereuropäischen Sprachen. Die Voraussetzung dafür ist die Verfügbarkeit passender diachroner Korpora bzw. Datensätze.

Eine Herausforderung bei der Datierung chinesischsprachiger Texte ist das Fehlen klarer Wortgrenzen.[205] Zudem sind durch die Rigidität der chinesischen Schrift orthographische Änderungen, die Datierungsaufgaben sonst zuträglich sind,[206] auf ein absolutes Minimum reduziert. Gerade bei digitalen, in der Regel normalisierten Ausgaben, in denen die – über die Jahrhunderte durchaus vorhandenen – Änderungen am graphischen Stil der Schrift nicht sichtbar werden, lassen sich derartige Veränderungen nicht nutzbar machen. Vergleichbar ist hier lediglich die Einführung der Kurzzeichen (*jiantizi* 簡 [简] 體 [体] 字) ab 1956. Da jedoch *alle* digitalen Versionen nach 1956 erstellt wurden und in der Volksrepublik auch ältere Texte oft in Kurzzeichen wiedergegeben werden, eignet sich auch die Berücksichtigung reformierter Zeichen kaum für Datierungszwecke.[207]

Die für die hier beschriebenen Methoden und Aspekte der Textdatierung benötigten computerlinguistischen Grundlagen, v. a. die Berechnung von Worthäufigkeiten zur Erzeugung von *Bag of Words*-Repräsentationen, die Verfügbarkeit von (diachronen) Korpora, sowie die Erkennung von Personennamen und *temporal expressions* für schriftsprachliches Chinesisch werden im folgenden Kapitel erörtert.

197 Siehe ebd., S. 1, siehe auch S. 4.
198 Kanhabua und Nørvåg 2008; Guo Siyuan et al. 2015; Gumpen und Nygard 2017; Bamman et al. 2017.
199 Garcia-Fernandez et al. 2011.
200 de Jong, Rode und Hiemstra 2005.
201 Graliński et al. 2017.
202 Zampieri, Malmasi und Dras 2016.
203 Toner und Han Xiwu 2019.
204 Strötgen und Gertz 2010.
205 Siehe auch Abschnitt 4.4, ab S. 73.
206 Siehe Garcia-Fernandez et al. 2011, S. 7; siehe auch Graliński et al. 2017, S. 32.
207 Da die Reformen in mehreren Schritten erfolgt sind, die teilweise rückgängig gemacht wurden, können vereinfachte Zeichen wie *xue* 彐 für 雪 bei gedruckten Ausgaben oder Scans derselben u. U. sogar eine sehr genaue zeitliche Einordnung ermöglichen. Vgl. auch John DeFrancis 1984: *The Chinese Language – Fact and Fantasy*. Honolulu: University of Hawaií Press, S. 261.

4 Computerlinguistische Methoden für schriftsprachliches Chinesisch

> *„Method and even the necessity of classical Chinese word segmentation is still an open question."*[1]
>
> Dong Yubing

Bag of Words (*BoW*)-Modelle oder vergleichbare Repräsentationen von Texten bilden in der Regel die Grundlage für die in Kapitel 3 vorgestellten computerlinguistischen Datierungsmethoden.[2] Voraussetzung dafür ist die Segmentierung bzw. Tokenisierung der so betrachteten Texte, d. h. die Zerlegung in einzelne Wörter oder Lexeme bzw. *tokens*, damit Worthäufigkeiten berechnet werden können. Während eine genaue Tokenisierung von Texten für die meisten westlichen Sprachen mit heute selbstverständlichen, computerlinguistischen Werkzeugen durchgeführt werden kann, stehen diese für schriftsprachliches Chinesisch nicht in geeigneter Weise zur Verfügung. In diesem Kapitel wird ein Blick auf die relevante Forschungslandschaft geworfen. Besondere Aufmerksamkeit verdienen das Segmentieren klassischer bzw. schriftsprachlicher chinesischer Texte und Alternativen dazu. Ebenfalls relevant für die zeitliche Einordnung von Texten sind die Erkennung von z. B. Personen- oder Ortsnamen (*Named Entity Recognition*, *NER*) (Kapitel 4.7, ab S. 97), sowie von *temporal expressions*, also Phrasen, die Zeitangaben enthalten (Kapitel 4.8, ab S. 103).

Für das Training von Tokenizern interessant ist zunächst die Verfügbarkeit schriftsprachlicher Textkorpora, die auch für die Evaluation von Datierungsmethoden unverzichtbar sind. Aufgrund des Mangels an diachronen Korpora werden zudem Alternativen diskutiert (Kapitel 4.2, ab S. 62).

Die Notwendigkeit der Tokenisierung chinesischsprachiger Texte kann auch infrage gestellt werden. Meng Yuxian et al. zeigen, dass sogar für modernes Chinesisch Zeichen- bzw. *n*-Gramm basierte Textrepräsentationen besser für computerlinguistische Methoden geeignet sein können als wortbasierte.[3] Dass reine *n*-Gramm Repräsentationen einer klassischen *BoW* vorzuziehen sind, gilt allerdings nicht uneingeschränkt für die Datierung von Texten.[4]

Ein Test der für das Chinesische verfügbaren Segmentierungssoftware im Hinblick auf ihre Eignung für historische Entwicklungsstufen der chinesischen Schriftsprache (Kapitel 4.5, ab S. 80) legt ebenfalls nahe, sich nicht auf vorhandene Tokenizer zu verlassen, sondern die zu verwendenden *types* und *tokens* durch eine *n*-Gramm-Zerlegung zu bestimmen. In Abschnitt 4.5.2 (ab S. 91) wird daher kurz auf die *n*-Gramm Tokenisierung in *Python* eingegangen, die im Rah-

1 Dong Yubing 2012: *CSCI 562 Final Project, Building a machine translation system that translates Modern Chinese into Classical Chinese*. GitHub Repository. URL: `https://github.com/tomtung/nlp-class/blob/master/final/report.pdf`, S. 2.

2 Siehe Kapitel 3.3, S. 47.

3 Die Autoren vergleichen dabei die Eignung bei der Verwendung in unterschiedlichen Aufgaben wie Maschinenübersetzung und Klassifizierung von Texten. Siehe Meng Yuxian et al. 2019.

4 Entsprechende Versuche werden in Kapitel 6.1 (ab S. 156) durchgeführt.

Veröffentlicht in: Tilman Schalmey: *Computerlinguistische Datierung schriftsprachlicher chinesischer Texte*. Heidelberg: HASP, 2023. DOI: https://doi.org/10.11588/hasp.1153

men dieser Arbeit zur Anwendung kommt. In Abschnitt 4.5.3 (ab S. 94) wird ein Kompromiss zwischen Segmentierung und *n*-Gramm-Zerlegung vorgeschlagen. Als Vorstufe der Zerlegung der zu untersuchenden Texte in einzelne *n*-Gramme, Wörter oder Lexeme kann die Standardisierung bzw. Normalisierung der verwendeten digitalen Ausgaben gesehen werden, die in Kapitel 4.3 (ab S. 69) diskutiert wird.

4.1 Forschungslandschaft

Für die moderne chinesische Hochsprache wird die Entwicklung von Methoden für computerlinguistische Anwendungen wie Tokenisierung, *Part-of-Speech* (*PoS*) *Tagging* und *NER* seit Ende der 1980er Jahre stark vorangetrieben.[5] Dabei werden für komplexe Anwendungen wie z. B. Maschinenübersetzung Ergebnisse erzielt, die vergleichbar mit denen für europäische Sprachen sind.[6]

Innerhalb der internationalen *Association for Computational Linguistics*[7] hat sich mit *SIGHAN* eine eigene *special interest group* gebildet, die sich den besonderen Herausforderungen bei der computerlinguistischen Verarbeitung chinesischer Sprache widmet.[8] Ein Team von Wissenschaftler:innen und Softwareentwickler:innen an der Academia Sinica (*Zhongyang yanjiu yuan* 中央研究院) in Taipeh 台北, die CKIP (*Chinese Knowledge and Information Processing*)-Gruppe (*[Zhongwen] ci [zhishi] ku xiaozu* [中文] 詞 [知識] 庫小組) ist mit der Entwicklung von *Natural Language Processing* (*NLP*)-Tools befasst, die ständig verbessert und erweitert werden und seit kurzem teilweise auch *Open Source* verfügbar sind.[9]

Während am Schnittpunkt zwischen moderner chinesischer Sprachwissenschaft und Computerlinguistik in den vergangenen Jahrzehnten ein stetig wachsendes Forschungsfeld entstanden ist, gibt es immer noch vergleichsweise wenig computerlinguistische Arbeiten zu schriftsprachlichem bzw. klassischem Chinesisch. Eine Ursache ist sicherlich das geringere kommerzielle Interesse an älteren Entwicklungsstufen der Sprache. In diesem Umfeld besteht auch ein Mangel an klassischen oder sogar diachronen Korpora, die in der Regel zum Training von Werkzeugen wie Tokenizern oder *PoS-Taggern* eingesetzt werden. Zudem war um die Jahrtausendwende noch die Annahme verbreitet, Klassisches Chinesisch sei „too difficult to

5 Siehe z. B. Huang Chu-ren 黃居仁, Tokunaga Takenobu 德永健伸 und Sophia Yat Mei Lee 2006: „Asian language processing: current state-of-the-art". In: *Language Resources & Evaluation* 30, S. 203–218, S. 205. Japanische Wissenschaftler:innen beschäftigten sich sogar bereits während der 1960er Jahre mit *Natural Language Processing* für asiatische Sprachen.

6 Für einen zusammenfassenden Überblick zu Sprachressourcen und Tools, siehe Huang Chu-ren 黃居仁 und Xue Nianwen 2019: „Digital Language Resources and NLP tools". In: *The Routledge Handbook of Chinese Applied Linguistics*. Hrsg. von Huang Chu-ren 黃居仁, Zhuo Jing-Schmidt und Barbara Meisterernst. Abingdon, Oxon & New York: Routledge, S. 461–482; Siehe z. B. auch Huang Chu-ren 黃居仁, Tokunaga Takenobu 德永健伸 und S. Y. M. Lee 2006, S. 204; Für einen frühen Überblick zu computerlinguistischen Arbeiten zur chinesischen Sprache empfehlen sich als Orientierung die Bibliographien von Cornelia Schindelin, in welchen gängige statistische Untersuchungen für die moderne Hochsprache zusammengefasst sind. Siehe Cornelia Schindelin 2005b: „Zur Geschichte quantitativ-linguistischer Forschungen in China". In: *Quantitative Linguistik – Quantitative Linguistics – An International Handbook / Ein internationales Handbuch*. Hrsg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. Berlin & New York: Walter de Gruyter, S. 96–115, S. 113–115; Cornelia Schindelin 2005a: „Die quantitative Erforschung der chinesischen Sprache und Schrift". In: *Quantitative Linguistik – Quantitative Linguistics – An International Handbook / Ein internationales Handbuch*. Hrsg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. Berlin & New York: Walter de Gruyter, S. 947–970, S. 968–970.

7 Association for Computational Linguistics: *Association for Computational Linguistics*. Website. URL: https://www.aclweb.org/ (besucht am 29. 09. 2018).

8 SIGHAN 2005–: *SIGHAN Home Page*. URL: http://sighan.cs.uchicago.edu/ (besucht am 29. 09. 2018).

9 Siehe CKIP Lab 2020: *CKIP Lab*. Website. URL: https://ckip.iis.sinica.edu.tw/ (besucht am 21. 05. 2021), siehe auch S. 81.

process".[10] In den vergangenen Jahren sind aber, begünstigt durch gezielte Förderung der *Digital Humanities*, sowie durch das große Engagement oft einzelner *Aficionados*[11], beeindruckende Projekte entstanden, die ein breites Spektrum abdecken.

Neben Plattformen, die Wissenschaftler:innen bei der Analyse oder Lektüre von chinesischen Texten unterstützen sollen, z. B. *MARKUS. Text Analysis and Reading Platform*,[12] oder die umfangreiche Such- und Statistikfunktionen bereitstellen und dabei weitere, externe Ressourcen über *Application Programming Interfaces* (*APIs*)[13] einbinden, wie etwa die *Academia Sinica Digital Humanities Research Platform (Zhongyang yanjiuyuan shuwei renwen yanjiu pingtai* 中央研究院數位人文研究平台)[14] und Online-Textsammlungen wie das *Chinese Text Project*[15] und *A Database of Mediaval Chinese Texts*[16] zeugen Gründungen von *Journals* wie *Asiascape: Digital Asia*, *Shuzi Renwen* 数字人文 (*Digital Humanities*) oder dem *Digital Orientalist*[17] von einer wachsenden Community. Eine steigende Zahl an Einzelarbeiten zu grundlegenden Methoden oder speziellen Forschungsfragen befassen sich vermehrt gerade mit klassischen bzw. schriftsprachlichen Texten, da urheberrechtliche Themen in diesem Kontext eine geringere Rolle spielen.[18] Die Vielfalt der darin bearbeiteten Bereiche kann im Folgenden nur angedeutet werden.[19]

Aus dem Bereich der Stilometrie sei an dieser Stelle nochmals der Aufsatz „Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature" von Paul Vierthaler erwähnt, der stilistische Unterschiede zwischen literarischen, als *xiaoshuo* 小說 eingestuften, und zwei Gattungen von historischen Texten, *yeshi* 野史 (inoffizielle Geschichten) und offizielle Dynastiegeschichten (*zhengshi* 正史), untersucht.[20] Statt sich mit der problematischen Segmentierung der Texte aufzuhalten, kommt Vierthaler ebenfalls mit einer Zerlegung der untersuchten Texte in *n*-Gramme und den daraus erzeugten Häufigkeitslisten zu überzeugenden Ergebnissen.[21] In einer Arbeit zu *Topic modelling* für *Lunyu* 論語, *Xunzi* 荀子 und *Mengzi* 孟子, stellen Nichols et al. fest, dass diese stilistischen *topics* potenziell unter anderem

10 Huang Liang et al. 2002a: „PCFG Parsing for Restricted Classical Chinese Texts". In: *COLING-02: The First SIGHAN Workshop on Chinese Language Processing*. ACL Anthology. DOI: 10.3115/1118824.1118830, S. 6.

11 Die Verwendung des Begriffes in diesem Kontext ist von Christoph Harbsmeier übernommen.

12 Hou-Ieong Brent Ho und Hilde DeWeerdt. 2014–: *MARKUS. Text Analysis and Reading Platform*. URL: https://dh.chinese-empires.eu/beta/.

13 *APIs* sind Programmierschnittstellen, die genutzt werden können, um Softwaresysteme miteinander zu verbinden, etwa um Funktionen, die eine Software zur Datenverarbeitung oder Transformation bereitstellt, zu nutzen und das Ergebnis an anderer Stelle zu verwenden, oder um zur Verfügung gestellte Inhalte in einem geeigneten Format abzurufen.

14 Academia Sinica, Center for Digital Cultures 中央研究院數位文化中心 2018: *Academia Sinica Digital Humanities Research Platform (Zhongyang yanjiuyuan shuwei renwen yanjiu pingtai* 中央研究院數位人文研究平台*)*. URL: https://dh.ascdc.sinica.edu.tw/ (besucht am 24. 01. 2021).

15 Donald Sturgeon, Hrsg. 2011: *Chinese Text Project*. URL: https://ctext.org (besucht am 24. 04. 2021); siehe auch Donald Sturgeon 2019: „Chinese Text Project: a dynamic digital library of premodern Chinese". In: *Digital Scholarship in the Humanities* 0.0, S. 1–12. DOI: 10.1093/llc/fqz046.

16 Christoph Anderl et al. 2015–: *A Database of Mediaval Chinese Texts*. URL: https://www.database-of-medieval-chinese-texts.be/ (besucht am 30. 04. 2021).

17 Siehe Florian Schneider, Hrsg. 2014–: *Asiascape: Digital Asia*. URL: https://brill.com/view/journals/dias/dias-overview.xml (besucht am 29. 05. 2021); Cheng Yizhong 程毅中 et al., Hrsg. 2020–: *Shuzi renwen* 数字人文 *Digital Humanities*. Beijing 北京: Zhonghua shuju 中華書局; Cornelis van Lit et al., Hrsg. 2015–: *Digital Orientalist, The*. URL: https://digitalorientalist.com/ (besucht am 29. 05. 2021).

18 Vgl. auch Vierthaler 2020, S. 8.

19 Für einen aktuellen Überblick siehe aber Vierthaler 2020, *passim*. Einen prägnanten Überblick über Plattformen, Ressourcen und Tools der chinesischen *DH* gibt Peter K. Bol 2020: „Introduction to the Utilities". In: *Journal of Chinese History* 4.2, S. 483–486. DOI: 10.1017/jch.2020.10.

20 Vierthaler 2016a, Siehe auch Kapitel 2.3, ab S. 20.

21 Siehe Vierthaler 2016a, S. 7; das zugehörige Kurzzeichen-Textkorpus ist online abrufbar. Siehe Paul Vierthaler 2016b: *Late Imperial Chinese Texts: The Corpus for Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature*. DOI: 10.7910/DVN/GDYFAG. URL: https://doi.org/10.7910/DVN/GDYFAG.

zur Untersuchung der Überlieferungsgeschichte, zur Datierung und zur Klassifizierung von Texten eingesetzt werden können.[22] Sie schlagen vor zu „untersuchen, ob die in der Sekundärliteratur vertretenen Meinungen über die relative Datierung der *Shangshu* [尚書] Kapitel auf der Grundlage ihrer sprachlichen Ähnlichkeit bestätigt werden können."[23] Entsprechende Ergebnisse werden bisher aber nicht vorgelegt.[24] Eine Arbeit zu Maschinenübersetzung vom Modernen ins Klassische Chinesische,[25] sowie die Entwicklung einer Programmiersprache, die Klassisch-Chinesische Befehle verwendet,[26] sind weitere Beispiele für den Enthusiasmus einzelner Forscher:innen für die Verbindung der klassischen Philologie und der Informatik.

4.2 Korpora

Sprachkorpora werden für zahlreiche computerlinguistische Anwendungen eingesetzt. Je nach Einsatzgebiet können sie breit angelegt, oder auf ein bestimmtes Textgenre, eine Epoche oder z. B. gesprochene Sprache beschränkt sein.[27] Bestenfalls sind sie gut dokumentiert und liegen in einem annotierten und segmentierten Format vor.

Bei der Digitalisierung älterer Texte kann zwischen einer diplomatischen, einer normalisierten und einer modernisierten Transkription unterschieden werden. Die diplomatische Transkription versucht – insbesondere bei Handschriften –, möglichst viele Eigenschaften des Urtextes zu bewahren bzw. kenntlich zu machen. Bei normalisierten Versionen werden in der Regel typo- und paläographische Eigenheiten, sowie das ursprüngliche Layout, außer Acht gelassen. Bei einer modernisierten Fassung fehlen zumeist alle Eigenschaften des Urtextes – abgesehen von den eigentlichen Wörtern.[28]

In der Korpuslinguistik ist der *XML*-Standard der Text Encoding Initiative (TEI) verbreitet. Er erlaubt die Erstellung von Transkriptionen, die in unterschiedlichen Abstufungen zwischen diplomatischer und normalisierter Version des Textes verwendet werden können.[29]

22 Siehe Nichols et al. 2018, S. 39.

23 Ebd., S. 23, übersetzt durch den Verfasser.

24 Siehe Nichols et al. 2018, S. 23; In einer früheren Studie kündigen die Autoren für die zitierte Studie hingegen voreilig an, man habe „successfully reproduced the basic scholarly consensus concerning the dating of individual chapters of the text (demonstrating the reliability of the technique), but suggested areas in which the consensus might be wrong (adding to our scholarly knowledge)." Kristoffer Nielbo, Ryan Nichols und Edward Slingerland 2018: „Mining the Past – Data-Intensive Knowledge Discovery in the Study of Historical Textual Traditions". In: *Journal of Cognitive Historiography* 3.1–2, S. 93–118. DOI: 10.1558/jch.31662, S. 115.

25 Dong Yubing 2012.

26 Siehe Charles Q. Choi 2020: *World's First Classical Chinese Programming Language*. URL: https://spectrum.ieee.org/tech-talk/computing/software/classical-chinese (besucht am 12. 09. 2020), Den Hinweis auf die Existenz der klassisch-chinesischen Programmiersprache *wenyan-lang* verdanke ich Christian Soffel. Ein weiteres Beispiel ist *Zhongshuyu* 中書玲 (*PerlYuYan*) von Tang Feng 唐鳳, womit sich *Perl*-Anwendungen mit einer klassisch-chinesischen Syntax schreiben lassen. Siehe Tang Feng 唐鳳 2009: *Lingua::Sinica::PerlYuYan – Perl in Classical Chinese in Perl – Zhongshuyu* 中書玲. URL: https://metacpan.org/pod/release/AUDREYT/Lingua-Sinica-PerlYuYan-1257340475/lib/Lingua/Sinica/PerlYuYan.pm (besucht am 12. 09. 2020).

27 Siehe z. B. Anatol Stefanowitsch 2020: *Corpus Linguistics: A Guide to the Methodology*. Textbooks in Language Sciences 7. Berlin: Language Science Press. DOI: 10.5281/zenodo.3735822, S. 22–23: „In corpus linguistics, the term […] refers to a collection of samples of language use with the following properties: […] *authentic*, […] *representative* of the language or language variety under investigation [and] […] *large*."

28 Matthew J. Driscoll 2007: *Electronic Textual Editing: Levels of transcription*. URL: http://www.tei-c.org/Vault/ETE/Preview/driscoll.html (besucht am 25. 09. 2018); Dieser Unterscheidung liegt Trennung zwischen dem Inhalt eines Texts (*substantives*) und seinen „formalen" Eigenschaften (*accidentals*) zugrunde. Siehe Walter W. Greg 1950/51: „The Rationale of Copy-Text". In: *Studies in Bibliography* 3, S. 19–36. URL: https://www.jstor.org/stable/40381874, S. 21.

29 TEI Consortium 2019: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Website. URL: http://www.tei-c.org/Guidelines/P5/ (besucht am 07. 05. 2019).

Digitalisierte Fassungen schriftsprachlicher chinesischer Texte sind leider kaum in dieser Qualität verfügbar. Ausnahmen bilden einzelne buddhistische Texte, die im Rahmen der *Database of Medieval Chinese Texts* veröffentlicht werden.[30]

Ein wichtiger Verwendungszweck von Korpora ist das Trainieren von Tools im Rahmen von *Natural Language Processing* (*NLP*), gerade für die Tokenisierung und vor allem *PoS-Tagging*, wozu auch das Erkennen von Namen (*Named Entity Recognition*, *NER*) gezählt werden kann, aber auch für komplexere Anwendungen wie Maschinenübersetzung.

Für die moderne Hochsprache steht mit der *Chinese Treebank 9.0* (*CTB*)[31] ein Korpus zur Verfügung, mit dem z. B. der *Stanford Word Segmenter* trainiert wurde.[32] Es wurde 1998 an der University of Pennsylvania begonnen und enthält in der aktuellen Version inzwischen über 3.000 Dokumente, etwa 2 Mio. Wörter bzw. 3 Mio. Kurzzeichen. Abgedeckt werden darin unterschiedliche Genres: Zeitungsartikel, Meldungen von Nachrichtenagenturen wie *Xinhua* 新華, Artikel aus Zeitschriften, Regierungsdokumente sowie auch transkribierte Telefongespräche als Beispiele für gesprochene Sprache. Erstere werden hier für die Tokenizer-Teste als modernes Vergleichsmaterial herangezogen.[33]

Verfügbarkeit schriftsprachlicher Korpora

Wie eingangs erwähnt, sind umfangreiche diachrone Korpora für die Evaluation computerlinguistischer Textdatierungsmethoden unabdingbar. Um eine entsprechende Aufarbeitung schriftsprachlicher Texte ist es jedoch deutlich schlechter bestellt. Die wenigen vorhandenen, frei zugänglichen, segmentierten Korpora sind sehr klein. Das sogenannte *Ancient Chinese Corpus* enthält lediglich Teile des *Zuozhuan* 左傳 mit Segmentierung und *PoS*-Tags.[34] Das *Sheffield Corpus of Chinese* wurde immerhin tatsächlich aus antiken, mittelalterlichen und spätkaiserzeitlichen Texten zusammengestellt, um Erkenntnisse über die diachrone Entwicklung der chinesischen Sprache zu gewinnen.[35] Leider ist es nur noch fragmentarisch abrufbar.[36]

Darüber hinaus existieren zwei Korpora der Academia Sinica, die online als einzelne Abschnitte einseh- bzw. durchsuchbar sind. Eine Downloadmöglichkeit wird leider nicht öffentlich

30 Siehe Marcus Bingenheimer und Zhang Boyong 張伯雍 2017: *XML Data for „Four Early Chan Texts from Dunhuang - A TEI-based Edition“*. XML-Datensatz. Version Dec 2017. DOI: 10.5281/zenodo.1133490. (Besucht am 30. 04. 2021); siehe dazu auch Christoph Anderl 2020: „Some Reflections on the Database of Medieval Chinese Texts as a Multi-Purpose Tool for Research, Teaching, and International Collaboration“. In: *Corpus-Based Research on Chinese Language and Linguistics*. Hrsg. von Bianca Basciano, Franco Gatti und Anna Morbiato. Sinica venetiana 6. Venezia: Edizioni Ca'Foscarini, S. 339–358. DOI: 10.30687/978-88-6969-406-6/011; bzw. Anderl et al. 2015–.

31 Xue Nianwen et al. 2016: *Chinese Treebank 9.0*. URL: https://catalog.ldc.upenn.edu/LDC2016T13 (besucht am 15. 06. 2016); Eine ausführliche Beschreibung der Konzeption der ersten Version findet sich in Xue Nianwen et al. 2005: „The Penn Chinese TreeBank: Phrase structure annotation of a large corpus“. In: *Natural Language Engineering* 11.2, S. 207–238.

32 Siehe Stanford Natural Language Processing Group 2015: *Stanford Word Segmenter*. URL: http://nlp.stanford.edu/software/segmenter.shtml (besucht am 14. 01. 2016), siehe dazu auch Abschnitt 4.5, ab S. 89.

33 Siehe Abschnitt 4.5, ab S. 79.

34 Chen Xiaohe et al. 2017: *Ancient Chinese Corpus*. URL: https://catalog.ldc.upenn.edu/LDC2017T14 (besucht am 18. 10. 2017).

35 Siehe Hu Xiaoling, Williamson und McLaughlin 2005.

36 Siehe Hu Xiaoling, Nigel Williamson und Jamie McLaughlin 2004: *Sheffield Corpus of Chinese*. DOI: 10.1093/llc/fqi034. URL: http://purl.ox.ac.uk/ota/2481 (besucht am 09. 02. 2019), Eine Anfrage nach dem Verbleib der restlichen Korpusdaten beim Verantwortlichen des *Oxford Text Archive*, Martin Wynn, lässt darauf schließen, dass klassischen bzw. schriftsprachlichen chinesischen Korpora nur geringe Bedeutung beigemessen wird: „I'm afraid that we don't have anything more, and I don't know what has happened to the rest of the corpus.“ (pers. Kommunikation, 11.02.2019).

gemacht.[37] Das ab 1990 zusammengestellte und ab 1995 segmentierte und getaggte klassische *Academia Sinica Ancient Chinese Corpus* 中央研究院古漢語標記語料庫 enthält 48 Texte, darunter frühe Kanonklassiker wie *Shangshu* 尚書, *Shijing* 詩經 und *Zhou yi* 周易, philosophische Klassiker wie *Lunyu* 論語, *Mengzi* 孟子 und *Zhuangzi* 莊子, sowie teilweise deutlich umfangreichere Han 漢-zeitliche Texte wie das *Shiji* 史記.[38] Der online verfügbare Teil des *Academia Sinica Tagged Corpus of Early Mandarin Chinese* 中央研究院近代漢語語料庫 enthält primär Ming 明- und Qing 清-zeitliche Romane. Der älteste enthaltene Text ist das *Zutang ji* 祖堂集 („Anthologie der Ahnenhalle"), eine buddhistische Gesprächssammlung aus dem 10. Jh.

Ungeachtet der eingeschränkten Zugänglichkeit reichen weder der abgedeckte Zeitraum, noch die Anzahl der für unterschiedliche Zeiträume verfügbaren Texte der erwähnten Korpora aus, um etwa für die gesamte schriftsprachliche Texttradition statistische Sprachmodelle für die Datierung zu berechnen.[39] Die verfügbaren Textabschnitte eignen sich aber als diachrone Goldstandard-Beispieltexte zum Test verschiedener Tokenizer mit klassischem bzw. schriftsprachlichem Textmaterial.[40]

Korpora in dieser Studie

Um in der gegebenen Situation in einem für die Evaluation von Datierungsmethoden angemessenen Umfang Textmaterial aus unterschiedlichen Zeiträumen zur Verfügung zu haben, muss auf Alternativen zurückgegriffen werden. Zu diesem Zweck werden daher unterschiedliche Behelfskorpora eingesetzt. Dazu gehören:

1. Aus unterschiedlichen Quellen zusammengestellte digitalisierte Texte, die lediglich als unsegmentierter *Plain Text* zur Verfügung stehen,
2. die im *DHYDCD* enthaltenen Beispielsätze, ebenfalls als *Plain Text*, sowie
3. von CrossAsia bereitgestellte Datensätze mit n-Gramm Häufigkeiten.

Besonderheiten, Vorzüge und Schwächen dieser eher provisorischen diachronen Korpora werden in diesem Abschnitt diskutiert. Tabelle 4.1 gibt zunächst einen Überblick über alle in dieser Studie verwendeten Korpora, ihr Datenformat und den Entstehungszeitraum der jeweils enthaltenen Texte. Zusätzlich werden die Anzahl der verwendbaren Texte und ihr Gesamtumfang in Schriftzeichen angegeben.[41] Die letzte Spalte, „Kapitel", verweist zudem auf Verwendungen des jeweiligen Korpus im Rahmen der vorliegenden Studie.

Plain Text Korpora

Als *Plain Text* werden Datenformate bezeichnet, die lediglich die Zeichen bzw. Wörter eines Textes enthalten. Formatierungen oder andere Zusatzinformationen wie Metadaten oder

37 Huang Chu-ren 黃居仁 et. al. 1990: *Academia Sinica Ancient Chinese Corpus* 中央研究院古漢語標記語料庫. URL: http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh (besucht am 10. 02. 2019); Huang Chu-ren 黃居仁 et. al. 2001: *Academia Sinica Tagged Corpus of Early Mandarin Chinese* 中央研究院近代漢語語料庫. URL: http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/pkiwi/kiwi.sh (besucht am 17. 02. 2019).

38 Eine vollständige Liste ist über http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh einsehbar (besucht am 25.04.2021).

39 Vgl. Kapitel 3.3, ab S. 45.

40 Siehe Kapitel 4.5, ab S. 79.

41 „Verwendbar" bezieht sich auf Auswahlkriterien, die für die Verwendung der Texte z. B. als Trainings- oder Testkorpus berücksichtigt werden. Beispielsweise enthält der vollständige *difangzhi*-Datensatz *n*-Gramm Häufigkeiten zu 11.081 Titeln. 6.914 dieser Lokalmonographien erfüllen dabei die auf S. 67 beschriebenen Anforderungen an die Vollständigkeit und Qualität der Zählungs- und Metadaten.

Tabelle 4.1 Übersicht aller verwendeten Korpora

Korpus	Datentyp	Zeitspanne	# Texte	Mio. Zeichen	Kapitel
Dynastiegeschichten (*zhengshi* 正史)	*Plain Text*	91 v. u. Z.–1928	25	26	2.3; 5.5.4; 6.1; 6.2; 6.3
Loewe	*Plain Text*	ca. 1000 v. u. Z.–110	65	4,9	5.5.4; 6.1; 6.3
DHYDCD Belegstellen	*Plain Text* Fragmente	ca. 1000 v. u. Z.–1992	47.066	11,3	6.1
Lokalchroniken (*difangzhi* 地方誌)	*n*-Gramm Frequenzen	1072–1949	6.914	1.527	5.5.4; 6.1; 6.2; 6.3
Xu xiu si ku quan shu 續修四庫全書	*n*-Gramm Frequenzen	960–1936	2.315	327	6.1; 6.2; 6.3
Academia Sinica *Ancient*	*Plain Text* mit *PoS*-Tags	ca. 1000–6 v. u. Z.	48	nicht bek.	4.5
Academia Sinica *Early Mandarin*	*Plain Text* mit *PoS*-Tags	ca. 9.–1800	20	nicht bek.	4.5
Chinese Treebank 6.0	*XML*	1996–2001	2.036	1,2	4.5

Kommentare können nicht gespeichert werden. Durch den im Vergleich zu getaggten bzw. segmentierten Korpora deutlich geringeren Produktionsaufwand, sind *Plain Text*-Fassungen zahlreicher schriftsprachlicher Texte online frei zugänglich. Vollständige diachrone Korpora mit Metadaten existieren allerdings nicht, so dass die einzelnen Texte aus unterschiedlichen Quellen zusammengestellt und Metadaten manuell ergänzt werden müssen. Die Nachteile so aufgebauter Textsammlungen liegen auf der Hand. Neben allgemeinen Qualitätsproblemen kommt es zu uneinheitlicher Verwendung von Kurz- und Langzeichen sowie von Zeichenvarianten.[42] Auch kann die gerade für Texte mit einer langen Überlieferungsgeschichte oft essenzielle Unterscheidung zwischen Haupttext und später eingefügten Kommentaren teils nicht getroffen werden.[43]

Ein *Plain Text* Korpus der **Dynastiegeschichten** (*zhengshi* 正史) wurde bereits in Kapitel 2.3 verwendet, um Aspekte des Sprachwandels zu untersuchen. Diese Textgattung „offizieller" chinesischer Geschichtsschreibung bildet mit ihrem normierten Charakter eine Tradition, die über einen Zeitraum von etwa 2.000 Jahren gepflegt wurde.[44] Das Textmaterial kann zudem zum Test von Datierungstechniken für eben diesen Zeitraum eingesetzt werden.[45] Für eine Auflistung der enthaltenen Texte, ihrer genauen zeitlichen Einordnung und eine kurze inhaltliche Einführung sei ebenfalls auf Kapitel 2.3 verwiesen.[46] Die hier verwendeten Versionen sind unterschiedlichen Online-Textsammlungen entnommen.[47] Um eine minimale Qualitätssicherung zu gewährleisten, wurden alle Texte stichprobenartig auf Übereinstimmung mit der *Scripta Sinica*-Version der *Zhonghua shuju* 中華書局-Ausgabe geprüft.[48]

42 Siehe dazu auch Kapitel 4.3, ab S. 69.

43 In gedruckten Ausgaben wird diese Unterscheidung in der Regel anhand unterschiedlicher Schriftgrößen ermöglicht. In einem *XML*-Format wäre eine solche Unterscheidung problemlos möglich.

44 Siehe ab S. 20.

45 Siehe Kapitel 6, ab S. 6, v. a. 6.1.3, ab S. 171, sowie 6.3, ab S. 6.3.

46 Siehe v. a. S. 20–23.

47 *Shiji* 史記 und *Han shu* 漢書 aus *Project Gutenberg* 1971–. URL: https://www.gutenberg.org/ (besucht am 07. 12. 2021); *Sanguo zhi* 三國志 und *Hou Han shu* 後漢書 aus *Weiji wenku* 維基文庫 *(Wikisource)* 2003–. Website. URL: https://www.gutenberg.org/; die übrigen 21 Texte von *Wenxue 100* 文學 *100* 2015–. Website. URL: http://www.wenxue100.com/ (besucht am 07. 12. 2021).

48 Siehe Academia Sinica 中央研究院 1984–: *Han ji dianzi wenxian ziliaoku* 漢籍電子文獻資料庫 *(Scripta Sinica database)*. Website. URL: http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm (besucht am 07. 12. 2021).

Neben der bereits erwähnten Untersuchung von Sprachwandel finden die *zhengshi* auch zur Verbesserung der im Rahmen von Kapitel 5.5 erzeugten Lexemdatenbank Verwendung. Hierzu werden automatisiert frühere Belege für im *DHYDCD* zu spät datierte Lexeme bzw. Zeichenkombinationen ergänzt, um eine bessere Datengrundlage für die lexembasierte Datierung zu schaffen.[49] Zudem können sie als Testdaten für die in Kapitel 6 evaluierten Textdatierungsmethoden eingesetzt werden.[50]

Eine Sammlung digitaler Versionen der Texte, die in Michael Loewes *Early Chinese Texts: A Bibliographical Guide* vorgestellt werden, bezeichne ich im folgenden als **Loewe-Korpus**.[51] Die insgesamt knapp fünf Millionen Zeichen dieser 65 für die klassische Periode repräsentativen Texte aus den Anfängen der Schrifttradition bis zum Ende der Han-Zeit (220) eignen sich ebenfalls zur Ergänzung des *DHYDCD* um frühere Belege.[52] Hierbei darf nicht vergessen werden, dass einige der enthaltenen Texte nicht mit befriedigender Genauigkeit datierbar sind, so dass mit ungefähren Zeiträumen gearbeitet werden muss.[53]

Die **Einträge des *DHYDCD*** enthalten neben der zur Datierung der Lexeme verwendeten *Loci classici* weitere *attestations*, die sich ebenfalls zeitlich einordnen lassen. Die Erzeugung eines Behelfskorpus aus diesen Belegstellen bzw. Beispielsätzen wird in Kapitel 5.6 beschrieben.[54] Trotz seines provisorischen Charakters als arbiträr zusammengestelltes und gewichtetes Hypertext-Potpourri[55] hat dieses Behelfskorpus zwei entscheidende Vorteile:[56] Der gesamte Zeitraum von den Anfängen des überlieferten Schrifttums bis ins 20. Jh. wird abgedeckt. Das enthaltene Material deckt dabei ein relativ breites Spektrum an Textgenres ab und kann zur Erzeugung entsprechender temporaler Sprachmodelle verwendet werden.[57]

Datensätze mit *n*-Gramm Häufigkeitslisten

Eine in den *Digital Humanities* relativ neue Erscheinung sind Datensätze mit *n*-Gramm-Häufigkeiten. Im Vergleich zu Volltexten oder sogar annotierten Korpora sind die Möglichkeiten der Verwendung stark eingeschränkt und die Lektüre der Texte als solche unmöglich gemacht. Durch die Abstraktion ist eine Veröffentlichung der Daten aber mit deutlich weniger urheberrechtlichen Bedenken verbunden.[58]

CrossAsia stellt unter anderem *n*-Gramm-Daten von **Lokalchroniken (*Difangzhi* 地方誌, *DFZ*)**, sowie des *Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*) zur Verfügung.[59]

---

49 Siehe v. a. ab S. 134.

50 Siehe v. a. Kapitel 6.1, S. 172 u. 6.2, S. 206. In Kapitel 6.3 (ab S. 210) dienen die *zhengshi* hingegen primär als Trainingsdatensatz.

51 Eine vollständige Liste der Texte und ihrer Quellen findet sich in T. Schalmey 2009, S. 104–106.

52 Siehe Kapitel 5.5.4, S. 134.

53 Vgl. Loewe 1993, S. xi.

54 Siehe ab S. 137; siehe auch Kapitel 5.5, S. 120 und Kapitel 5.5.1, S. 123.

55 Zur Auswahl und Gewichtung der Belege im *HYDCD* siehe Kapitel 5.7, ab S. 138. Ein Beispiel für das so entstandene Material wird in Kapitel 6.1.3 (ab S. 171) wiedergegeben.

56 Ausführlicher siehe Kapitel 5.6, ab S. 137. Die Verwendung von Belegstellen aus diachronen Wörterbüchern wird anhand des *Oxford English Dictionary* diskutiert in Sebastian Hoffmann 2004: „Using the OED quotations database as a corpus – a linguistic appraisal". In: *ICAME* 28, S. 17–30.

57 Siehe Kapitel 6.1, ab S. 156, v. a. 6.1.3, S. 171.

58 Eine ausführliche Diskussion dieses und anderer abgeleiteter Textformate findet sich in Christof Schöch et al. 2020: „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2020_006, Siehe insb. S. 5, S. 15–16, S. 18.

59 CrossAsia, Staatsbibliothek zu Berlin 2019a: *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌). Datenset, Version 0.0.2-20190408. DOI: 10.5281/zenodo.2594596 (im Folgenden zit. als *DFZ*); CrossAsia, Staatsbi-

Erstere sind – wie die zuvor erwähnten *zhengshi* – eine spezialisierte, historiographische Textgattung. Anders als jene sind die *DFZ* nicht an eine bestimmte Herrscherdynastie gebunden. Sie thematisieren historisch relevante Aspekte auf unterschiedlichen lokalen Ebenen, teilweise auch über Dynastiewechsel hinweg. Joseph Dennis definiert sie als einen „cumulative record of a territorial unit published in book format, generally by a local government, and arranged by topics such as topography, institutions, population, taxes, biographies, and literature."[60] Der sehr umfangreiche Datensatz enthält Listen mit den jeweils absoluten Häufigkeiten der Uni-, Bi- und Trigramme von insgesamt 11.081 Texten. Eine Besonderheit, die für eine hohe Qualität der zugrundeliegenden Digitalisierung spricht, besteht darin, dass eine Vielzahl von Variantenzeichen (*yitizi* 異體字) enthalten sind.[61] Begleitend stehen Metadaten mit Informationen über die ursprüngliche Veröffentlichung des jeweiligen Textes, den darin beschriebenen Zeitraum, sowie weiteren bibliographischen Angaben wie Titel, Autor, Provinz und Regierungsdevise zur Verfügung. Damit lässt es sich eingeschränkt sowohl für die Erzeugung und Verwendung temporaler Sprachmodelle, als auch zum Trainieren und Testen von lexikographischen Textdatierungsmethoden nutzen.[62] Die Rohdaten enthalten dabei eine Zeile pro *n*-Gramm im Format

n-Gramm	Anzahl Vorkommen, also z. B.
在山下	84

Dateinamen der Häufigkeitslisten sind über einen 32-stelligen hexadezimalen Primärschlüssel mit den Metadaten verknüpft. Leider weist der Datensatz einige Mängel auf, so dass davon betroffene Texte bei der Verwendung ausgeschlossen werden müssen:

1. Bei knapp 15 % der Texte fehlen die 2-Gramm-Häufigkeiten.[63] Diese Texte werden kategorisch ausgeschlossen.

bliothek zu Berlin 2019b: *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書. Datenset. Version 0.0.1-20190307. DOI: 10.5281/zenodo.2586940 (im Folgenden zit. als *XXSKQS*), die Veröffentlichung erfolgt in Form von Listen mit den 1-, 2- und 3-Grammen und deren absoluten Häufigkeiten pro Text.

60 Joseph Dennis 2015: *Writing, Publishing, and Reading Local Gazetteers in Imperial China, 1100–1700*. Harvard East Asian Monographs 379. Cambridge, MA & London: Harvard University Asia Center, Harvard University Press, S. 1.

61 So findet sich z. B. in 7.978 Texten das Zeichen *li* 歷, in 8.505 *li* 歷, das hier als orthodox angesehen wird. Ebenfalls wird in 5.586 Texten die hier orthodoxe Variante *li* 曆, in 4.029 Texten *li* 曆 geschrieben, in 187 Texten wird zudem das heutige Kurzzeichen *li* 历 genutzt. Zusätzlich findet man in 644 Texten die Variante *li* 厤. Dabei kann ein Zusammenhang mit der Tabuisierung von *li* 曆 vermutet werden. Da Hongli 弘曆 der Name von Kaiser Qianlong 乾隆 (reg. 1736–1796) ist, wurde zu dessen Lebzeiten zur Vermeidung des Zeichens 曆 oft ohne den unteren Bestandteil (*ri* 日), also 厤, geschrieben. Siehe Piotr Adamek 2012: „Good Son is Sad If He Hears the Name of His Father: The Tabooing of Names in China as a Way of Implementing Social Values". Diss. Leiden: Leiden University, S. 287. Siehe auch Kapitel 4.3, ab S. 70. Ob tatsächlich ein Zusammenhang mit dem Namenstabu besteht, bleibt unklar, da die Variante 厤 schon deutlich älter ist und auch in früheren Texten des Korpus Verwendung findet.

62 Siehe Kapitel 6, ab S. 155, v. a. 6.1.1, ab S. 158; siehe auch 6.2.5, S. 197. Beispiele für weitere *DH*-Studien an den Volltexten der chinesischen Lokalchroniken im Rahmen der am Max-Planck-Institut für Wissenschaftsgeschichte entwickelten *LoGaRT*-Plattform werden in Chen Shih-Pei 陳詩沛 et al. 2020: „Local Gazetteers Research Tools: Overview and Research Application". In: *Journal of Chinese History* 4.2, S. 544–558. DOI: doi:10.1017/jch.2020.26, beschrieben; siehe auch Chen Shih-Pei 陳詩沛 et al. 2017: *LoGaRT: Local Gazetteers Research Tools*. Software. Berlin: Max-Planck-Institut für Wissenschaftsgeschichte. URL: https://www.mpiwg-berlin.mpg.de/research/projects/logart-local-gazetteers-research-tools (besucht am 22. 10. 2021), Eine Studie zu zeitlichen Trends in den thematischen Kategorien der *DFZ* von derselben Autorin ist in Arbeit.

63 Von diesem Fehler sind 1.637 Texte betroffen. Im 2-Gramm-Ordner des Datensatzes ist zwar jeweils eine Liste enthalten, diese enthält jedoch nur die Vorkommen der Einzelzeichen. Auch in der Version 0.0.2 bleibt das Problem bestehen.

2. Einige Primärschlüssel sind doppelt vergeben, so dass Metadaten und *n*-Gramme nicht mehr eindeutig einander zugeordnet werden können. [64]
3. Einzelne Texte sind mit ungültigen Zeitangaben wie „9999“ versehen.

Der Datensatz des ***Xu xiu si ku quan shu*** weist eine sehr ähnliche Struktur auf. Diese von 1931–1945 zusammengestellte Sammlung von insgesamt 5.420 Texten[65] gilt als Fortführung des *Si ku quan shu* 四庫全書 (*SKQS*).[66] Das *XXSKQS* ergänzt dabei Texte, die im Zeitraum nach der Zusammenstellung des *SKQS* bis 1927 entstanden sind. Aufgenommen wurden auch zahlreiche ältere Texte, die in der Qing-Zeit nicht berücksichtigt oder verboten waren oder von denen eine besser oder vollständiger erhaltene Ausgabe gefunden wurde, sowie einige Erzählungen (*xiaoshuo* 小說) und Lokalchroniken (*Difangzhi*).

Wie beim Qing-zeitlichen Vorbild sind die Texte im *XXSKQS* in vier Kategorien („Kammern“, *ku* 庫) eingeteilt: „Klassiker“ (*jing* 經), Geschichtswerke (*shi* 史), philosophische Texte („Meister“, *zi* 子) und Anthologien literarischer Texte (*ji* 集), so dass das Material z. B. für stilometrische Analysen von Interesse sein kann. Technische Mängel, wie sie der *DFZ*-Datensatz aufweist, bestehen kaum. Problematisch ist aber, dass die bereitgestellten Metadaten den Zeitpunkt der Veröffentlichung der im Korpus enthaltenen Manifestationen von Texten enthalten. Dieser kann viele Jahrhunderte nach der ursprünglichen Entstehung sein, so dass die Verwendung des Korpus im Kontext der Textdatierung als problematisch einzustufen ist.[67]

Während mit den *zhengshi*, den Belegstellen aus dem *DHYDCD* und vor allem den *n*-Gramm Datensätzen von CrossAsia zur Evaluation von Datierungsmethoden geeignetes schriftsprachliches Textmaterial vorhanden ist, muss vom Training eines Tokenizers oder sogar *PoS-Taggers* für schriftsprachliche Texte im Rahmen dieser Arbeit bedauerlicherweise Abstand genommen werden, da keine ausreichenden Daten verfügbar sind.[68]

Dank der genauen Metadaten und des großen Umfangs ist der *DFZ*-Datensatz das einzige unter den hier betrachteten Datensätzen bzw. Korpora, aus dem sowohl Trainings- als auch Testdaten für die Entwicklung und Evaluation von Datierungsmethoden entnommen werden können. Obwohl die frühesten Texte aus dem 11. Jh. stammen, sind dafür aber erst ab Ende des 15. Jh. ausreichend Texte vorhanden. Eine weitere Limitation ergibt aus der Spezialisierung auf ein einzelnes Textgenre. Ein bedeutend größerer Zeitraum kann mit den *DHYDCD*-Belegstellen abgedeckt werden, die zudem auch Material unterschiedlicher Textgattungen enthalten. Da die Segmente dieses Korpus aber aus einzelnen Sätzen zusammengesetzt sind, stehen keine passenden Volltexte als Testdaten zur Verfügung. Hierfür muss auf die übrigen *n*-Gramm bzw. *Plain Text*-Korpora ausgewichen werden.

64 Die Primärschlüssel wurden offensichtlich nicht als Sequenz, sondern zufällig vergeben, so dass es zur wiederholten Vergabe einzelner IDs gekommen ist. Ein Beispiel dafür ist `fbfd08097f0af325b0bf72b64df1a2bb`, die sowohl für das *Fuchuan xian zhi* 富川縣志 von 1890, als auch für das *Yuezhou fu zhi* 岳州府志 aus dem Jahr 1685 vergeben wurde.

65 Tatsächlich enthält das *XXSKQS* inzwischen über 33.000 Werke, die größtenteils offensichtlich nicht Teil des vorliegenden Datensatzes sind. Vgl. Wilkinson 2000, S. 275.

66 Das *Si ku quan shu* wurde als umfassende Textsammlung Ende des 18. Jhs. zwischen 1773 und 1782 auf Betreiben von Kaiser Qianlong 乾隆 (reg. 1735–1796) zusammengestellt und umfasst 3.461 Texte. Siehe ebd., S. 274.

67 Siehe Kapitel 6.1, S. 171, insb. auch ab S. 174.

68 Ausführlicher dazu siehe Kapitel 4.4, ab S. 73.

4.3 Vorverarbeitung und Normalisierung

In der quantitativen Korpuslinguistik werden – je nach Anwendungsfall – neben dem für die Textdatierung eher nachteilhaften Entfernen von *stop words*[69] – häufig Bearbeitungsschritte ausgeführt, die der Vereinheitlichung des untersuchten Materials dienen. So können für Worthäufigkeitsanalysen alle Groß- in Kleinbuchstaben konvertiert,[70] die Orthographie vereinheitlicht, oder ein *Stemming*[71] durchgeführt werden.[72] Für das Chinesische sind diese Arten des *pre-processing* hinfällig.

Bei der Textdatierung sind Normalisierungen nicht zwangsläufig hilfreich. Guo Siyuan et al. haben gezeigt, dass z. B. bestimmte *OCR*-Fehler (*Optical Character Recognition*, Texterkennung) (z. B. „f“ für „f“ und damit Falschschreibungen wie „fuch“ anstatt „fuch“) für eine bestimmte Zeit typisch sein können.[73] Auch Wissen über Rechtschreibreformen kann hilfreich sein, abweichende Schreibweisen zeitlich zu lokalisieren und damit für die Datierung nutzbar zu machen.[74] Änderungen an der graphischen Gestalt chinesischer Zeichen können für die Datierung ebenfalls relevant sein, in digitalen Ausgaben sind sie jedoch kaum nutzbar (s. u.). Das Chinesische bringt zudem weitere Besonderheiten mit sich, die bei der Vorverarbeitung zu beachten sind. Das gilt umso mehr, wenn mit eklektisch zusammengestellten Textsammlungen gearbeitet wird.

Codierung

Codierungen sind Konventionen, wie Repräsentationen von Zeichen digital gespeichert werden. Während für zeitgenössische englischsprachige Texte mit *ASCII* (*American Standard Code for Information Interchange*) bereits ab 1963 eine heute noch gängige Standardcodierung durchgesetzt werden konnte, ist für fast alle anderen Sprachen die Beschäftigung mit unterschiedlichen *encodings* relevant.[75] Für die chinesische Sprache sind *GB* (*guobiao* 國標, Abk. für *guojia biaozhun* 國家標準), *GBK* (*guobiao kuozhan* 國標擴展), eine Erweiterung des *GB*-Standards für traditionelle Langzeichen, sowie *Big5*, ein *encoding* für Langzeichen, das vor allem in Taiwan, Hong Kong und Macau eingesetzt wird, gängig.[76] Zunehmend setzt sich als internationaler Standard der *Unicode* durch,

69 Vgl. dazu Kapitel 2, ab S. 11.

70 Für Worthäufigkeitsanalysen ist es üblich, alle Großbuchstaben in Kleinbuchstaben umzuwandeln, so dass alle Instanzen zum gleichen *type* gezählt werden.

71 Beim *Stemming* werden morphologisch bedingte Wortendungen entfernt, so dass lediglich der Wortstamm zurückbleibt, „sprachen“, „sprechen“ und „sprich“ würde also z. B. zu „sprech-“.

72 Bei der Normalisierung werden allgemein alle in einem Korpus auftretenden Nicht-Standardvarianten eines Wortes auf einen einheitlichen Standard gebracht. Vgl. z. B. Richard W. Sproat et al. 2001: „Normalization of non-standard words“. In: *Computer Speech & Language* 15.3, S. 287–333. DOI: 10.1006/csla.2001.0169, S. 287–288.

73 Siehe Guo Siyuan et al. 2015, S. 4–6; Insgesamt scheinen *OCR*-Fehler sich aber eher negativ auf die Datierungsgenauigkeit auszuwirken. Siehe Graliński et al. 2017, S. 33.

74 Siehe Garcia-Fernandez et al. 2011, S. 7; siehe auch Graliński et al. 2017, S. 32.

75 Im *ASCII*-Standard werden alle Zeichen mit einer Länge von 7 Bit codiert, so dass max. 128 (2^7) unterschiedliche Zeichen codiert werden können, von denen etliche als Steuerzeichen für Fernschreiber benötigt wurden. Mit einer Länge von nur einem *Byte* pro Zeichen ist ASCII damit vielen neueren Codierungen überlegen, was den geringen Speicherplatzbedarf angeht. Siehe Fotis Jannidis 2017b: „Zahlen und Zeichen“. In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 59–67, S. 63–64.

76 Siehe Wong Kam-Fai 黃錦輝 et al. 2010: *Introduction to Chinese Natural Language Processing*. Hrsg. von Graeme Hirst. Synthesis Lectures on Human Language Technologies. San Rafael: Morgan & Claypool, S. 30–31; Siehe auch Lu Qin 2019: „Computers and Chinese writing systems“. In: *The Routledge Handbook of Chinese Applied Linguistics*. Hrsg. von Huang Chu-ren 黃居仁, Zhuo Jing-Schmidt und Barbara Meisterernst. Abingdon, Oxon & New York: Routledge, S. 461–482, S. 466–470.

der „Definitionen für alle größeren Sprachen der Welt“[77] enthält und in dem theoretisch über eine Million Zeichen definiert werden können. Sowohl Lang- als auch Kurzzeichen, sowie inzwischen etliche (wenngleich bei weitem nicht alle) Zeichenvarianten (*yitizi* 異體字) werden darin unterstützt.[78] Im Rahmen dieser Arbeit kommt UTF-8 (8 Bit Unicode Transformation Format) als Textdateiformat für *Unicode*-Daten zum Einsatz.[79] Durch den Umfang der *Unicode*-Definition ist eine Konvertierung von Texten aus *GB* oder *Big5* in Richtung *Unicode* – anders als in Gegenrichtung – in der Regel nicht nur problemlos, sondern auch verlustfrei möglich.

Zeichenvarianten

Wie auch die *Zeichen* der lateinischen Schrift hat sich die chinesische Schrift seit Vereinheitlichung der sogenannten Kanzleischrift (*lishu* 隸書) während der frühen Han-Zeit in ihrer graphischen Gestalt kaum verändert.[80] Anders als alphabetische Schriftsysteme bleibt sie auch in der Verwendung von Lautverschiebungen und dialektalen Variationen weitestgehend unberührt.[81] Eine Herausforderung für die quantitative Linguistik stellen jedoch graphische Zeichenvarianten dar. Solche *yiti zi* 異體字 existieren bereits in den frühesten Entwicklungsstufen der chinesischen Schrift.[82] Auch die ab 1956 schrittweise eingeführten Kurzzeichen (*jiantizi* 簡[简]體[体]字),[83] sowie lokale bzw. dialektale Abwandlungen[84] können als graphische Varianten aufgefasst werden. Zudem muss – gerade im Kontext der Textdatierung – auf Namenstabus (*bihui* 避諱) eingegangen werden.[85]

Neben der Verwendung eines einheitlichen *encodings*, sollte für den sinnvollen Vergleich von Worthäufigkeiten in einem heterogenen, diachronen Textkorpus eine Normalisierung der Texte auf Standardzeichen, sowie Kurz- oder besser Langzeichen (*fantizi* 繁體字) bzw. die Neutralisierung von im gewählten *encoding* unterstützten, bedeutungsgleichen Zeichenvarianten erwogen werden.[86]

77 Siehe Jannidis 2017b, S. 64.

78 In der aktuellen Version 9.0 sind ca. 128.000 Zeichen definiert. Die Zeichen werden mit einer vierstelligen Hexadezimalzahl adressiert, üblicherweise in der Notation `U+00DF` („ß“). Dass dem Unicode-Konsortium, welches diesen Standard definiert, etliche Organisationen und Einzelpersonen, aber auch große IT-Firmen wie IBM, Microsoft und Google angehören, spricht für seine langfristige Verfügbarkeit. Siehe ebd., S. 64–66.

79 Bei UTF-8 Dateien wird für jedes Zeichen nur die benötigte Anzahl an *Bytes* verwendet – *ASCII*-kompatible Zeichen benötigen also nur 1 Byte, wohingegen Zeichen mit höheren Positionen in der Codetabelle einen entsprechend höheren Speicherplatzbedarf haben. Siehe ebd., S. 64.

80 Siehe Qiu Xigui 裘錫圭 2000: *Chinese Writing. übersetzt von Gilbert L.* Mattos *und Jerry* Norman. Berkeley: The Society for the Study of Early China; The Institute of East Asian Studies, S. 113; siehe auch Norman 1988, S. 65; für einen historischen Abriss, der auch die Vereinheitlichung der Schrift durch Qin Shihuang 秦始皇 (reg. 221–210 v. u. Z.; davor 247–221 v. u. Z. reg. als König von Qin) abdeckt, siehe auch Roberto Nespeca-Moser 2005: „Auf dem Weg zu einem Lexikon, Tagger und Parser für das Antikchinesische“. Lizenziatsarbeit. Zürich: Universität Zürich, S. 27–30.

81 Vgl. z. B. Henry Rogers 2005: *Writing systems: a linguistic approach*. Blackwell textbooks in linguistics 18. Malden, MA & Oxford: Blackwell, S. 194.

82 Siehe Imre Galambos 2015: „Variant Characters“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill; Neben *yiti zi* 異體字 sind auch andere Arten von Schreibvarianten gängig. Ein konziser Überblick findet sich in Wilkinson 2000, S. 417–423.

83 Siehe dazu z. B. DeFrancis 1984, S. 257–261.

84 Siehe z. B. Cheung Kwan-hin 張韋顯 und Robert S. Bauer 2002: *The Representation of Cantonese with Chinese Characters*. Bd. 18. Journal of Chinese Linguistics Monograph Series. Hong Kong: Chinese University Press, für einen Einblick in die Verschriftlichung des Kantonesischen.

85 Als Nachschlagewerk dazu kann Wang Yankun 王彦坤, Hrsg. 1997: *Lidai bihuizi huidian* 歷代避諱字匯典 *(Geschichtliches Lexikon von Tabuzeichen)*. Zhongzhou guji chubanshe 中州古籍出版社, herangezogen werden.

86 Andernfalls werden nicht die Vorkommen von semantisch eigentlich identischen Zeichen, sondern die Vorkommen der Varianten gezählt, z. B. *wei* 為, 爲 und 为. Für Fragestellungen wie „Kommt in Korpus *C* häufiger Variante *a* oder Variante *b* vor?“ ist von einer Normalisierung natürlich abzusehen.

Für eine Normalisierung auf Langzeichen spricht, dass im Rahmen dieser Arbeit fast ausschließlich Texte betrachtet werden, die ursprünglich vor der offiziellen Einführung von Kurzzeichen verfasst wurden. Zudem werden für die Einträge des hier eingesetzten *DHYDCD* ebenfalls Langzeichen verwendet.[87] Mit *mafan* 麻煩 steht eine *Python*-Bibliothek zur Konvertierung zwischen Kurz- und Langzeichen zur Verfügung.[88] Bei dieser Art der Normalisierung muss ein Datenverlust in Kauf genommen werden, da in manchen Fällen zwei oder mehrere traditionelle Zeichen zu einem Kurzzeichen zusammengefasst wurden. Die beiden Zeichen *zhi* 誌 („Aufzeichnungen", in *difangzhi* 地方誌) und *zhi* 志 („Wille, Absicht") etwa wurden mit der Schriftreform zu 志 zusammengeführt, das zuvor auch bereits in beiden Bedeutungen verwendet wurde.

Komplexer ist die Berücksichtigung im *Unicode* vorgesehener *yiti zi*, da für Langzeichen weder durch das Unicode Consortium noch eine staatliche Stelle eindeutige offizielle Standardvarianten festgelegt werden.[89] Bei der Verwendung von Langzeichen innerhalb von *NLP*-Aufgaben muss daher – abhängig von den verwendeten Daten und der Zielsetzung – eine individuell passende Lösung für das Problem der Normalisierung gefunden werden.[90]

Im Rahmen dieser Arbeit werden daher die in den Worteinträgen des *DHYDCD* verwendeten Zeichen als Standard definiert. Um eine Ersetzungsliste für die folglichen Variantenzeichen zu erzeugen, wird die *Unihan*-Datenbank als Quelle genutzt.[91]

Terminologisch wird darin zwischen *y*-, und *z*-Varianten unterschieden. Als *y*-Varianten sind Zeichen definiert, die semantisch gleich, aber graphisch unterschiedlich sind und auch unterschiedlich dargestellt werden müssen („non-unifiable shapes") – „sinologischer" ausgedrückt meist Kurz- und Langzeichenvarianten von Schriftzeichen. *z*-Varianten werden lediglich graphisch unterschieden. `U+4E3A` *wei* 为 ist z. B. eine *y*-Variante von `U+70BA` *wei* 爲.[92] Eine zuverlässige Umwandlung zwischen Lang- und Kurzzeichen auf Basis dieser *y*-Varianten (`kSimplifiedVariantTable`) ist nicht nur wegen der fehlenden Bijektivität, sondern auch wegen mangelnder Einheitlichkeit unmöglich. Zusammen mit den Daten zu semantischen Varianten (`kSemanticVariantTable`) lassen sich die *z*-Varianten (`kZVariantTable`) aber zur (unvollständigen) Ermittlung von *yitizi* wie z. B. *wei* 爲 und 為, sowie *shuo* / *shui* / *yue* 說 und 説 nutzen.[93]

Mittels einer *SQL*-Abfrage auf die genannten Varianten-Tabellen der *Unihan*-Datenbank lassen sich 16.301 Zeilen mit Kandidaten für die Normalisierung ermitteln.

87 Zur gemischten Verwendung von Lang- und Kurzzeichen im *DHYDCD* siehe auch Kapitel 5.3, ab S. 113.

88 Schaaf 2017, `mafan` ermöglicht die (nicht immer verlustfreie) Konvertierung in beide Richtungen.

89 Für Kurzzeichen werden durch die *Diyipi yitizi zhengli biao* 第一批異體字整理表 orthodoxe Zeichen festgelegt. Siehe Zhonghua renmin gongheguo wenhuabu 中华人民共和国文化部 und Zhongguo wenzi gaige weiyuanhui 中国文字改革委员会, Hrsg. 1988 [1955]: *Diyipi yitizi zhenglibiao* 第一批异体字整理表 *(Erste Tabelle mit Standardformen für Zeichen mit Varianten)*. Beijing 北京: Zhonghua renmin gongheguo wenhuabu 中华人民共和国文化部 und Zhongguo wenzi gaige weiyuanhui 中国文字改革委员会; in der für Langzeichen vergleichbaren *Yitizi biao* des Bildungsministeriums der Republik China können mehrere Standardzeichen (*zhengtizi* 正體字) pro Variante definiert sein. Vgl. Jiaoyubu 教育部 (Bildungsministerium [der Republik China]) 2017: *Yitizi biao* 異體字表 *(Variantenzeichentabelle)*. Yitizi zidian 異體字字典 (Variantenzeichenwörterbuch). URL: `https://dict.variants.moe.edu.tw/variants/rbt/variant_modified_record_tiles.rbt` (besucht am 14. 07. 2021).

90 Ein Beispiel für eine sehr konservative Normalisierung mit nur 395 Ersetzungen ist das Vorgehen auf *Ctext.org*, siehe Sturgeon 2011, `https://ctext.org/faq/normalization`.

91 Die *Unihan*-Datenbank definiert die *Unicode*-Zeichenbereiche für CJK-Symbole, also chinesische, japanische und koreanische Zeichen. Sie ist über ein Sourceforge-Projekt von Chen Dingyi im *SQLite*-Format erhältlich. Siehe Chen Dingyi 2013: *libUnihan*. URL: `https://sourceforge.net/projects/libunihan/` (besucht am 30. 11. 2016).

92 Siehe Inc. Unicode 2016: *Glossary of Unicode Terms*. URL: `http://www.unicode.org/glossary/` (besucht am 30. 11. 2016).

93 Vgl. ebd.

```
select
 z.code code_left, u1.utf8 char_left, z.variantCode code_right, u2.utf8 char_right, 'z_variants
     ' source from kZVariantTable z
 left join utf8Table u1 on z.code = u1.code
 left join utf8Table u2 on z.variantCode = u2.code
union select
 z.variantCode code_left, u2.utf8 char_left, z.code code_right, u1.utf8 char_right, 'z_variants
     ' source from kZVariantTable z
 left join utf8Table u1 on z.code = u1.code
 left join utf8Table u2 on z.variantCode = u2.code
union select [...] from kZVariantTableExtra [...]
union select [...] from kSemanticVariantTable [...]
union select [...] from kSemanticVariantTableExtra [...]
union select [...] from kSimplifiedVariantTable [...]
order by code_left;
```

Um aus den Varianten-Einträgen einen Standard festzulegen, der sich am *DHYDCD* orientiert, müssen Listen mit den jeweils „zusammengehörigen“ Zeichen gebildet (z. B. *qi* 丌, *qi* 亓, *qi* 其) und dann zu jeder Liste ein Eintrag im *DHYDCD* als orthodoxe Variante bestimmt werden. Dass eine unkritische Verwendung so erzeugter Listen problematisch ist, sei am Beispiel von *li* 厲 erläutert: Hier ist *li* 厉 als vereinfachte Variante von *li* 厲 angegeben, jedoch ist auch *li* 历 als semantische Variante von 厲 aufgeführt. Da 历 wiederum die vereinfachte Variante von *li* 曆 bzw. *li* 歷 ist, käme es implizit zu einer semantisch falschen Gleichsetzung von 厲 („krass“) und 歷 („Erfahrung“).

Für die Normalisierung wird daher folgendes Vorgehen gewählt: Ist im *DHYDCD* nur eine der ermittelten Varianten aufgeführt, bzw. existieren nur zu einer der Varianten Worteinträge mit mehreren Zeichen, so wird dieses Zeichen als Standard betrachtet. Ist keine der Varianten aufgeführt, kann keine Standardisierung auf das *DHYDCD* vorgenommen werden.[94] Werden mehrere Variantenzeichen mit eigenen Untereinträgen aufgeführt, so werden beide als Standard akzeptiert und lediglich die übrigen Zeichen der Liste auf die häufigste Variante standardisiert. Für das Beispiel 其 (26 Worteinträge), 丌 (1 Worteintrag) und 亓 (nur Zeicheneintrag) ergeben sich also folgende Normalisierungen: alle Vorkommen von 其 und 丌 werden beibehalten, alle Vorkommen von 亓 werden durch 其 ersetzt. Alle Vorkommen von *wei* 為, 爲 und 为 werden zu 爲 vereinheitlicht, bei *z*-Varianten von *jian* 劍 („Schwert“) werden z. B. Vorkommen von 剑, 劒, 劔, 釼 und 釰 durch 劍 ersetzt.[95] 厉 wird auf 厲 normalisiert, 歷 und 历 auf 歷, sowie 曆 auf 曆. Eine Normalisierung von 曆 auf 歷 findet nicht statt, da beide Zeichen zahlreiche eigene Untereinträge aufweisen. Insgesamt über 1.200 Normalisierungen werden so als *Python*-Funktion `hydcd_standardize(str)` bereitgestellt, die vor der weiteren Verarbeitung von Texten oder *n*-Gramm-Listen alle Ersetzungsvorgänge vornehmen kann.

Tabuzeichen

Für die Datierung von Inschriften, Drucken und Handschriften spielt der Aspekt der Namenstabus (*bihui* 避諱) bzw. Tabuzeichen eine wichtige Rolle.[96] Diese Tradition ist beinah seit den frühesten schriftlichen Überlieferungen und bis zum Ende der Kaiserzeit 1912 belegt.[97] Dabei

94 Dasselbe gilt für Varianten, die nicht in der *Unihan*-Datenbank gelistet sind.
95 釰 ist zwar als Einzelzeichen im *DHYDCD* aufgeführt, hat jedoch keine eigenen Worteinträge.
96 Siehe Piotr Adamek 2015 [2012]: *Good Son is Sad If He Hears the Name of His Father: The Tabooing of Names in China as a Way of Implementing Social Values*. Monumenta Serica Monograph Series 66. St. Augustin & London [Leiden, Diss.]: Monumenta Serica & Routledge, S. 18, S. 241.
97 Siehe Adamek 2015 [2012], S. 3; Vergleichbare Namenstabus, die Auswirkungen bis zur Veränderung des Grundwortschatzes mit sich bringen können, sind auch aus anderen Kulturen bekannt, siehe z. B. Keller 2003, S. 17.

war es üblich, die persönlichen Namen regierender Herrscher oder der eigenen Ahnen nicht zu schreiben oder auszusprechen, was während bestimmter Zeiträume zur Ersetzung entsprechender Schriftzeichen durch andere phonologisch oder semantisch ähnliche, oder sogar eigens angepasste Zeichen führte, oder in der Auslassung von Zeichen resultierte.[98] Ein anschauliches Beispiel, bei dem das Tabu durch Weglassung des letzten Strichs befolgt wird, ist der Vorname des Kaisers Kangxi 康熙 (reg. 1654–1722), Xuanye 玄燁: Statt *xuan* 玄 wird 玄 geschrieben, auch in Zeichen wie *miao* 玅 („exquisit, subtil"), in denen 玄 als Komponente enthalten ist;[99] *ye* 燁 wird zu 煠.[100] Dank der umfassenden Dokumentation dieser Tradition ist die Analyse von Namenstabus ein wertvolles Werkzeug für die Datierung historischer Textausgaben.[101] In digitalisierten und dadurch oft standardisierten, modernen Ausgaben schriftsprachlicher Texte sind solche Varianten allerdings noch selten vorzufinden. Dass immer mehr Zeichenvarianten wie *xuan* 玄 und *ye* 煠 in den *Unicode* aufgenommen werden, stellt eine Verbesserung dieser Situation in Aussicht. Leider stehen noch keine digitalen Ressourcen zur Verfügung, die das vorhandene Wissen über Tabuzeichen für die Korpuslinguistik nutzbar machen. Denkbar wäre z. B. eine Datenbank mit Tabuzeichen und -gründen und den Zeiträumen der Anwendung des jeweiligen Tabus.[102] Ohne solche Daten muss – neben den ebenfalls in *Plain Text* nicht wiederzugebenden materiellen Eigenschaften einer Ausgabe – mit den Tabuzeichen ein weiterer datierungsrelevanter Aspekt der chinesischen Schrift außen vor gelassen werden.

Die hier verwendete Normalisierung schriftsprachlicher Texte umfasst damit neben der Verwendung von *Unicode* und Langzeichen die Reduzierung einiger semantisch identischer, aber graphisch unterschiedlicher Zeichenvarianten auf einen einheitlichen Standard, falls dieser durch das *DHYDCD* abgedeckt wird. Je nach Anwendungsfall wird die Normalisierung lediglich für den Abgleich mit den *DHYDCD*-Worteinträgen genutzt und auf eine Normalisierung der Einzelzeichen verzichtet.

4.4 Tokenisierung & *Part-of-Speech* Tagging

In den meisten gängigen Schriftsystemen verwenden wir Leerzeichen und unterschiedliche Arten von Interpunktion, um Wörter, Sätze und Satzteile voneinander abzugrenzen. Diese Art der Abgrenzung erleichtert das Erkennen sprachlicher Einheiten nicht nur für menschliche Leser, sie ermöglicht es auch, einfache Regeln für eine automatisierte Worterkennung bzw. -trennung festzulegen und damit eine Tokenisierung des Textes durchzuführen.[103]

Während Satzgrenzen in modernen Ausgaben chinesischsprachiger Texte durch ein eindeutiges Zeichen, den Satzendepunkt „。", klar markiert sind, ist das Segmentieren der Sätze in

98 Siehe Adamek 2015 [2012], v. a. S. 49–56.

99 Siehe Adamek 2015 [2012], S. 54–55; siehe auch Aisin-Gioro Xuanye 愛新覺羅 · 玄燁 (als Shengzu ren huangdi 聖祖仁皇帝), Hrsg. 1922 [1716]: *Yuding Kangxi zidian* 御定康熙字典 *(„Kaiserliches Kangxi-Zeichenwörterbuch")*. Shanghai 上海: Tongwen shuju 同文書局, S. 725.

100 Siehe Aisin-Gioro Xuanye 愛新覺羅 · 玄燁 (als Shengzu ren huangdi 聖祖仁皇帝) 1922 [1716], S. 682.

101 Siehe Adamek 2015 [2012], S. 18.

102 Eine gut aufbereitete chronologische Liste von Tabuzeichen für Herrschernamen findet sich in ebd., S. 337–356, allerdings ohne Angabe der zur Ersetzung verwendeten Zeichen.

103 Trotz vorhandener Leer- und Satzzeichen gibt es aber auch für westliche Sprachen Herausforderungen beim *Tokenizing* bzw. der Segmentierung. Im Deutschen zählen dazu etwa Abkürzungen wie „z. T." („zum Teil"), da der Punkt nach dem „z", gefolgt von dem Großbuchstaben „T" rein typographisch betrachtet ein Satzende suggeriert.

ihre einzelnen *Wörter* alles andere als trivial,[104] denn das Chinesische wird, wie etwa auch die klassische griechische Sprache, als eine Art *Scriptura continua* geschrieben. Die daraus resultierende Segmentierungsaufgabe ähnelt dem Versuch, einen seiner Leerzeichen beraubten Text ohne Kenntnis der Wortbedeutung zu lesen:

> bescheidnewahrheitsprechichdirwennsichdermenschdiekleinenarrenweltgewöhnlichfüreinga
> nzeshältichbineinteildesteilsderanfangsalleswareinteilderfinsternisdiesichdaslichtgebar[…][105]

Im Klassischen Chinesischen repräsentieren die einzelnen Schriftzeichen (*zi* 字) in der Regel Morpheme und stellen damit eine zuverlässige, visuell erkennbare sprachliche Einheit dar, die auch potenzielle Wortgrenzen markiert.[106] Zusätzlich können bei der Segmentierung etliche Partikel wie *yi* 矣, *ye* 也 und *yu* 與 hilfreich sein, die häufig das Satzende markieren und gleichzeitig noch Auskunft über die Art des Satzes geben *können* – aber nicht müssen.[107]

Schriftsprachliche, aber bereits auch klassische chinesische Texte können zahlreiche mehrsilbige *tokens* enthalten,[108] auch wenn gerade die antike Sprachform immer wieder als Beispiel für ausgeprägte Monosyllabizität angeführt wurde.[109] Zwar enthalten selbst Texte, die in der modernen Umgangssprache verfasst sind, einen relativ hohen Anteil einsilbiger Wörter,[110] doch ihre Bedeutung bzw. Verwendung als Einzelzeichen kann stark von Vorkommen in Komposita bzw. unterschiedlichen Kontexten abweichen. Die größte Herausforderung besteht also in der unterschiedlichen *Länge* der Wörter. Hinzu kommt, dass auch Muttersprachler bei einer manuellen Segmentierung von Texten zwar sinnvolle sprachliche Einheiten wählen, aber nicht unbedingt zum selben Ergebnis kommen.[111] Zur Veranschaulichung seien an dieser Stelle zwei moderne Beispiele gegeben:

- Für die Phrase 江澤民主席 schlagen Huang Liang et al. zwei „gültige“ Segmentierungen vor:
  1. *Jiang Zemin zhuxi* 江澤民 // 主席 („Der Vorsitzende Jiang Zemin“)
  2. *jiangze minzhu xi* 江澤 // 民主 // 席 („Flüsse und Sümpfe, Demokratie, Sitz“)

  „Apparently, the second segmentation is nonsense.“[112] Das muss aber nicht so sein, wie das nächste Beispiel zeigt:

104 Selbst die in heutzutage gedruckten Ausgaben übliche Interpunktion ist eher eine Erscheinung des 20. Jahrhunderts. Für eine ausführlichere Diskussion von Wort- und Satzgrenzen im Chinesischen siehe Christoph Harbsmeier 1998: *Language and Logic*. Hrsg. von Kenneth Robinson. Science and Civilization in China Volume 7, Part 1. Cambridge: Cambridge University Press, S. 174–184; ebenfalls zitiert in Nespeca-Moser 2005, S. 98.

105 Johann Wolfgang von Goethe 1871 [1808]: *Faust: Eine Tragödie*. Berlin: G. Grote'sche Verlagsbuchhandlung, S. 53. Leer- und Satzzeichen vom Verfasser entfernt und Großbuchstaben durch Kleinbuchstaben ersetzt. Ein ähnliches Beispiel findet sich auch in Harbsmeier 1998, S. 174.

106 Siehe z. B. Harbsmeier 1998, S. 175.

107 Siehe z. B. ebd., S. 174.

108 Siehe dazu auch die Graphik zur Lexemlänge auf S. 149.

109 Für eine aktuelle Diskussion zur Monosyllabizität bzw. eben *Nicht*-Monosyllabizität des klassischen Chinesischen siehe z. B. Wolfgang Behr 2018: „»Monosyllabism« and Some Other Perennial Clichés“. In: *Asia and Europe – Interconnected: Agents, Concepts, and Things*. Hrsg. von Angelika Malinar und Simone Müller. Wiesbaden: Harrassowitz, S. 155–209, v. a. S. 176-185; siehe auch George A. Kennedy 1951: „The Monosyllabic Myth“. In: *Journal of the American Oriental Society* 71.3, S. 161–166. doi: 10.2307/595185 , S. 161–166; sowie DeFrancis 1984, S. 104–118.

110 Siehe dazu auch Kapitel 5.7.3, ab S. 146.

111 Siehe dazu Richard W. Sproat et al. 1996: „A Stochastic Finite-State Word-Segmentation Algorithm for Chinese“. In: *Computational Linguistics* 22.3, S. 377–404, S. 393–394. In einem Experiment wird hier gezeigt, dass eine manuelle Segmentierung durch sechs Muttersprachler in einer 76-prozentigen Übereinstimmung untereinander resultiert, wodurch die Ergebnisse quantitativer Analysen verzerrt werden können.

112 Huang Liang et al. 2002b: „Statistical Part-of-Speech Tagging for Classical Chinese“. In: *Text, Speech and Dialogue: 5th International Conference, TSD 2002, Brno, Czech Republic September 9-12, 2002*. Hrsg. von Petr Sojka, Ivan Kopecek und Karel Pala. Berlin & Heidelberg: Springer, S. 115–122, S. 119.

- Die Phrase 發展中國家[113] lässt sich auf folgende Weisen zerlegen:
  1. *fazhan zhong guojia* 發展 // 中 // 國家 („Entwicklungsländer“)
  2. *fazhan zhongguo jia* 發展 // 中國 // 家 („chinesische Familien entwickeln“, oder gar „Chinaentwicklungsexpert:innen“)

Für die Segmentierung schriftsprachlicher Texte ergeben sich damit drei wesentliche Herausforderungen:

1. Wortgrenzen sind nicht an Leerräumen erkennbar und Wörter unterschiedlich lang.
2. Interpunktion steht, besonders bei klassischen Texten, nicht immer zur Verfügung.
3. Es gibt keine eindeutige und universell anerkannte Definition des Wortbegriffs.[114]

In den vergangenen Jahren ist die Entwicklung von Tokenizern und *PoS-Taggern* für das Chinesische stark vorangeschritten. Ein Großteil der vorhandenen Software ist zwar für die Segmentierung moderner Texte vorgesehen, seit kurzem existieren aber auch einige wenige auf schriftsprachliches bzw. klassisches Chinesisch spezialisierte Tokenizer bzw. Trainingsdatensätze. Selbst den für modernes Chinesisch verfügbaren Tokenizern attestieren Meng Yuxian et al. allerdings immer noch: „state-of-the-art word segmentation performance is far from perfect.“[115]

Eine frühe Arbeit zum *PoS-Tagging* für Klassisches Chinesisch auf Basis des Hidden-Markov-Modells (HMM) stammt von Huang Liang et al. (2002).[116] Für das Erlernen der statistischen Wahrscheinlichkeiten für lexikalische Kategorien der einzelnen Zeichen werden manuell getaggte Trainingsdaten aus *Dao de jing* 道德經 und *Lunyu* 論語 eingesetzt.[117] Sehr problematisch ist dabei die stark vereinfachende Annahme, dass „most words are written in the single-character […] form, thus no word segmentation is required.“[118] Grundlagen für Parsing und *PoS-Tagging* des klassischen Chinesisch wurden zudem auch von Nespeca-Moser untersucht.[119] Mit *UD-Kanbun* von Yasuoka Kōichi 安岡孝一 liegt inzwischen ein in *Python* geschriebener *Open Source Dependency Parser* für Klassisches Chinesisch vor, der neben Segmentierung und *PoS-Tagging* auch die Struktur klassischer Sätze visualisieren kann.[120]

113 Eva Lüdi Kong 2018: „“随文入观”: 古文的阅读、理解与翻译 (Hinein in den Text: Lesen, Verstehen und übersetzen klassischer Chinesischer Texte)“. Konferenzbeitrag vom 15. Dezember 2018 im Rahmen des *International Symposium on the Teaching of Classical Chinese* in Bonn.

114 Eine ausführliche Diskussion dazu findet sich in Jiang Shaoyu 蔣绍愚 2015: *Hanyu lishi cihui xue gaiyao* 汉语历史词汇学概要 *(Outline of the History of Chinese Lexicology)*. Beijing 北京: Shangwu yinshuguan 商务印书馆 (The Commercial Press), S. 41–53; siehe z. B. auch Schindelin 2005a, S. 960; Roger Lass weist allerdings darauf hin, dass der Wortbegriff auch allgemein nicht unproblematisch bzw. eindeutig ist. Siehe z. B. Lass 1997, S. 93.

115 Meng Yuxian et al. 2019, S. 3243.

116 Huang Liang et al. 2002b, HMMs sind stochastische Modelle, die auf Korpusdaten basieren und häufig im Rahmen von *PoS-Taggern* in Verbindung mit dem Viterbi-Algorithmus zum Einsatz kommen. Dabei „erlernt“ der Tagger aus dem Korpus Wahrscheinlichkeiten des Auftretens bestimmter Wortarten nach bzw. vor Wörtern oder Wortfolgen, hier 2–3-Gramm-Kombinationen. Vom selben Autorenteam stammt zudem eine Arbeit zum *Probabilistic Context-Free Grammar*-Modell, in der sie – auf Basis desselben getaggten „Korpus“ aus 1.000 Sätzen, mit einer Genauigkeit von 82,3 % Strukturbäume für Sätze, überwiegend aus *Xunzi* 荀子 und *Hanfeizi* 韓非子, generieren. Siehe Huang Liang et al. 2002a.

117 Für Testdaten aus denselben Texten wird ein *F*-Score des resultierenden Taggers von bis zu 97,6 % berichtet. Siehe Huang Liang et al. 2002b, S. 119–120. Insgesamt beschränken sich die Autoren auf 6.000 „Wörter“, wovon 5.500 als Trainingsdaten verwendet werden. Leider sind weder die resultierende Software noch der Quellcode zugänglich.

118 Ebd., S. 116. Die Autoren gehen sogar noch weiter: „Especially in Classical Chinese, a word is a single character, so no separation of word[s] is possible.“ (S. 117).

119 Siehe Nespeca-Moser 2005, *passim*. Da die Arbeit von Robert Gassmann betreut wurde, spricht Nespeca-Moser stets von „Antikchinesisch“.

120 Yasuoka Kōichi 安岡孝一 2019: „Universal Dependencies Treebank of the Four Books in Classical Chinese“. In: *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, S. 20–28, Siehe; Yasuoka Kōichi 安

Auf Basis der bei Google entwickelten Sprachmodelle *Bidirectional Encoder Representations from Transformers* (*BERT*),[121] und der *Python*-Bibliothek 🤗 HuggingFace *Transformers*[122] ist mit *CKIP Transformers*[123] unlängst ein *Segmenter* und *PoS-Tagger* veröffentlicht worden, der neben den enthaltenen auch mit anderen *BERT*-Modellen genutzt werden kann.[124]

Bei der Segmentierung chinesischer Texte lassen sich zwei wesentliche Herangehensweisen unterscheiden: statistisch, d. h. auf Basis von Korpora-Trainingsdaten, und lexikonbasierte. Erstere hätten sich zwar in aktuellen Studien als effektiver erwiesen – jedoch nur, wenn ausreichend passende Trainingsdaten in Form manuell segmentierter Korpora für die entsprechende Textgattung vorliegen.[125] Ein *maximum matching* mit Wörterbuchdaten als „simplest but remarkably robust model"[126] kann für Sprachentwicklungsstufen, für die keine geeigneten Trainingsdaten vorliegen, dennoch die beste verfügbare Möglichkeit des *Tokenizing* sein.[127] Dabei werden Texte Zeichen für Zeichen nach der längstmöglichen Übereinstimmung mit einem gegebenen Lexikon abgeglichen. Es kann von vorne nach hinten (*maximum forward match)*, von hinten nach vorne (*maximum backward match*) oder bidirektional segmentiert werden.[128]

Noch mehr als für die Segmentierung ist man im Bereich des *PoS-Tagging* auf Trainingsdaten angewiesen. Eine lexikonbasierte Zuordnung von *PoS-Tags* kommt wegen des „extraordinary freedom that almost any word enjoys to enter into […] atypical syntactic functions"[129] nicht in Betracht, denn „nouns can function like verbs; verbs and adjectives, likewise, may be used like nouns or adverbs […]".[130] Nur etwa 36 % aller Wörter im klassischen Chinesischen sind dabei nicht mehrdeutig.[131]

Die nachfolgende Evaluation zeigt jedoch, dass die verfügbaren Tools leider (noch) nicht für die gesamte schriftsprachliche Tradition geeignet sind. Im Rahmen dieser Arbeit wird auf *PoS-Tagging* daher verzichtet. Dass *PoS-Tagging* eigentlich gerade auch für die zeitliche Einordnung schriftsprachlicher Texte hilfreich sein dürfte, steht außer Frage. Bereits am Beispiel von *zhi* 之 lassen sich unterschiedliche linguistische Trends für die betrachteten Funktionen als Pronomen und als Subordinationspartikel erkennen.[132] Ebenfalls verzichtet wird hier auf die Erkennung unterschiedlicher Wortbedeutungen (*word sense disambiguation*), die – entsprechende Daten vor-

岡孝一 2019–: *UD-Kanbun*. GitHub Repository. URL: https://github.com/KoichiYasuoka/UD-Kanbun (besucht am 25. 05. 2021), siehe auch Kapitel 4.5, ab S. 88.

121 Siehe Jacob Devlin et al. 2019: „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *ArXiv* 1810.04805. DOI: 10.18653/v1/N19-1423.

122 Siehe Thomas Wolf et al. 2020: „Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, S. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

123 Mu Yang 慕楊 und Ma Wei-Yun 馬偉雲 2020–: *CKIP Transformers*. GitHub Repository. URL: https://github.com/ckiplab/ckip-transformers (besucht am 30. 05. 2021).

124 Auf der Website sind unzählige Sprachmodelle für mehr als 150 Sprachen verfügbar. Siehe Hugging Face 🤗 2020–: *Hugging Face Models*. Website. URL: https://huggingface.co/models (besucht am 15. 07. 2021); Ein verfügbares Modell für klassisches Chinesisch ist Ethan-yt 2020: *GuwenBERT Guwen yu xunlian moxing* 古文预训练模型. GitHub Repository. URL: https://github.com/Ethan-yt/guwenbert (besucht am 25. 05. 2021).

125 Siehe Huang Chu-ren 黃居仁 und Xue Nianwen 2019.

126 Meng Yuxian et al. 2019, S. 3244.

127 Siehe dazu auch Kapitel 4.6, ab S. 95. Vgl. auch Wong Kam-Fai 黃錦輝 et al. 2010, S. 43–57.

128 Siehe z. B. William J. Teahan et al. 2000: „A Compression-based Algorithm for Chinese Word Segmentation". In: *Computational Linguistic* 26.3, S. 375–393. DOI: 10.1162/089120100561746, S. 377–378.

129 Norman 1988, S. 88.

130 Ebd.

131 Siehe Huang Liang et al. 2002b, S. 121, vgl. auch S. 117.

132 Siehe Kapitel 2.3, v. a. Abb. 2.4, S. 26.

ausgesetzt – sicherlich ebenfalls zur Verbesserung der Genauigkeit von Datierungsmethoden genutzt werden könnte.[133]

4.5 Tokenisierung für schriftsprachliches Chinesisch: eine Feldstudie

Da nur wenige Tokenizer mit einer Spezialisierung auf schriftsprachliches Chinesisch zur Verfügung stehen, wird im Folgenden auch eine Auswahl der für die moderne Hochsprache frei verfügbaren Software (siehe Tabelle 4.2) vorgestellt. Die Genauigkeit dieser Segmenter wird für unterschiedliche Entwicklungsstufen der Schriftsprache anhand repräsentativer Texte, sowie eines modernen Referenz-Korpustexts untersucht. Da viele der modernen Komposita in schriftsprachlichen bzw. klassischen Texten als getrennte Wortformen auftreten können (z. B. *guojia* 國家, „Land" aus *guo* 國, „Staat" und *jia* 家, „Familie")[134], wird erwartet, dass für modernes Chinesisch entwickelte Tokenizer einige Wortgrenzen „übersehen".

Tabelle 4.2 Getestete Tokenizer

	Software	Version	Sprache	quelloffen	kostenlos	PoS-Tagging
1	*CKIP Zhongwen duan ci xitong* 中文斷詞系統	k. A. (2019)	k. A. (*C*, *C++*)	nein	beschränkt	ja
2	*CKIP Tagger*	0.2.1	*Python*	ja	ja	ja
3	*CKIP Transformers*	0.2.4	*Python*	ja	ja	ja
4	*GuwenBERT*	k. A. (2020)	*Python*	ja	ja	ja
5	*IK Analyzer*	5.0 (2012)	*Java*	ja	ja	ja
6	*Jieba* 結巴	0.39	*Python*	ja	ja	ja
7	*NLPIR-ICTCLAS*	k. A. (2019)	*Java*	ja	beschränkt	ja
8	*Paoding's Knives*	2.0.4	*Java*	ja	ja	nein
9	*Stanford Segmenter*	3.6	*Java*	ja	ja	ja
10	*Wenlin*	4.2	*C*	nein	nein	nein
11	*UD-Kanbun*	3.2.3	*Python*	ja	ja	ja

Ein vergleichbarer diachroner Test von Tokenizern für das Chinesische wurde bisher nicht durchgeführt.[135] Ähnliche Studien über *NLP*-Tools für Chinesisch, sogenannte *bake-offs* werden unregelmäßig von der *SIGHAN*-Gruppe veranstaltet, jedoch nicht für schriftsprachliches Chinesisch.[136]

133 Vgl. Kanhabua und Nørvåg 2008, S. 361; Eine Bestandsaufnahme der *state of the art* für die computerlinguistische Erkennung von semantischem Wandel findet sich in Nina Tahmasebi, Lars Borin und Adam Jatowt 2019: „Survey of Computational Approaches to Lexical Semantic Change Detection". In: *arXiv [cs. CL]* arXiv:1811.06278v2, S. 1–55.

134 Vgl. z. B. Ulrich Unger 1985b: *Einführung in das Klassische Chinesisch, Teil II: Erläuterungen*. Wiesbaden: Harrassowitz, S. 14.

135 Stand: Dezember 2021. Eine kurze Diskussion der Thematik findet sich aber in Mariana Zorkina 2021: „Defining word boundaries for Modern and Classical Chinese". In: *The Digital Orientalist*. URL: https://digitalorientalist.com/ (besucht am 16. 05. 2021).

136 Siehe auch Kapitel 4.1, S. 60. In „International Chinese Word Segmentation Bakeoffs" wird seit 2003 die Performance von Segmentern auf Testdaten von unterschiedlichen chinesischsprachigen Korpora systematisch verglichen. Siehe Richard W. Sproat und Thomas Emerson 2003: „The first international Chinese word segmentation Bakeoff". In: *Proceedings of the second SIGHAN workshop on Chinese language processing* 17, S. 133–143. DOI: 10.3115/1119250.1119269, S. 133–136.

Vorgehensweise

Um die Eignung unterschiedlicher Tokenizer für die Analyse schriftsprachlicher Texte zu prüfen, wird ihre Performance bei der Tokenisierung sogenannter Goldstandards, d. h. händisch vorsegmentierter Texte, untersucht. Da kein geeignetes einheitliches Korpus zur Verfügung steht, wird das Material aus unterschiedlichen Korpora zusammengestellt. Mit bekannten Texten soll so ein möglichst breites Spektrum an Sprachentwicklungsstufen abgedeckt werden. Die verwendeten Fragmente bestehen aus insgesamt über 32.000 *tokens* in acht Texten (Tabelle 4.3).[137] Jeder der ausgewählten Textabschnitte, vom frühen Schrifttum (*Shangshu* 尚書) bis in die Qing 清-Zeit (*Ru lin wai shi* 儒林外史) wird mit den in Tabelle 4.2 aufgeführten Tools segmentiert. Da diese überwiegend für moderne Texte entwickelt wurden, wird zum Vergleich auch die Segmentierung einiger Meldungen der Nachrichtenagentur *Xinhua* 新華 aus den späten 1990er Jahren getestet.[138] Anpassungen am Programmcode oder an der Ausgabe werden nur vorgenommen, wenn die Vergleichbarkeit der Ergebnisse dies erforderlich macht. Unberücksichtigt bleiben bei der hier durchgeführten Erhebung die Performance bei der Erkennung sogenannter *out of vocabulary*-Wörter,[139] sowie die Geschwindigkeit der Software.[140]

Auf Basis des jeweiligen Goldstandards werden *Precision*, *Recall* und *F-Score* gemessen.[141] Die *Precision* gibt an, welcher Anteil an gefundenen *tokens* tatsächlich relevant, d. h. im Goldstandard ebenfalls vorhanden ist. Der *Recall* gibt an, welcher Anteil an im Goldstandard vorhandenen *tokens* vom jeweiligen Tokenizer gefunden wurde. Der *F-Score*, ist ein künstliches Vergleichsmaß für *Information-Retrieval*-Systeme, in welchem *Precision* und *Recall* gleichberechtigt berücksichtigt werden. Als harmonisches Mittel dieser beiden Werte wird er so berechnet, dass sich jeweils ein Wert zwischen 0 (sehr schlecht) und 1 (sehr gut) ergibt:[142]

$$F\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Für eine Eignung von Tokenizern, die *tokens* eines beliebigen schriftsprachlichen Texts zu ermitteln, wird ein *Recall* von mindestens 0,9 angestrebt.

137 In einer vergleichbaren Studie wird ein Korpus mit etwa 15.000 *tokens* verwendet. Siehe He Ying und Mehmet Kayaalp 2006: *A Comparison of 13 Tokenizers on MEDLINE*. Technical Report. DOI: 10.1.1.216.2433, 4. Die Studie vergleicht insgesamt 13 Tokenizer auf ihre Eignung für die Tokenisierung von Abstracts englischsprachiger medizinischer Fachartikel.

138 Diese stammen aus der Martha Palmer et al. 2007: *Chinese Treebank 6.0*. URL: https://catalog.ldc.upenn.edu/LDC2007T36.

139 Bei der Evaluation von Tokenizern ist es üblich, *Precision*, *Recall* und *F-Score* zusätzlich für *in vocabulary* und *out of vocabulary*-Wörter, d. h. Wörter, die in den Trainingsdaten bzw. Wortlisten des jeweiligen Segmenters vorhanden (*in*), bzw. nicht vorhanden (*out of vocabulary*) sind, separat zu ermitteln. Siehe auch Sproat und Emerson 2003, Nicht alle hier getesteten Programme lassen gleichermaßen die Modifikation von Wortlisten durch Anwender:innen zu.

140 Mit Ausnahme der in der Berechnung aufwändigeren *BERT*-Modelle segmentieren alle getesteten Tokenizer die gegebenen Textabschnitte in (teils deutlich) unter einer Sekunde. Ein Geschwindigkeitsvergleich macht deutlich umfangreicheres Material erforderlich, einige der hier betrachteten Tokenizer werden dahingehend zudem oberflächlich untersucht in: Cao Liang, Wu Weiming und Gu Yonghao 2011: „The Research of Performance of Lucene's Chinese Tokenizer". In: *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*. IEEE, S. 7398–7401. DOI: 10.1109@AIMSEC.2011.6011478, Für die Frage nach der Eignung für schriftsprachliches Chinesisch ist die Geschwindigkeit zunächst irrelevant.

141 Zur Berechnung von *Precision* und *Recall* bzw. zum Abgleich mit dem Goldstandard werden hier zur Vereinfachung nicht die einzelnen tatsächlichen Auftreten von *tokens* in ihrer ursprünglichen Reihenfolge abgeglichen, sondern ihre Vorkommenshäufigkeit gezählt und verglichen.

142 Siehe Sasaki Yutaka 佐々木裕 2007: „The Truth of the F-measure". In: *Toyota Technological Institute (Toyota Kōgyō Daigaku* 豊田工業大学*)*. URL: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf (besucht am 04. II. 2022), S. 1–2.

Goldstandard-Textmaterial

In Ermangelung eines einheitlichen, breit aufgestellten diachronen Textkorpus werden die Goldstandards für die Tokenizer-Evaluation aus folgenden Quellen entnommen und in das Format „Wort | Wort | Wort | Satzzeichen | Wort |…“ vereinfacht.[143]

1. Auszüge aus der Online-Version des *Academia Sinica Ancient Chinese Corpus* 中央研究院古漢語標記語料庫 und des *Academia Sinica Tagged Corpus of Early Mandarin Chinese* 中央研究院近代漢語語料庫.[144] Die Normalisierung der *tokens* in das gewünschte Format erfolgt mittels regulärer Ausdrücke.[145]
2. *Sheffield Corpus of Chinese for Diachronic Linguistic Study*.[146] Die Sätze und *tokens* werden aus dem *XML*-Format des Korpus extrahiert. d. h. ein Satz in der ursprünglichen Formatierung:

```
<s>
  <noun type="polysyllabic" pinyin="zaishi">
    <preposition type="" pinyin="zai">在</preposition>世</noun>
  <verb type="verb_object_polysyllabic" pinyin="weiren">
    <preposition type="" pinyin="wei">爲</preposition>
    <pronoun type="" pinyin="ren">人</pronoun></verb>
  <verb type="monosyllabic" pinyin="bao">保</verb>
  <number type="definite" pinyin="qi">七</number>
  <classifier type="" pinyin="xun">旬</classifier>
  <punctuation type="" pinyin="">, </punctuation>
</s>
```

   wird hier verwendet als „ 在世 | 為人 | 保 | 七 | 旬 | , |…“.[147]
3. *Chinese Treebank 6.0* (*CTB*).[148] Die *tokens* werden ebenfalls aus dem *XML*-Format extrahiert.

Tabelle 4.3 Goldstandard-Texte für den Tokenizer-Vergleich

	Text (Abschnitt)	Quellkorpus	Einordnung	# *tokens*
1	*Shangshu* 尚書 (*Yao Dian* 堯典)	*Sinica*	ca. 7. Jhdt. v. u. Z.	1.487
2	*Lunyu* 論語 (1–4)	*Sinica*	ca. 5. Jhdt. v. u. Z.	2.860
3	*Mengzi* 孟子 (*Liang Hui wang* 梁惠王)	*Sinica*	ca. 3. Jhdt. v. u. Z.	6.087
4	*Shiji* 史記 (*Taishigong zixu* 太史公自序)	*Sinica*	94 v. u. Z.	7.716
5	*Zu tang ji* 祖堂集 (1–7)	*Sinica*	952	4.490
6	*Zhuzi yu lei* 朱子語類 (*Xue 6* 學六)	*Sheffield*	1270	3.702
7	*Ru lin wai shi* 儒林外史 (8)	*Sheffield*	1749	4.655
8	*Xinhua* 新華 (*Penn CTB 1-5*)	*Penn CTB*	1996	1.421

Einschränkungen

Die gewählte Herangehensweise führt zu Limitationen in der Vergleichbarkeit der einzelnen Tokenizer:

143 Ausführlichere Informationen zu den verwendeten Korpora finden sich in Kapitel 4.2, ab S. 62.
144 Huang Chu-ren 黃居仁 et. al. 1990; Huang Chu-ren 黃居仁 et. al. 2001, Da die Korpora der Academia Sinica nicht als Volltextdownload zur Verfügung stehen, können lediglich Auszüge verwendet werden.
145 „ 孟子 (NB1)[+prop] 見 (VK) 梁惠王 (NB1)[+prop]。王 (NA1) 曰 (VE):「叟 (NH) 不 (DC) 遠 (VP) 千 (S) 里 (NF) 而 (C) 來 (VA)…“ wird hier also verwendet als: „ 孟子 | 見 | 梁惠王 | 。| 王 | 曰 | : | 「| 叟 | 不 | 遠 | 千 | 里 | 而 | 來 | …“
146 Hu Xiaoling, Williamson und McLaughlin 2005.
147 Der XML-Baum wird mit *BeautifulSoup* verarbeitet. Leonard Richardson 1996–2020: *BeautifulSoup*. Python module. URL: https://www.crummy.com/software/BeautifulSoup/ (besucht am 02. 02. 2020).
148 Palmer et al. 2007.

1. Durch die Korpussituation bleibt der Zeitraum zwischen der östlichen Han- 漢 (25–220) und Tang 唐-Zeit (618–907) unterrepräsentiert.
2. Die uneinheitliche Ausgabe der verglichenen Tokenizer muss für die Evaluation standardisiert werden.
3. Die Korpora, denen die Goldstandards entnommen sind, können jeweils eigenen Richtlinien für die Segmentierung folgen.
4. Ein Vergleich der von einigen Tokenizern erzeugten *Part-of-Speech Tags* kann nicht erfolgen, da weder die *Tags* der Goldstandards noch die *PoS*-Definitionen der Tokenizer einheitlich sind. Für eine sinnvolle diachrone Evaluation von *PoS-Tagging* wären ein umfangreicheres Testframework und zumindest einheitliche Korpusdaten wünschenswert.[149]

Ergebnisse des Tokenizer-Vergleichs

Zur Veranschaulichung der Ergebnisse wird für die im Test besten Tokenizer die Ausgabe eines bekannten Abschnitts aus dem Werk *Mengzi* 孟子 (ca. 4 Jh. v. u. Z.), der Anfang des Kapitels *Liang Hui wang* 梁惠王, wiedergegeben.[150] Die angeführten Angaben zu *F-Score* (*F*), *Precision* (*P*) und *Recall* (*R*) beziehen sich jeweils auf das gesamte Kapitel. Ein Gesamtüberblick der Ergebnisse wird in Abschnitt 4.5.1 skizziert.[151]

> 孟子見梁惠王。王曰:「叟不遠千里而來，亦將有以利吾國乎?」孟子對曰:「王何必曰利? 亦有仁義而已矣。王曰『何以利吾國』? 大夫曰『何以利吾家』? 士庶人曰『何以利吾身』? 上下交征利而國危矣。萬乘之國弑其君者，必千乘之家；千乘之國弑其君者，必百乘之家。萬取千焉，千取百焉，不為不多矣。苟為後義而先利，不奪不饜。未有仁而遺其親者也，未有義而後其君者也。王亦曰仁義而已矣，何必曰利?」[152]
>
> (Mit Interpunktion 186 Zeichen)
>
> Mong Dsï [Mengzi] trat vor den König Hui von Liang. Der König sprach: „Alter Mann, tausend Meilen waren Euch nicht zu weit, um herzukommen, da habt ihr wohl auch einen Rat für mich, um meinem Reich zu nützen.“ Mong Dsï erwiderte und sprach: „Warum wollt Ihr durchaus vom Nutzen reden, o König? Es gibt doch auch einen Standpunkt, daß man einzig und allein nach Menschlichkeit und Recht fragt. Denn wenn der König spricht: ‚Was dient meinem Reiche zum Nutzen?‘ so sprechen die Adelsgeschlechter: ‚Was dient unserm Hause zum Nutzen‘ und die Ritter und die Leute des Volkes sprechen: ‚Was dient unserer Person zum Nutzen?‘ Hoch und niedrig sucht sich gegenseitig den Nutzen zu entwinden, und das Ergebnis ist, daß das Reich in Gefahr kommt. Wer in einem Reich von zehntausend Kriegswagen den Fürsten umzubringen wagt, der muss sicher selber über tausend Kriegswagen verfügen. Wer in einem Reich von tausend Kriegswagen den Fürsten umzubringen mag, der muss sicher selber über hundert Kriegswagen verfügen. Von zehntausend Kriegswagen tausend zu besitzen, von tausend Kriegswagen hundert zu besitzen, das ist an sich schon keine geringe Macht. Aber so man das Recht hintansetzt und den Nutzen voranstellt, ist man nicht befriedigt, es sei denn, daß man den anderen das Ihre wegnehmen kann. Auf der anderen Seite ist es noch nie vorgekommen, dass ein liebevoller Sohn seine Eltern im Stich läßt, oder daß ein pflichttreuer Diener seinen Fürsten vernachlässigt. Darum wollet auch Ihr, o König, Euch auf den Stand-

149 Siehe dazu auch Kapitel 4.2, S. 62, sowie 4.4, ab S. 73.
150 Die verwendete Version stammt aus dem klassischen Textkorpus der Academia Sinica Mengzi 孟子 1990: „Mengzi 孟子“. In: Academia Sinica 中央研究院. Kap. 1. URL: http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh (besucht am 10. 02. 2019), Im Rahmen der Tokenizer-Tests wurde das vollständige Kapitel verwendet.
151 Siehe ab S. 89.
152 Mengzi 孟子 1990.

> punkt stellen: ‚Einzig und allein Menschlichkeit und Recht!' Warum wollt Ihr durchaus vom Nutzen reden?"[153]

Das *Ancient Chinese Corpus* sieht folgende Tokenisierung vor:

> 孟子‖見‖梁惠王‖。‖王‖曰‖:‖「‖叟‖不‖遠‖千‖里‖而‖來‖,‖亦‖將‖有‖以‖利‖吾‖國‖乎‖?‖」‖孟子‖對‖曰‖:‖「‖王‖何‖必‖曰‖利‖?‖亦‖有‖仁義‖而已‖矣‖。‖王‖曰‖『‖何以‖利‖吾‖國‖』‖?‖大夫‖曰‖『‖何以‖利‖吾‖家‖』‖?‖士‖庶人‖曰‖『‖何以‖利‖吾‖身‖』‖?‖上下‖交‖征‖利‖而‖國‖危‖矣‖。‖萬乘‖之‖國‖弑‖其‖君‖者‖,‖必‖千乘‖之‖家‖;‖千乘‖之‖國‖弑‖其‖君‖者‖,‖必‖百乘‖之‖家‖。‖萬‖取‖千‖焉‖,‖千‖取‖百‖焉‖,‖不‖為‖不‖多‖矣‖。‖苟‖為‖後‖義‖而‖先‖利‖,‖不‖奪‖不‖饜‖。‖未‖有‖仁‖而‖遺‖其‖親‖者‖也‖,‖未‖有‖義‖而‖後‖其‖君‖者‖也‖。‖王‖亦‖曰‖仁義‖而已‖矣‖,‖何‖必‖曰‖利‖?‖」[154] (Mit Interpunktion 168 *tokens*)

CKIP / Academia Sinica

Die CKIP-Gruppe (*ciku xiaozu* 詞庫小組) entwickelt *Segmenter* und *PoS-Tagger* für das Chinesische an der an der Academia Sinica in Taipeh 台北. Das schon länger verfügbare *Zhongwen duan ci xitong* 中文斷詞系統 (*Chinese Word Segmentation System*) kann online verwendet und für private bzw. akademische Nutzung auch ein kostenloser Download beantragt werden.[155]

Die kürzlich veröffentlichten, neueren *CKIP Tagger*,[156] sowie *CKIP Transformers*[157] werden *Open Source* auf *GitHub* bereitgestellt. Letztere können auch mit anderen *BERT*-Sprachmodellen wie *GuwenBERT*[158] eingesetzt werden. Sowohl *CKIP Tagger* als auch *CKIP Transformers* werden als *Python*-Bibliotheken bereitgestellt, so dass sie nahtlos innerhalb eigener *NLP*-Workflows einsetzbar sind.

Die Ausgabe aller *CKIP*-Tools erfolgt mit Leerzeichen als Trennzeichen und *PoS-Tags* in Klammern:[159]

```
  孟子(Nb)  見(VE)  梁惠王(Nb)  。(PERIODCATEGORY)
--------------------------------------------------------------------------------
  王曰(Na)  ：(COLONCATEGORY)
--------------------------------------------------------------------------------
  「(PARENTHESISCATEGORY)  叟(FW)  不遠千里(D)  而(Cbb)  來(D)  ，(COMMACATEGORY)
--------------------------------------------------------------------------------
[…]
```

Normalisiert und auf die die Wortsegmentierung reduziert zunächst die Ausgabe des *Chinese Word Segmentation System*:

> 孟子|見|梁惠王|。|王曰|:|「|叟|不遠千里|而|來|,|亦|將|有|以利|吾|國|乎|?|」|孟子|對|曰|:|「|王|何必|曰|利|?|亦|有|仁義|而已|矣|。|王曰|『|何以|利|吾|國|』|?|

153 Richard Wilhelm 1982: *Mong Dsï: Die Lehrgespräche des Meisters Meng K'o*. Köln: Eugen Diederichs, S. 40f.

154 Mengzi 孟子 1990.

155 Der Download enthält eine *Python*-Klasse, mittels der die kompilierten Programmpakete auch über ein *API* (*Application programming interface*) angesprochen werden können. Damit eignet sich das *Chinese Word Segmentation System* auch zum Einsatz in eigenen Workflows. Die Algorithmen zur Segmentierung von *tokens*, Erkennung unbekannter *tokens* und Zuweisung der *Part of Speech*-Tags werden in zahlreichen Publikationen beschrieben. Siehe v. a. Ma Wei-Yun 馬偉雲 und Chen Keh-Jiann 陳克健 2003: „Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff". In: *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, S. 168–171; Tsai Yu-Fang und Chen Keh-Jiann 陳克健 2004: „Reliable and Cost-Effective Pos-Tagging". In: *International Journal of Computational Linguistics & Chinese Language Processing* 9.1, S. 83–96.

156 Li Peng-Hsuan 李朋軒 und Ma Wei-Yun 馬偉雲 2019–.

157 Mu Yang 慕楊 und Ma Wei-Yun 馬偉雲 2020–.

158 Ethan-yt 2020.

159 Aus Platzgründen ist nicht die gesamte Ausgabe des Beispiel-Abschnitts wiedergegeben.

> 大夫|曰|『|何以|利|吾|家|』|?|士|庶|人曰|『|何以|利|吾|身|』|?|上下|交征利|而|國危|矣|。|萬乘|之|國弒|其|君者|,|必|千|乘|之|家|;|千|乘|之|國弒|其|君者|,|必|百乘|之|家|。|千取百|焉|,|千取百|焉|,|不|為|不|多|矣|。|苟|為|後義|而|先|利|,|不|奪|不|饜|。|未有|仁|而|遺|其|親|者|也|,|未|有義|而後|其|君者|也|。|王|亦|曰|仁義|而已|矣|,|何必|曰|利|?|」(Mit Interpunktion 146 *tokens*, *F* 0,80, *P* 0,86, *R* 0,75)

Beide Namen, Mengzi und König Hui von Liang, werden ebenso korrekt erkannt wie ein Großteil der einsilbigen Nomen und Verben. Allerdings werden einige Phrasen als einzelnes *token* erkannt, darunter *bu yuan qian li* 不遠千里 („tausend *li* nicht für weit halten") und sogar *jiao zhengli* 交征利 (mod. eher „Dividenden auszahlen", in der Übersetzung von Wilhelm als separate *tokens* erkennbar: „… sucht sich gegenseitig den Nutzen zu entwinden […]") Die nominalisierende Partikel *zhe* 者 wird als Suffix betrachtet, so dass 君者 als ein *token* gewertet wird („der Fürst") – wohingegen *zhe* hier eigentlich die gesamte Phrase nominalisiert: *wan cheng zhi guo shi qi jun* ***zhe*** 萬乘之國弒其君者 („jemand, der in einem Land von zehntausend Kriegswagen seinen Fürsten umbringt"). Davon abgesehen ist die Segmentierung des klassischen Textmaterials durch das *Chinese Word Segmentation System* brauchbar, mit einem *F-Score* von 0,8 kann aber keine Empfehlung für die Verwendung als Segmenter für Klassisches Chinesisch ausgesprochen werden.

Eine leichte Verbesserung zeigt bereits der modernere *CKIP Tagger*:

> 孟子|見|梁惠王|。|王|曰|:|「|叟|不遠千里|而|來|,|亦|將|有|以利|吾|國|乎|?|」|孟子|對|曰|:|「|王|何必|曰|利|?|亦|有|仁義|而已|矣|。|王|曰|『|何以利|吾|國|』|?|大夫|曰|『|何以|利|吾|家|』|?|士庶人|曰|『|何以利|吾|身|』|?|上下|交征|利|而|國|危|矣|。|萬|乘|之|國|弒|其|君|者|,|必|千|乘|之|家|;|千|乘|之|國|弒|其|君者|,|必|百|乘|之|家|。|萬取|千|焉|,|千取百焉|,|不|為|不|多|矣|。|苟為|後義|而|先利|,|不|奪|不|饜|。|未|有|仁|而|遺|其|親|者|也|,|未|有|義|而後|其|君|者|也|。|王|亦|曰|仁義|而已|矣|,|何必|曰|利|?|」(Mit Interpunktion 154 *tokens*, *F* 0,85, *P* 0,88, *R* 0,82)

Einige der im älteren *Word Segmentation System* fehlenden Segmentierungen erfolgen nun korrekt, an anderen Stellen bleibt die Problematik fälschlich als mehrsilbig erkannter Ausdrücke aber bestehen.

Als für klassische Sprache vielversprechend kann die Segmentierung des auf *BERT* basierenden *CKIP Transformers* unter Verwendung des zugehörigen *CKIP BERT Base Chinese*-Sprachmodells[160] angesehen werden:

> 孟子|見|梁惠王|。|王|曰|:|「|叟|不遠千里|而|來|,|亦|將|有|以利|吾|國|乎|?|」|孟子|對|曰|:|「|王|何必|曰|利|?|亦|有|仁義|而已|矣|。|王|曰|『|何以|利|吾|國|』|?|大夫|曰|『|何以|利|吾|家|』|?|士庶人|曰|『|何以|利|吾|身|』|?|上下|交|征|利|而|國|危|矣|。|萬|乘|之|國|弒|其|君|者|,|必|千|乘|之|家|;|千|乘|之|國|弒|其|君|者|,|必|百|乘|之|家|。|萬|取|千|焉|,|千|取|百|焉|,|不|為|不|多|矣|。|苟|為|後義|而|先|利|,|不|奪|不|饜|。|未|有|仁|而|遺|其|親|者|也|,|未|有|義|而後|其|君|者|也|。|王|亦|曰|仁義|而已|矣|,|何必|曰|利|?|」|(Mit Interpunktion 163 *tokens*, *F* 0,89, *P* 0,91, *R* 0,87)

Nur noch an wenigen Stellen bleibt hier überhaupt der Einfluss des modernen Lexikons erkennbar: *er hou* 而後 („und danach") wird weiterhin als feststehender Ausdruck erkannt, ebenso das bereits genannte, im modernen Chinesischen als *chengyu* 成語 verwendete *bu yuan qian li* 不遠千里.[161]

160 Mu Yang 慕楊 2020: *CKIP BERT Base Chinese*. BERT Modell. URL: https://huggingface.co/ckiplab/bert-base-chinese (besucht am 13. 10. 2021).

161 *Chengyu* sind viergliedrige, zum Sprichwort gewordene Phrasen, die ihren Ursprung oft in klassischen Geschichten haben – wie der hier zitierten Stelle aus dem *Mengzi*. *Bu yuan qian li*, „Tausend *li* nicht für weit halten", beschreibt in der

Eine Stärke der *Transformers* bzw. *BERT*-Plattform besteht darin, dass auch andere, kompatible Sprachmodelle als Trainingsdaten verwendet werden können. *CKIP Transformers* können so also auch mit dem für klassische Sprache trainierten, am Beijing Institute of Technology (*Beijing ligong daxue* 北京理工大學) entwickelten *GuwenBERT*[162] verwendet werden:

> 孟|子|見|梁|惠|王|。|王|曰|:|「|叟|不|遠|千|里|而|來|,|亦|將|有|以|利|吾|國|乎|?|」|孟|子|對|曰|:|「|王|何|必|曰|利|?|亦|有|仁|義|而|已|矣|。|王|曰|『|何|以|利|吾|國|』|?|大|夫|曰|『|何|以|利|吾|家|』|?|士|庶|人|曰|『|何|以|利|吾|身|』|?|上|下|交|征|利|而|國|危|矣|。|萬|乘|之|國|弒|其|君|者|,|必|千|乘|之|家|;|千|乘|之|國|弒|其|君|者|,|必|百|乘|之|家|。|萬|取|千|焉|,|千|取|百|焉|,|不|為|不|多|矣|。|苟|為|後|義|而|先|利|,|不|奪|不|饜|。|未|有|仁|而|遺|其|親|者|也|,|未|有|義|而|後|其|君|者|也|。|王|亦|曰|仁|義|而|已|矣|,|何|必|曰|利|? (Mit Interpunktion 185 *tokens*, *F* 0,84, *P* 0,79, *R* 0,89)

Dabei wird der Text meist vollständig in Einzelzeichen segmentiert,[163] womit sich nur bei rein klassischen Texten relativ gute Ergebnisse erzielen lassen, die *Precision* bleibt jedoch auch für den klassischen Text *Mengzi* deutlich hinter der des nativen Sprachmodells des *CKIP Transformers*, *CKIP BERT Base Chinese* zurück.

Jieba 結巴

Jieba (chin. für „stottern", eigentlich *Jieba zhongwen fenci* 結巴中文分詞) ist eine in *Python* geschriebene *OpenSource*-Bibliothek, die Funktionen für Segmentierung und *PoS-Tagging* zur Verfügung stellt. Die Entwickler haben es sich zum Ziel gesetzt, das „beste Python-Modul zur Segmentierung chinesischer Wörter"[164] bereitzustellen.[165] Es können benutzerdefinierte Wörterbücher bzw. Wortlisten eingesetzt und zur Laufzeit verändert werden.[166] Diese enthalten zudem Häufigkeiten der *types* im Trainingskorpus, sowie mit *ICTCLAS*[167]-kompatible *PoS-Tags*. Wie der *CKIP Tagger* und *CKIP Transformers* kann auch *Jieba* flexibel in andere *Python*-Programme integriert werden. Es werden unterschiedliche Modi unterstützt:

— 1. Im ***accurate mode*** werden Ambiguitäten über Wahrscheinlichkeiten bzw. Worthäufigkeiten aufgelöst, was in der Regel zu einer höheren *Precision* führen sollte. Zudem versucht *Jieba* standardmäßig auf Basis des Hidden-Markov-Modells (HMM) und des Viterbi-Algorithmus mittels der Wahrscheinlichkeiten aus dem Trainingskorpus „Wörter" zu erkennen, die nicht in der verwendeten Wortliste enthalten sind. So können Wortbildungen mit im *Jieba*-Trainingskorpus vorkommenden Prä- und Suffixen erkannt werden.[168] Diese Option lässt sich deaktivieren, indem der Parameter `HMM` auf `False` gesetzt wird.

---

modernen Hochsprache gemeinhin die Bereitschaft, für etwas oder jemanden einen weiten Weg auf sich zu nehmen. „Tausend *li*" werden dabei bereits im *Zuozhuan* 左傳 als Abstraktion einer größeren Distanz verwendet. Siehe auch *DHYDCD*, 千里.

162 Ethan-yt 2020.

163 Bei mehreren Durchläufen sind die Ergebnisse nicht immer exakt reproduzierbar.

164 Sun Junyi 2018.

165 Vgl. auch Meng Yuxian et al. 2019, S. 3242. Die Autoren bezeichnen *Jieba* als „most widely-used open-sourced Chinese word segmentation system".

166 Siehe Sun Junyi 2018, In der Standarddistribution ist eine umfangreiche Wortliste von etwa 350.000 Wörtern enthalten, zudem kann alternativ eine umfangreiche Wortliste mit 600.000 Zeichenkombinationen in Kurz- und Langzeichen verwendet werden.

167 Siehe S. 85.

168 Womit das Modell trainiert wurde, wird leider nicht angegeben.

> 孟子|見|梁惠王|。|王|曰|:|「|叟|不遠|千里|而來|,|亦|將有|以利|吾|國乎|?|」|孟子|對|曰|:|「|王|何必|曰|利|?|亦|有仁義|而已|矣|。|王|曰|『|何以|利吾國|』|?|大夫|曰|『|何以|利吾家|』|?|士|庶人|曰|『|何以|利吾身|』|?|上下交征|利而國|危矣|。|萬乘|之國|弒|其君者|,|必千乘|之家|;|千乘|之國|弒|其君者|,|必|百乘|之|家|。|萬取|千焉|,|千取|百焉|,|不為|不|多|矣|。|苟為後義而|先利|,|不|奪|不|饜|。|未有|仁而|遺其|親者|也|,|未有|義而後其|君者|也|。|王|亦|曰|仁義|而已|矣|,|何必|曰|利|?|」(HMM aktiv, 120 *tokens*, *F* 0,64, *P* 0,73, *R* 0,56)

> 孟子|見|梁惠王|。|王|曰|:|「|叟|不|遠|千里|而|來|,|亦|將|有|以利|吾|國|乎|?|」|孟子|對|曰|:|「|王|何必|曰|利|?|亦|有|仁|義|而已|矣|。|王|曰|『|何以|利|吾|國|』|?|大夫|曰|『|何以|利|吾|家|』|?|士|庶人|曰|『|何以|利|吾|身|』|?|上下交征|利|而|國|危|矣|。|萬|乘|之|國|弒|其|君|者|,|必|千|乘|之|家|;|千|乘|之|國|弒|其|君|者|,|必|百乘|之|家|。|萬|取|千|焉|,|千|取|百|焉|,|不|為|不|多|矣|。|苟|為|後|義|而|先|利|,|不|奪|不|饜|。|未有|仁|而|遺|其|親|者|也|,|未有|義|而|後|其|君|者|也|。|王|亦|曰|仁|義|而已|矣|,|何必|曰|利|?|」(HMM deaktiviert, 165 *tokens*, *F* 0,87, *P* 0,87, *R* 0,88)

Bei Segmentierung des klassischen Abschnitts wirkt sich das offensichtlich mit modernem Sprachmaterial trainierte HMM negativ aus: *Jieba* „erkennt“ durch vermeintliche Suffixe wie *jia* 家 (mod. „Spezialist“) und *guo* 國 („Land“) *tokens* wie *liwujia* 利吾家 („Mir-nütz-Spezialist“) oder *liwuguo* 利吾國 („Mir-nütz-Land“). Die Gesamtperformance von *Jieba* im *accurate mode* bei abgeschaltetem HMM kann für klassisches Chinesisch allerdings als überraschend gut gewertet werden. Abbildung 4.1 vergleicht die *F*-Scores der *Jieba*-Modi bei der Segmentierung aller Texte aus Tabelle 4.3 (S. 79).

— 2. Der ***search mode*** dient primär der Erstellung von Suchmaschinenindizes – dabei werden keine Ambiguitäten aufgelöst, sondern alle durch die Wortlisten in Frage kommenden Zeichenkombinationen, sowie zusätzlich die einzelnen Zeichen ausgegeben. Auf Kosten der *Precision* sollte das zu einem höheren *Recall* führen.

> 孟子|見|梁惠王|。|王|曰|:|「|叟|不|遠|千里|而|來|,|亦|將|有|以利|吾|國|乎|?|」|孟子|對|曰|:|「|王|何必|曰|利|?|亦|有|仁|義|而已|矣|。|王|曰|『|何以|利|吾|國|』|?|大夫|曰|『|何以|利|吾|家|』|?|士|庶人|曰|『|何以|利|吾|身|』|?|上下|上下交征|利|而|國|危|矣|。|萬|乘|之|國|弒|其|君|者|,|必|千|乘|之|家|;|千|乘|之|國|弒|其|君|者|,|必|百乘|之|家|。|萬|取|千|焉|,|千|取|百|焉|,|不|為|不|多|矣|。|苟|為|後|義|而|先|利|,|不|奪|不|饜|。|未有|仁|而|遺|其|親|者|也|,|未有|義|而|後|其|君|者|也|。|王|亦|曰|仁|義|而已|矣|,|何必|曰|利|?|」(HMM deaktiviert, 166 *tokens*, *F* 0,87, *P* 0,86, *R* 0,88)

Der Unterschied zum *accurate mode* mit deaktiviertem HMM fällt mit der Kürze des gegebenen Beispiels kaum ins Gewicht: *shangxia* 上下 und *shangxiajiaozheng* 上下交征 werden beide als *tokens* akzeptiert. Bei Texten mit nur wenigen Segmentierungsambiguitäten ist der *F-Score* bei minimal schlechterer *Precision* und besserem *Recall* erwartungsgemäß nahezu identisch (Abb. 4.1).

— 3. Der ***full mode*** ist auf hohe Verarbeitungsgeschwindigkeit ausgelegt und arbeitet ohne HMM. Es werden alle sich aus den verwendeten Wortlisten ergebenden *tokens* ohne *PoS-Tagging* und Interpunktion ausgegeben. Wie in Abb. 4.2 und 4.3 zu sehen, bietet der *search mode* wie erwartet für alle getesteten Goldstandards beim *Recall*, der *accurate mode* wiederum für die *Precision* die besseren Ergebnisse bei der Segmentierung. Insgesamt schneiden für alle vormodernen Texte *search* und *accurate mode* am besten ab. Bei der Verarbeitung der neueren Texte (ab *Zhuzi yu lei* 珠子語類, ca. Anfang 13. Jh.) wirkt sich die Verwendung des HMM im *accurate mode* nicht

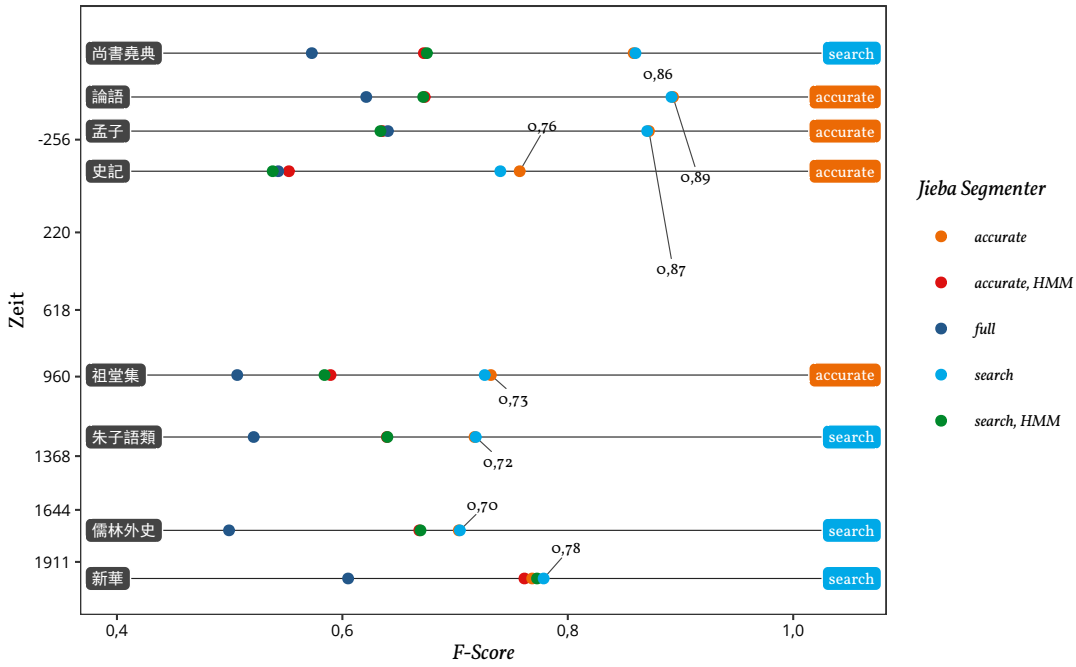


Abbildung 4.1 *F-Scores* der *Jieba-Segmenter*: links ist der segmentierte Goldstandardtext angegeben, rechts der Modus mit der jeweils besten Performance.

mehr oder kaum noch negativ aus. Im *search mode* verschlechtert das HMM auch den *Recall* für die modernen Textabschnitte, da falsche *tokens* anstatt der ursprünglichen Segmentierung ausgegeben werden.

NLPIR-ICTCLAS

NLPIR-ICTCLAS (*Natural Language Processing and Information Retrieval, Institute of Computing Technology Chinese Lexical Analysis System*) wurde zum ersten Mal 2003 vorgestellt[169] und wird von ZHANG Huaping 張華平 (Kevin ZHANG) entwickelt.[170] Die primäre Entwicklungssprache ist *Java*, es stehen aber auch APIs zu anderen Programmiersprachen zur Verfügung. Die Kernfunktionaltität bilden Tokenisierung und *PoS-Tagging*. Die hier durchgeführten Tests beziehen sich auf die 2018 auf der Website nutzbare Version, die Texte mit einer maximalen Länge von bis zu 3.000 Zeichen verarbeitet.[171] Ein freier Download der Software oder des Programmcodes wird nicht angeboten.

Der Beispielabschnitt aus dem *Mengzi* 孟子 wird wie folgt verarbeitet:

> 孟子/nr 见/v 梁惠王/nr。/wj 王/n 曰/vg：/wp「/w 叟/w 不远千里/vl 而/cc 来/vf，/wd 亦/d 将/d 有/vyou 以/p 利/n 吾/rr 国/n 乎/y?/ww」/w 孟子/nr 对/p 曰/vg：/wp「/w 王/n 何必/d 曰/vg 利/n?

169 ZHANG Huaping 張華平 et al. 2003: „HHMM-based Chinese Lexical Analyzer ICTCLAS“. In: *Proceedings of the Second Workshop on Chinese Language Processing, SIGHAN 2003, Sapporo, Japan, July 11-12, 2003*. URL: https://aclanthology.info/papers/W03-1730/w03-1730.

170 ZHANG Huaping 張華平 2018: *NLPIR-ICTCLAS 汉语分词系统 (NLPIR-ICTCLAS Chinese lexical analysis system)*. Website. URL: http://ictclas.nlpir.org/index.html (besucht am 18. 03. 2019), Die Software wird laut der offiziellen Website von etlichen großen Firmen eingesetzt und hat einen internationalen *bakeoff* gewonnen.

171 Ebd.

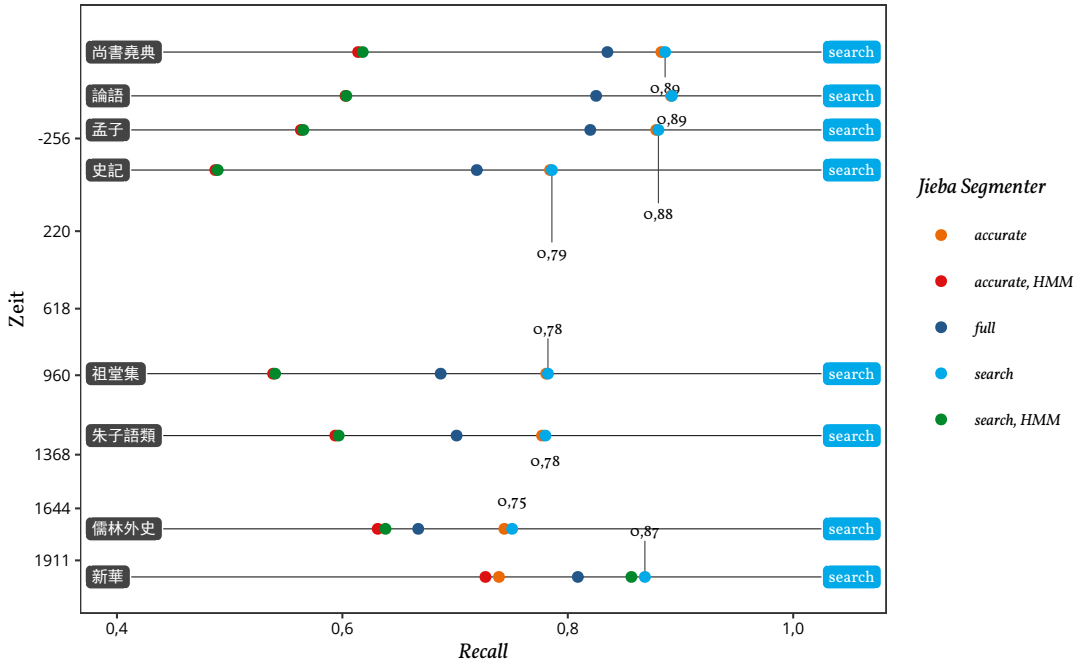


Abbildung 4.2 *Recall* der *Jieba* Modi, diachrone Goldstandards

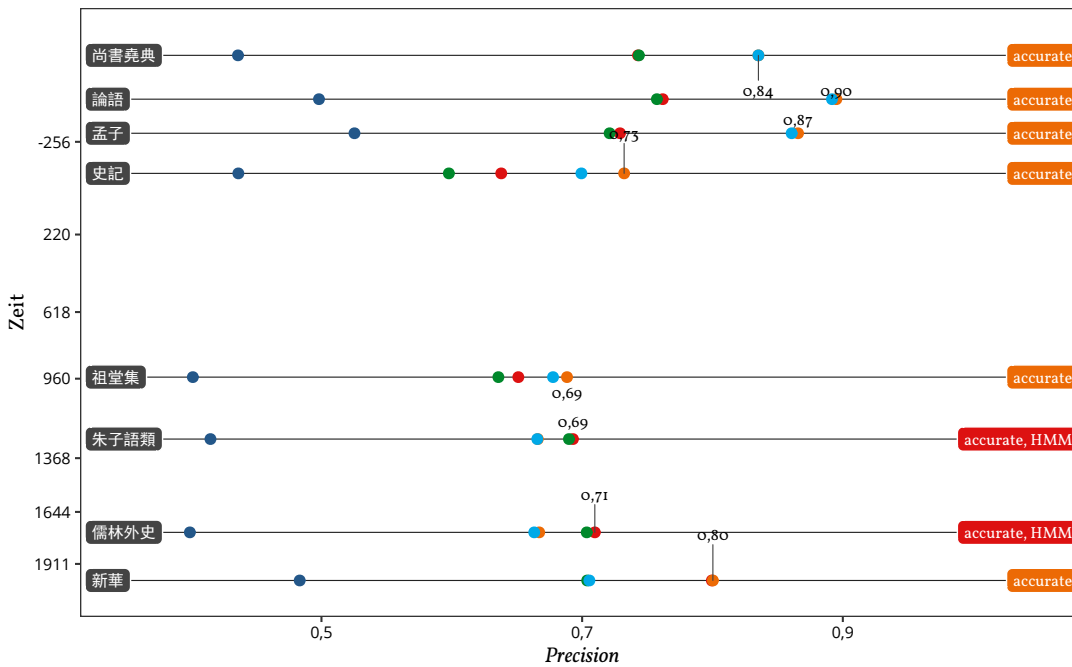


Abbildung 4.3 *Precision* der *Jieba* Modi, diachrone Goldstandards

/ww 亦/d 有/vyou 仁义/n 而已/y 矣/y。/wj 王/n 曰/vg『/wyz 何以/d 利/vg 吾/rr 国/n』/wyy? /ww 大夫/n 曰/vg『/wyz 何以/d 利/vg 吾/rr 家/n』/wyy? /ww 士/ng 庶人/n 曰/vg『/wyz 何以/d 利/vg 吾/rr 身/ng』/wyy? /ww 上下/n 交/ng 征/v 利/n 而/cc 国/n 危/ag 矣/y。/wj 万/m 乘/v 之/uzhi 国/n 弑/w 其君/nr2 者/k，/wd 必/d 千/m 乘/v 之/uzhi 家/n；/wf 千/m 乘/v 之/uzhi 国/n 弑/w 其君/nr2 者/k，/wd 必/d 百/m 乘/v 之/uzhi 家/n。/wj 万/m 取/v 千/m 焉/y，/wd 千/m 取/v 百/m 焉/y，/wd 不/d 为/v 不/d 多/a 矣/y。/wj 苟/ag 为/v 后/f 义/ng 而/cc 先/d 利/vg，/wd 不/d 夺/v 不/d 餍/w。/wj 未/d 有/vyou 仁/ag 而/cc 遗/vg 其/rz 亲/ng 者/k 也/d，/wd 未/d 有/vyou 义/ng 而/cc 后/f 其君/nr2 者/k 也/d。/wj 王亦曰/nr 仁义/n 而已/y 矣/y，/wd 何必/d 曰/vg 利/n？ /ww」/w (162 *tokens*, *F* 0,86, *P* 0,87, *R* 0,85, Alle Werte beziehen sich nur auf die Segmentierung.)

Auch das *CLAS* verarbeitet den klassischen Textabschnitt relativ gut. Allerdings wird das Ergebnis stets in Kurzzeichen ausgegeben. Dass *bu yuan qian li* 不远千里 („tausend *li* nicht für weit halten“) im Gegensatz zum Goldstandard als feststehender Ausdruck erkannt wird, ist wieder einem modernen Lexikon geschuldet. Noch problematischer ist aber die *NER*, die in dem kurzen Beispiel bereits an zwei Stellen „zugeschlagen“ hat: *qi jun* 其君 („seinen Fürsten“) wird als *Qijun* („Der Fürst von Qi 其“) tokenisiert, *wang yi yue* 王亦曰 (hier: „Saget auch Ihr, o König“…) wird zu Wang Yiyue. Für einen Tokenizer, der keinerlei Spezialisierung für die klassische Sprache hat, sind die Ergebnisse als gut zu bewerten.

Wenlin 文林

Die Segmentier-Funktion von *Wenlin* 文林[172] kann nicht ohne manuellen Aufwand in *NLP*-Workflows verwendet werden, da sie nur innerhalb einer *macOS* / *Windows* App zur Verfügung steht.[173] Zur genauen Implementierung werden keine Angaben gemacht, die Ergebnisse und die Tatsache, dass *Wenlin* in erster Linie eine Wörterbuchsoftware ist, lassen aber darauf schließen, dass ein wörterbuchbasiertes *maximum matching* eingesetzt wird.

Die Ausgabe beinhaltet die Anzeige von Ambiguitäten, so dass ein Eingriff durch die Anwender:in stattfinden kann – aber auch muss. Für die qualitative Bearbeitung von Texten ist dies sicherlich ein Mehrwert, bei quantitativen Analysen muss ein automatisierter Umgang mit den Auswahlmöglichkeiten gefunden werden. Die Performance von *Wenlin* wird hier jeweils mit allen möglichen erkannten *tokens* („all hits“), sowie mit automatischer Auswahl der ersten Auswahlmöglichkeit jeder Ambiguität („first hits“) berechnet. Dabei führt die erste Möglichkeit zu einem besseren *Recall*, die zweite zu einer potenziell höheren *Precision*.

Die Segmentierung des Veranschaulichungsbeispiels mit auswählbaren Ambiguitäten wird wie folgt ausgegeben:

孟子|見|梁|惠|王。王|曰:「叟|不遠千里|而|來，亦|將|有|以|利|吾|國|乎?」孟子|對|曰:「王|何必|曰|利? 亦|有|仁義| 【◎Fix:◎ 而已|矣;◎ 而|已矣】。王|曰『何以|利|吾|國』? 大夫|曰『何以|利|吾|家』? 【◎Fix:◎ 士庶|人;◎ 士|庶人】|曰『何以|利|吾身』? 上下交征利|而|國|危|矣。萬乘之國|弑|其|君|者，必|千乘|之|家；千乘之國|弑|其|君|者，必|百乘之家。萬|取|千|焉，千|取|百|焉，不|為|不|多|矣。苟|為|後|義|而|先|利，不|奪|不|饜。未有|仁|而|遺|其|親|者|也，未有|義|而後|其|君|者|也。王|亦|曰|仁義| 【◎Fix:◎ 而已|矣;◎ 而|已矣】，何必|曰|利?」(Mit Interpunktion 159 *tokens*, *F* 0,862, *P* 0,878, *R* 0,848)[174]

172 Wenlin Institute, Inc. 2015: *Wénlín* 文林 *Software for Learning Chinese, Version 4.2.0. macOS* App.

173 Die Funktionalität versteckt sich als „Segment Hanzi“ unter dem Menüpunkt „Edit“ > „Make transformed copy“

174 Da *Wenlin* Interpunktion im Gegensatz zum Goldstandard nicht als eigene *tokens* betrachtet, wurde diese in einem weiteren Bearbeitungsschritt mittels eines regulären Ausdrucks nachsegmentiert, um die Vergleichbarkeit der Ergebnisse zu gewährleisten.

Durch Klick auf die ◎ können Anwender:innen die gewünschte Option auswählen. Wird der erste Vorschlag angenommen, ergibt sich folgende Segmentierung:

> 孟子|見|梁|惠|王。王|曰:「叟|不遠千里|而|來，亦|將|有|以|利|吾|國|乎?」孟子|對|曰:「王|何必|曰|利? 亦|有|仁義|而已。王|曰『何以|利|吾|國』? 大夫|曰『何以|利|吾|家』? 士庶|人|曰『何以|利|吾身』? 上下交征利|而|國|危|矣。萬乘之國|弑|其|君|者，必|千乘|之|家；千乘之國|弑|其|君|者，必|百乘之家。萬|取|千|焉，千|取|百|焉，不|為|不|多|矣。苟|為|後|義|而|先|利，不|奪|不|饜。未有|仁|而|遺|其|親|者|也，未有|義|而後|其|君|者|也。王|亦|曰|仁義|而已|矣，何必|曰|利?」(Mit Interpunktion 153 *tokens*, *F* 0,86, *P* 0,884, *R* 0,837)

Dass der Segmentieralgorithmus von *Wenlin* bei mehreren der Goldstandardtexte gute und für das *Shangshu* sogar die besten Ergebnisse erzielt,[175] bestätigt, dass *maximum matching* für schriftsprachliche Texte mit geeigneten Lexikondaten momentan nicht schlechter geeignet ist als Methoden, die auf Trainingsdaten zurückgreifen. Die *Recall*-Marke von 0,9 wird jedoch erneut verfehlt.

UD-KanBun

Das zuerst 2019 veröffentlichte *UD-KanBun*[176] von Yasuoka Kōichi 安岡孝一[177] stellt neben der Segmentierung und *PoS-Tagging* auch eine Funktion zur Visualisierung der Satzstruktur bereit.

Im Folgenden ist das Ergebnis der Segmentierung des Beispielabschnitts durch *UD-Kanbun* wiedergegeben:

> 孟子|見|梁|惠|王|。|王|曰|:|「|叟|不|遠|千|里|而|來|，|亦|將|有|以|利|吾|國|乎|?|」|孟子|對|曰|:|「|王|何|必|曰|利|?|亦|有|仁|義|而|已|矣|。|王|曰|『|何|以|利|吾|國|』|?|大|夫|曰|『|何|以|利|吾|家|』|?|士|庶|人|曰|『|何|以|利|吾|身|』|?|上|下|交|征|利|而|國|危|矣|。|萬|乘|之|國|弑|其|君|者|，|必|千|乘|之|家|；|千|乘|之|國|弑|其|君|者|，|必|百|乘|之|家|。|萬|取|千|焉|，|千|取|百|焉|，|不|為|不|多|矣|。|苟|為|後|義|而|先|利|，|不|奪|不|饜|。|未|有|仁|而|遺|其|親|者|也|，|未|有|義|而|後|其|君|者|也|。|王|亦|曰|仁義|而|已|矣|，|何|必|曰|利|?|」(Mit Interpunktion 184 *tokens*, *F* 0,85, *P* 0,81, *R* 0,89

Anders als bei *GuwenBERT* werden auch mehrsilbige *tokens* wie Mengzi zugelassen, die Trennung von *Liang Hui wang* 梁惠王 in drei *tokens* zeigt jedoch direkt, dass die Trainingsdaten nur wenige Ausnahmen von der Monosyllabizität des *kanbun* 漢文 vorsehen. Die für klassische Texte insgesamt gute Performance bei der Segmentierung kann für mittelchinesisches oder noch späteres Textmaterial allerdings nicht erreicht werden.[178]

175 Vgl. auch Abb. 4.4, S. 90.

176 *Kanbun* 漢文 ist eine japanische Bezeichnung für klassisches Chinesisch, wobei wörtlich die Sprache der Han-Dynastie gemeint ist, der Begriff ist aber generell etwas weiter gefasst und schließt das frühe Schrifttum sowie etwa die Tang-Zeit mit ein. Siehe Astrid Brochlos 2004: *Kanbun* 漢文の基礎 – *Grundlagen der klassischen sino-japanischen Schriftsprache*. Wiesbaden: Harrassowitz, S. 9.

177 Yasuoka Kōichi 安岡孝一 2019; Yasuoka Kōichi 安岡孝一 2019–.

178 Siehe Abb. 4.4, S. 90.

Weitere Tokenizer

In der vorliegenden Untersuchung wurden auch einige in *Java* entwickelte *Open Source* Segmenter berücksichtigt. Der *IK Analyzer*,[179] sowie *Paoding's Knives* (chin. *Paoding jie niu* 庖丁解牛)[180] basieren auf den *Lucene*-Bibliotheken von Apache[181] und können mit eigenen Wörterbüchern erweitert werden. Trotz der klassischen Anspielung im Namen ist *Paoding's Knives* für die Zerlegung klassischen Textmaterials kaum geeignet. Auch der *IK Analyzer* bleibt in der Performance hinter den neueren Tokenizern zurück, so dass von einer detaillierten Betrachtung abgesehen werden kann.

Der *Stanford Segmenter* gehört zu einer Reihe von Programmbibliotheken, die von der Stanford NLP Group für unterschiedliche Sprachen entwickelt und veröffentlicht werden.[182] Als Trainingskorpus kommt die *Penn Chinese Treebank* zum Einsatz. So gut das „moderne Training" den *Stanford Segmenter* für die *Xinhua*-Texte aus der *CTB* macht,[183] so nachteilhaft wirkt es sich erneut auf die Segmentierung des älteren Textmaterials aus.

Den *Stanford Segmenter* mit umfangreichen Korpusdaten für die jeweilige Sprachentwicklungsstufe zu trainieren, könnte ein vielversprechender Ansatz sein – *out of the box* ist er aber für schriftsprachliche Texte ebenfalls nicht geeignet.

4.5.1 Gesamtvergleich der getesteten Segmenter

Abb. 4.4 zeigt zusammenfassend die *F*-Scores aus dem diachronen Tokenizer-Vergleich für alle Goldstandard-Texte.[184] Wie bereits diskutiert gehen mit der verfügbaren Software bei der Segmentierung von schriftsprachlichem Textmaterial stets mehr als 10 Prozent der vorhandenen *tokens* verloren. Der *Stanford-Segmenter* beeindruckt vor allem bei der Segmentierung des *Xinhua*-Texts aus seinen Trainingsdaten, ist für ältere Texte aber ohne entsprechendes Training ungeeignet. Ähnliches gilt umgekehrt für *UD-Kanbun*: der Textabschnitt aus dem *Lunyu* wird gut segmentiert, für spätere Texte, aber auch für das ältere *Shangshu*, sind die Ergebnisse weniger gut bzw. unbefriedigend.

179 Die Weiterentwicklung wurde vom ursprünglichen Autor 2012 eingestellt – Lin Liangyi 2012: *ik-analyzer IK-Analyzer java* 开源中文分词器. URL: `https://code.google.com/p/ik-analyzer/` (besucht am 13. 01. 2016); eine mit neueren Versionen von *Lucene* kompatible Version wurde aber von Eugene Su 2017: *IK Analyzer Solr 5*. URL: `https://github.com/EugenePig/ik-analyzer-solr5` (besucht am 07. 01. 2018), auf *GitHub* veröffentlicht.

180 Der Name *Paoding jie niu* 庖丁解牛 bezieht sich auf eine Stelle aus dem Buch *Zhuangzi* 莊子: ein Koch namens Paoding zerteilt Rinder so sehr im Einklang mit dem *dao* 道, dass sein Messer dabei nicht abstumpft. (*Zhuangzi* 3) Trey Lin 2013: *Paoding Analysis*. GitHub Repository. URL: `https://github.com/cslinmiso/paoding-analysis` (besucht am 26. 02. 2019).

181 Bei *Lucene* handelt es sich um eine in *Java* geschriebe *Open Source* Suchmaschinenbibliothek, die die Implementierung schneller Volltextsuchen vereinfachen soll. *Lucene* wird von Apache unter einer freien Lizenz veröffentlicht und kann auf deren Website kostenlos heruntergeladen werden. Vgl. Apache Software Foundation 2011–2016: *Lucene*. URL: `https://lucene.apache.org/core/` (besucht am 01. 05. 2018), Startseite.

182 Siehe dazu Stanford Natural Language Processing Group 2015; Der Quellcode ist über *GitHub* erhältlich: Stanford Natural Language Processing Group 2019: *Stanford CoreNLP*. GitHub Repository. URL: `https://github.com/stanfordnlp/CoreNLP` (besucht am 23. 03. 2019).

183 Bei der Verarbeitung des modernen Textbeispiels ist der *Stanford Segmenter* erwartungsgemäß „Testsieger" – schließlich ist die Segmentierung der eigenen Trainingsdaten gewissermaßen ein Heimspiel.

184 Siehe Tabelle 4.3, S. 79.

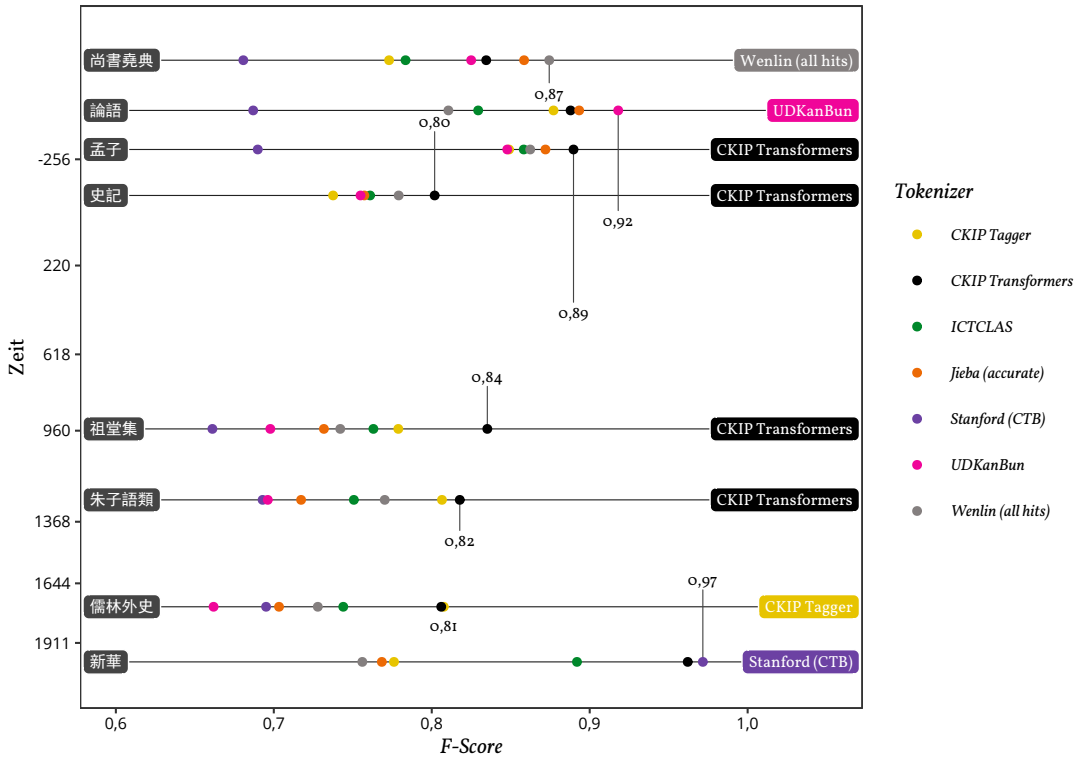


Abbildung 4.4 *F-Scores* aller getesten Tokenizer für alle Goldstandard-Texte. Links ist jeweils der Titel des Texts angegeben, rechts der Name des Tokenizers mit der jeweils besten Performance.

Einige Tokenizer wie der *CKIP Tagger* und *Jieba* liefern offensichtlich beim Segmentieren moderner Texte nicht zwangsläufig bessere Ergebnisse als bei klassischen oder schriftsprachlichen Texten. *NLPIR-ICTCLAS*, *IK-Analyzer* (ohne Abb.) und der *CKIP Tagger* der Academia Sinica bilden gemeinsam mit *Wenlin* und *Jieba* für die früheren Textbeispiele ein „Mittelfeld“. *CKIP Transformers* kann zusammen mit dem *CKIP BERT Base Chinese* Sprachmodell als Gesamttestsieger für die vormodernen Textabschnitte gewertet werden. Knapp dahinter folgen mit nur geringfügig schlechteren Ergebnissen *CKIP Tagger*, *Wenlin* und *Jieba*. Eine klare Empfehlung für die Segmentierung schriftsprachlichen Materials mit den verfügbaren Tools lässt sich nicht aussprechen.

Bei mittelchinesischen Texten und frühem Mandarin ist die Performance am schlechtesten. Für das moderne Referenzmaterial aus der *CTB* sind diejenigen Tokenizer am besten geeignet, die auf Basis von Trainingsdaten unbekannte Wörter erkennen und dadurch *F*-Scores von mehr als 0,9 erreichen können, die – geeignete Trainingskorpora vorausgesetzt – theoretisch sicherlich auch für andere Sprachentwicklungsstufen erreicht werden können. Ohne solche Trainingsdaten funktioniert eine einfache, lexikonbasierte Implementierung (*maximum matching*), wie sie bei *Wenlin* und *Jieba* mit abgeschaltetem HMM zum Einsatz kommt, für die früheren Sprachentwicklungsstufen insgesamt am besten. Segmentierfehler entstehen dann vor allem noch durch zu modernes Vokabular im verwendeten Lexikon. Eine Verbesserung der Ergebnisse sollte also durch den Einsatz einer zeitspezifischen Wortliste erzielt werden

können, die der Epoche des jeweils zu segmentierenden Texts angepasst ist.[185] Für undatiertes Textmaterial ist diese Lösung jedoch ungeeignet.

Auch ohne detaillierte Untersuchung der Performance der betrachteten Tokenizer beim Zuordnen von *PoS-Tags* muss davon ausgegangen werden, dass die Ergebnisse diejenigen der bloßen Segmentierung nicht übertreffen können. Für klassische und moderne Texte stehen also *Tagger* zur Verfügung, die akzeptable Ergebnisse liefern – nicht aber für Mittelchinesisch und frühes Mandarin.

Tabelle 4.4 Ranking der durchschnittlichen Performance aller getesteten Tokenizer mit allen vormodernen Goldstandard-Texten (ohne *Xinhua*), zur 1–4-Gramm Tokenisierung siehe die folgenden Abschnitte 4.5.2 und 4.5.3, zu *ChronLex* und *4ward* siehe Kapitel 4.6.

	Tokenizer	*F-Score*		Tokenizer	*Precision*		Tokenizer	*Recall*
1	*CKIP Transformers*	0,839	1	*CKIP Transformers*	0,859	1	*1–4 grams*	0,998
2	*ChronLex*	0,811	2	*CKIP Tagger*	0,838	2	*UD-KanBun*	0,838
3	*CKIP Tagger*	0,804	3	*4ward*	0,804	3	*1–4 gram words*	0,836
4	*Wenlin (all hits)*	0,795	4	*Wenlin (first hits)*	0,802	4	*ChronLex*	0,834
5	*Wenlin (first hits)*	0,795	5	*ICTCLAS*	0,797	5	*CKIP Transformers*	0,820
6	*4ward*	0,794	6	*Sinica*	0,795	6	*Jieba (accurate)*	0,820
7	*Jieba (accurate)*	0,790	7	*Wenlin (all hits)*	0,793	7	*Wenlin (all hits)*	0,799
8	*ICTCLAS*	0,784	8	*ChronLex*	0,790	8	*Wenlin (first hits)*	0,789
9	*UD-KanBun*	0,772	9	*Jieba (accurate)*	0,764	9	*4ward*	0,785
10	*1–4 gram words*	0,762	10	*Stanford (CTB)*	0,743	10	*CKIP Tagger*	0,774
11	*Sinica*	0,758	11	*UD-KanBun*	0,717	11	*ICTCLAS*	0,774
12	*IK Analyzer*	0,723	12	*IK Analyzer*	0,706	12	*IK Analyzer*	0,741
13	*GuwenBERT*	0,688	13	*1–4 gram words*	0,701	13	*Sinica*	0,724
14	*Stanford (CTB)*	0,672	14	*GuwenBERT*	0,671	14	*GuwenBERT*	0,709
15	*Paoding's Knives*	0,351	15	*Paoding's Knives*	0,419	15	*Stanford (CTB)*	0,614
16	*1–4 grams*	0,344	16	*1–4 grams*	0,208	16	*Paoding's Knives*	0,303

4.5.2 *n*-Gramm Zerlegung

Dass keiner der verfügbaren Tokenizer eine akzeptable Segmentierung schriftsprachlichen Textmaterials ermöglicht, legt nahe, auf alternative Strategien der Wortextraktion zurückzugreifen. Die Zerlegung in *n*-Gramme ermöglicht es, (fast) alle möglichen *tokens* aus einem Text zu extrahieren. Bei den Korpora, die bereits in dieser Abstraktionsstufe vorliegen,[186] ist eine Segmentierung bzw. *PoS-Tagging* sowieso nicht mehr möglich.

Wie die obigen Textbeispiele bereits andeuten, reicht die Verwendung von 1–4-Grammen aus, um knapp 100 % der in schriftsprachlichen Texten enthaltenen Wort-*types* zu identifizieren.[187] Obwohl in der modernen Hochsprache auch deutlich längere Wortbildungen möglich sind, wie das von Lü Shuxiang 呂淑湘 (1904–1998) bemühte Beispiel *tongbu wenxiang huixuan jiasuqi* 同步穩相回旋加速器 („Synchrocyclotron", eine Art Teilchenbeschleuniger) eindrucksvoll belegt,[188] ist eine Begrenzung auf vier Zeichen auch für moderne Texte noch sinnvoll, da der Anteil von

185 Dies wird in Kapitel 4.6 (ab S. 95) diskutiert.
186 *DFZ*; *XXSKQS*.
187 Vgl. auch Kapitel 5.7, S. 149.
188 Siehe Jiang Shaoyu 蒋绍愚 2015, S. 42–43.

Wörtern bzw. lexikalisierten Phrasen mit einer Länge von fünf oder mehr Zeichen zu jeder Zeit verschwindend gering ist.[189]

Tatsächlich sind im *HYDCD* auch ganze Phrasen lexikalisiert. Der Eintrag ***bushi dongfeng yaliao xifeng, jiushi xifeng yaliao dongfeng*** 【不是東風壓了西風，就是西風壓了東風】, ohne Interpunktion 16 Zeichen, stellt dabei das extremste Beispiel dar.[190] Auch wenn es sich dabei nicht um ein „Wort" handelt, spricht nichts dagegen, diese sprachliche Einheit, die das *HYDCD* mit dem Roman *Hong lou meng* 紅樓夢 belegt, als *type* für Textanalysen bzw. Sprachmodelle zu verwenden. Ein Blick auf die Längenverteilung der *DHYDCD*-Einträge beweist aber, dass die Berechnung von 1–16-Grammen unverhältnismäßig wäre.[191] Gleicht man alle 12 Millionen 1–16-Gramm-*types* im *Hong lou meng* mit den Einträgen des *DHYDCD* ab, finden sich 28.721 Lexem-*types*, von denen 28.652 (fast 99,8 %) eine Länge von 1–4 Zeichen haben.

Eine Berücksichtigung von 5+-Grammen wirkt sich auf die Erkennung zusätzlicher Lexem-*types* also nur marginal aus, wie Abb. 4.5 am Beispiel des *Hong lou meng* zeigt:

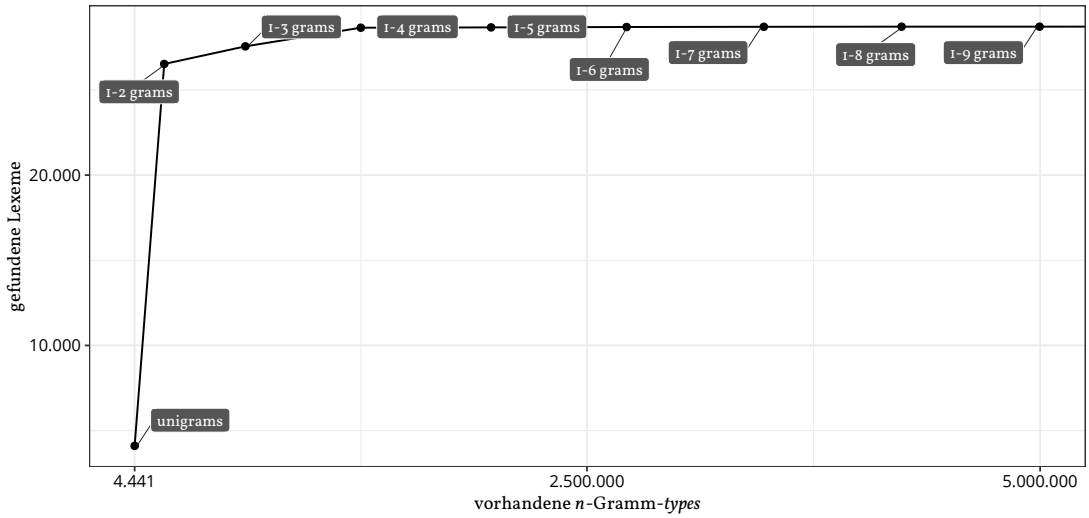


Abbildung 4.5 *n*-Gramm Effizienz am Beispiel von *Hong lou meng* 紅樓夢

Erzeugung von *n*-Gramm Häufigkeitslisten

Eine sehr performante Methode, *n*-Gramm-Listen mithilfe der *Python*-Funktion `zip` zu erzeugen, wird von Scott Triglia beschrieben.[192] Diese Implementierung wird hier im Wesentlichen

189 Eine statistische Untersuchung hierzu wird in Kapitel 5.7 (ab S. 138) erläutert; vgl. auch Abb. 5.13 (S. 149). Siehe auch die Untersuchungen zu Wortlängen im Chinesischen: Maria Breiter 1994: „Length of Chinese words in relation to their other systemic features". In: *Journal of quantitative linguistics* 1.3, S. 224–231; sowie Zhu Jinyang und Karl-Heinz Best 1998: „Wortlääufigkeiten in chinesischen Kurzgeschichten". In: *Asian and African Studies* 7, S. 45–51.

190 *HYDCD*, Bd. 1, S. 428.

191 Siehe Abb. 5.13, S. 149.

192 Siehe Scott Triglia 2013: *Elegant n-gram generation in Python*. Blog entry. URL: `http://locallyoptimal.com/blog/2013/01/20/elegant-n-gram-generation-in-python/` (besucht am 27.07.2016).

übernommen und an einem klassischen Beispiel kurz erläutert.[193] Zunächst wird der Text in eine Zeichenliste umgeformt:

```
>>> daodejing = "道可道，非常道。名可名，非常名。無，名天地之始；有，名萬物之母。"
>>> input_list = list(daodejing)
```

Damit steht eine Liste aller 1-Gramme von `daodejing` zur Verfügung.

```
>>> input_list
['道', '可', '道', '，', '非', '常', '道', '。', '名', '可', '名', '，', '非', '常', '名', '。'
    , '無', '，', '名', '天', '地', '之', '始', '；', '有', '，', '名', '萬', '物', '之', '母',
     '。']
```

Diese wird nun mit `zip` mit einer um den Index 1 verschobenen Liste (`input_list[1:]`) quasi im Reißverschlussverfahren zusammengeführt, um die Liste der 2-Gramme zu erzeugen.

```
>>> bigrams = list(zip(input_list, input_list[1:]))
>>> bigrams
[('道', '可'), ('可', '道'), ('道', '，'), ('，', '非'), ('非', '常'), ('常', '道'),  \mbox{[…
    ]} ('母', '。')]
```

Dieses Verfahren lässt sich für *n*-Gramme generalisieren, indem eine Liste der für `zip` zu verwendenden Listen aufgebaut und mit dem *-Operator wieder „entlistet“ wird:[194]

```
>>> def ngrams(input_list, n):
>>>     return zip(*[input_list[i:] for i in range(n)])
```

Die als `tuple` zurückgegebenen Elemente werden mit `join` wieder zusammengeführt, so dass die Elemente der Liste als *n*-Gramm-Strings zur Verfügung stehen:

```
>>> bigrams = (["".join(x) for x in list(find_ngrams(input_list, 2))])
>>> bigrams
['道可', '可道', '道，', '，非', '非常', '常道', […], '母。']
```

Durch Aufruf der `ngrams`-Funktion in einer `range`-Schleife von *m* bis *n*+1[195] kann nun eine Liste aller *m*–*n*-Gramme generiert werden:

```
>>> ngramlist, mingram, maxgram = [], 1, 4
>>> for n in range(mingram, maxgram+1):
>>>     ngramlist.extend(["".join(x) for x in list(ngrams(input_list, n))])
>>> ngramlist
 ['道', '可', '道', ... '，名萬物', '名萬物之', '萬物之母', '物之母。']
```

Aus der so erzeugten Liste von 122 1–4-Gramm-*tokens* von 103 *types*[196] kann nun mithilfe der Funktion `FreqDist().most_common(i)` aus der *Python*-Bibliothek `nltk`, die Häufigkeitsverteilung der *types* ermittelt werden, z. B.:[197]

```
>>> from nltk import FreqDist
>>> ngram_freqlist = FreqDist(ngrams).most_common(10)
>>> ngram_freqlist
[('名', 5), ('，', 4), ('道', 3), ('。', 3), ('可', 2), ('非', 2), ('常', 2), ('之', 2), ('，非
    ', 2), ('非常', 2)]
```

193 Laozi 老子 2009: *Lau-zi dao de jing* 老子《道德經》. eBook. URL: http://www.gutenberg.org/ebooks/7337 (besucht am 19. 05. 2019), Abschnitt 1. „道可道，非常道。名可名，非常名。無名天地之始；有名萬物之母。“ „Das *Dao*, das ausgesprochen werden kann, ist kein immerwährendes *Dao*. Namen, die genannt werden können, sind keine immerwährenden Namen. Die Nichtexistenz von Namen war am Anfang von Himmel und Erde; die Existenz von Namen ist die Mutter der zehntausend Dinge.“ (Interpretation des Verfassers.)

194 Siehe Triglia 2013.

195 Da die Aufrufparameter für `range` als *start* und *stop* definiert sind, muss z. B. der *start*-Wert 1 und der *stop*-Wert 5 sein, um eine Liste der 1–4-Gramme erzeugen.

196 Da die Beispiele lediglich der Veranschaulichung dienen, wird hier auf eine vollständige Wiedergabe verzichtet.

197 Siehe Steven Bird, Ewan Klein und Edward Loper 2009: *Natural Language Processing with Python*. 1. Aufl. Sebastopol: O'Reilly, S. 17; Steven Bird, Ewan Klein und Edward Loper 2014: *Natural Language Processing with Python*. 2. Aufl. URL: http://nltk.org/ (besucht am 12. 09. 2018), S. 19.

4.5.3 Zurück zur *Bag of Words*

Werden mit längeren Texten *alle* *n*–4-Gramme für die Berechnung von Sprachmodellen verwendet, erhält man nicht nur eine sehr große Anzahl, sondern auch einen hohen Anteil sinnentleerter Dimensionen. Dem kann durch eine bewusste *feature reduction* auf Basis der Häufigkeit entgegengewirkt werden, indem ein Anteil oder eine fixe Anzahl an häufigsten *n*-Grammen betrachtet oder eine Mindesthäufigkeit festgelegt wird. Für die Textdatierung können jedoch auch einzelne, seltene *types* im Zweifelsfall entscheidend sein, gerade dann, wenn der untersuchte Text wenig zeitgenössisches Vokabular aufweist.

Eine Alternative stellt daher die Reduktion auf genau diejenigen *features* dar, zu denen tatsächlich chronologische Daten vorliegen: die *types*, die im *DHYDCD* lexikalisiert sind.[198] Hierfür kann in *Python* die Schnittmenge der *n*-Gramm *types* und der *DHYDCD*-Lexeme (jeweils als `set`) mit dem Operator `&` ermittelt werden.[199]

```
words = (freq_grams & dict_entries)
```

Abb. 4.6 zeigt den *Recall* für die verwendeten Goldstandards bei Verwendung aller 1–4-Gramme („1–4 grams“) und mit der beschriebenen Reduktion der *features* („1–4 gram words“) im Vergleich mit den oben getesteten Tokenizern.

— 1. ***1–4 grams***. Ohne Beschränkung auf die *DHYDCD*-Lexeme werden für fast alle Textbeispiele 100 % der *tokens* gefunden. Erwartungsgemäß resultiert diese Vorgehensweise in einer extrem niedrigen *Precision* zwischen 0,14 für das moderne Textmaterial und etwa 0,2 für die schriftsprachlichen Texte.

— 2. ***1–4 gram words***. Die *Precision* kann für schriftsprachliche Texte auf 0,6 bis 0,86 erhöht werden, wenn – wie oben beschrieben – die Schnittmenge der gefundenen 1–4-Gramme mit der *DHYDCD*-Wortliste gebildet wird. Der *Recall* wiederum sinkt dadurch auf ein Niveau zwischen 0,75 und 0,92 für die schriftsprachlichen, sowie 0,42 für den modernen Vergleichstext und liegt damit für die älteren Texte immer noch über demjenigen der meisten Tokenizer.

198 Vgl. Kapitel 5.5, ab S. 120.

199 „In order to find an element in a set, a hash lookup is used (which is why sets are unordered). This makes contains (in operator) a lot more efficient for sets than lists.“ Python Software Foundation 2017: *Python 2.7.14 documentation*. URL: `https://docs.python.org/2/` (besucht am 26. 09. 2018), sets.html. Mehrere Millionen *n*-Gramm-*types* können so problemlos in wenigen 100-stel-Sekunden mit über 300.000 Wörterbucheinträgen verglichen werden.

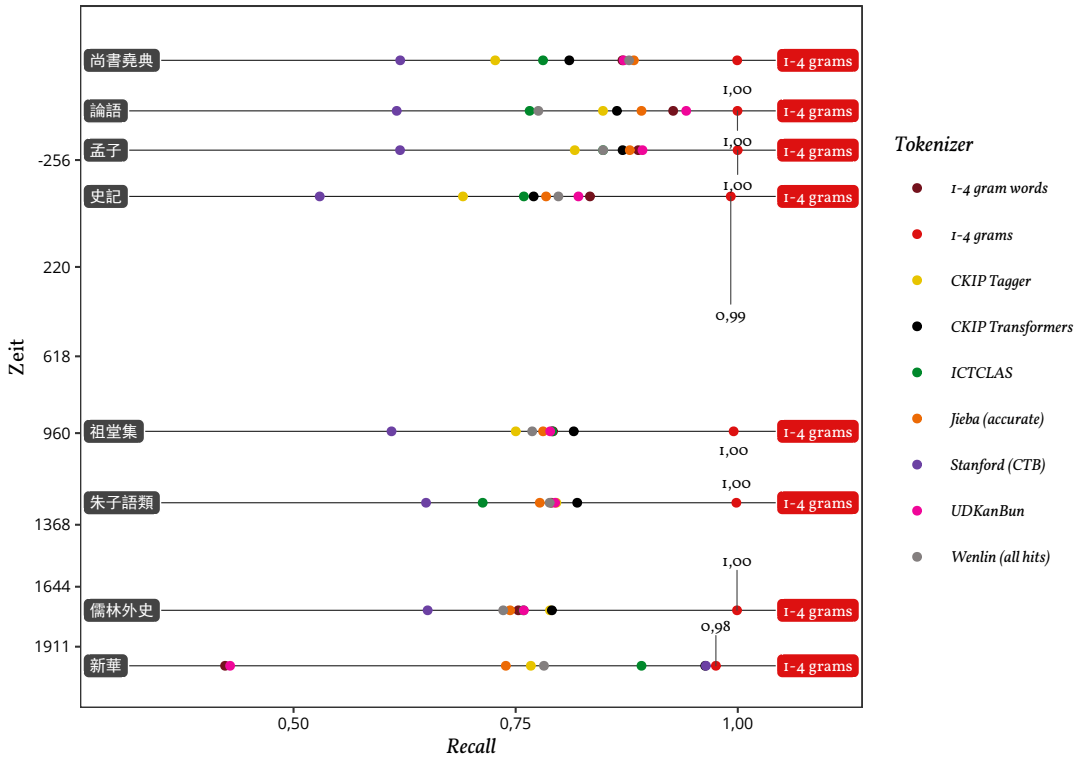


Abbildung 4.6 *Recall* der getesteten Tokenizer vs. Verwendung von 1–4-Grammen

Der *Recall* dieser zweiten Methode lässt auch vorsichtige Annahmen über die Vollständigkeit des *DHYDCD* in Bezug auf den Wortschatz unterschiedlicher Sprachentwicklungsstufen zu. Es deutet sich eine unterschiedlich gute Abdeckung an. Da hier nur einzelne Abschnitte ausgewählter Texte betrachtet werden, sollten aber keine voreiligen Schlüsse gezogen werden. Augenscheinlich ist eine geringere Erfassung des Vokabulars des ausgehenden 20. Jahrhunderts, was angesichts des Beginns der Kompilation in den 1970er Jahren wenig überrascht.[200]

Wenn – wie in Kapitel 6.2 und 6.3[201] – die Lexikalisierungsdaten aus dem *DHYDCD* als Datenquelle dienen, liefert die zweite Methode auch bei einem rechnerischen Recall von 0,75 bis 0,92 fast 100 % (bzw. 99,8 %) der tatsächlich nutzbaren *types*. Auch bei der Verwendung statistischer Sprachmodelle kann der Verlust der *out of vocabulary*-Dimensionen aber einer Verwendung aller *n*-Gramme vorgezogen werden.[202]

4.6 *ChronLex* – ein Segmenter-Experiment

Wie die Tokenizer-Evaluation in Kapitel 4.5 zeigt, liegt ein typisches Problem moderner Tokenizer mit klassischem bzw. schriftsprachlichem Textmaterial in der Erkennung erst später lexikalisierter Wörter bzw. Phrasen. So segmentiert z. B. das *CKIP Word Segmentation System* im

200 Siehe dazu auch Kapitel 5.1, ab S. 109.
201 Ab S. 179 bzw. ab S. 210
202 Siehe Kapitel 6.1 (ab S. 156).

Mengzi den Ausdruck *jiao zhengli* 交征利 („Dividenden auszahlen“) statt *jiao zheng li* („sich gegenseitig den Nutzen zu entwinden suchen“). Durch Verwendung eines bis zur Entstehungszeit des zu segmentierenden Textes eingeschränkten Vokabulars sollte also eine Verbesserung der Segmentierung erreicht werden. Hierzu muss die Anwender:in den Text zeitlich einordnen können und der *Segmenter* auf dieser Angabe basierende Wortlisten verwenden. Mit zunehmender Größe des verwendeten Lexikons und steigendem Anteil an mehrsilbigen *tokens* – also in jüngeren Texten – steigt dabei die Wahrscheinlichkeit für Ambiguitäten bzw. Segmentierfehler.

Um den potenziellen Nutzen dieser Maßnahme zu evaluieren, wird in *Python* ein einfaches *forward maximum matching* implementiert.[203] Die Benutzer:in wird aufgefordert, das Jahr der Veröffentlichung anzugeben und gemäß dieser Eingabe werden zeitgenössische und ältere Lexeme und Namen dynamisch zu einer passenden Wortliste zusammengestellt.[204] Segmentiert wird in Schritten von maximal vier Zeichen, d. h. immer die nächsten 4, 3 und dann 2 Zeichen werden auf einen Treffer in der diachronen Liste der verfügbaren *types* geprüft.[205] Zusätzlich werden Zahlwörter bis zu 6 Zeichen mittels eines regulären Ausdrucks erfasst. Wird keine Entsprechung gefunden, wird das Einzelzeichen als *token* angenommen und die Segmentierung beim nächsten Zeichen fortgesetzt. Diesen experimentellen Tokenizer bezeichne ich im Folgenden als *ChronLex* Tokenizer.

Die Segmentierung derselben Textabschnitte wie in Kapitel 4.5 ist vor allem für die klassische Periode vielversprechend (Abb. 4.7).[206] Als *Baseline* wird dieselbe Tokenisierung zusätzlich auch mit einer vollständigen, zeitunabhängigen Wort- und Namensliste ausgeführt. Diesen *Baseline*-Tokenizer bezeichne ich als *4ward* Tokenizer, da ebenfalls auf *tokens* mit einer Länge von 1–4 Zeichen in Leserichtung geprüft wird.

Ohne Trainingsdaten, Regeln oder statistische Modelle zu verwenden, wird so für die Textabschnitte aus *Lunyu* und *Mengzi* die *F-Score* Performance der jeweils besten in 4.5 getesteten Tokenizer beinahe erreicht, für den *Shiji*-Abschnitt sogar minimal übertroffen. Auch bei den anderen vormodernen Goldstandard-Textabschnitten reicht das Ergebnis nah an komplexere Tokenizer heran. Im Vergleich zur *Baseline* schneidet *ChronLex* besser ab, wobei dieser Trend sich bereits im *Ru lin wai shi* umkehrt und das Ergebnis der Segmentierung des modernen Textabschnitts schließlich deutlich schlechter ist, da eine große Menge an *out of vocabulary*-Wörtern und eine deutlich geringere Anzahl monosyllabischer *tokens* vorhanden sind. Die Nutzung diachroner Wort- und Namenslisten kann also für schriftsprachliche Texte helfen, die Segmentierung zu verbessern. Dieses Wissen lässt sich auch für die Verwendung von Tokenizern wie *Jieba* nutzen, bei denen die verwendeten Wortlisten angepasst werden können. Die Voraussetzung dafür ist, dass die Entstehungszeit des zu segmentierenden Textes bekannt ist.

203 Ebenfalls kann die Segmentierung von hinten nach vorne erfolgen (*maximum backward matching*, was im direkten Vergleich aber zu schlechteren Resultaten führt.

204 Zur dynamischen Erzeugung entsprechender Lexemlisten wird die in Kapitel 5.5 (ab S. 120) erstellte Datenbank konsultiert, Namenslisten werden durch eine entsprechende Abfrage auf die *China Biographical Database* (*CBDB*) (siehe Kapitel 4.7, S. 97) zusammengestellt.

205 Die Beschränkung auf Wörter mit max. 4 Zeichen dient der Laufzeitperformance. Eine Aufhebung dieser Beschränkung auf *n* Zeichen ist technisch problemlos möglich, bringt aber eine entsprechende Verlangsamung der Tokenisierung mit sich. Wie schon in Abb. 4.5 (S. 92) veranschaulicht, wäre der erzielte Effekt dabei absolut marginal. Siehe auch Abb. 5.13 (S. 149).

206 *Jieba* wurde hier im *accurate mode* mit abgeschaltetem *HMM* ausgeführt, da sich diese Einstellung für die Segmentierung des schriftsprachlichen Textmaterials am besten bewährt hat. Siehe den Abschnitt zu *Jieba* ab S. 83.

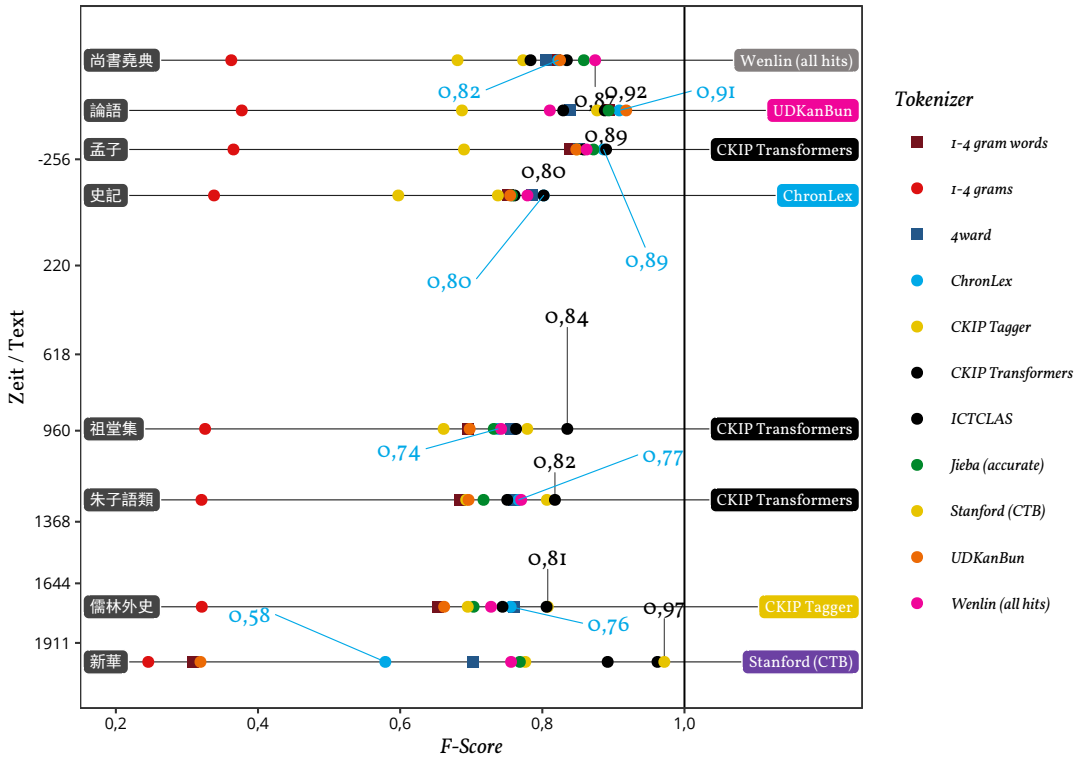


Abbildung 4.7 *F-Score* der getesteten Tokenizer vs. *ChronLex*

4.7 *Named Entity Recognition* (*NER*) und die *China Biographical Database* (*CBDB*)

Vorkommen von *Named Entities* können direkte Hinweise auf die zeitliche Einordnung von Texten liefern. Erwähnungen nicht fiktiver Personen deuten darauf hin, dass ein Text später zu datieren ist als auf das Geburtsjahr der erwähnten Person. Namen sind jedoch nicht ein-eindeutig, denn „in der chinesischen Geschichte gibt es enorm viele Personen, die denselben Namen tragen."[207] Eine Nutzung zu Datierungszwecken sollte also mit Bedacht geschehen. Hinzu kommt, dass chinesische Namen – deutlich häufiger als westliche – mit Zeichen(kombinationen) geschrieben werden, die auch in ihren lexikalisierten Bedeutungen in Texten vorkommen können. Ein unwahrscheinliches, aber anschauliches Beispiel ist Xi Jinping 習近平 (geb. 1953, reg. 2013–), dessen Name auch etwa mit „Übe, dem Frieden nah zu sein" übersetzt werden könnte.[208] Chinesische Namen sind daher potenziell in zweierlei Hinsicht ambig.[209]

Eine weitere Herausforderung für die *NER*, besonders in schriftsprachlichen Texten, ergibt sich aus der Tradition, dass Personen neben dem eigentlichen Vornamen (*benming* 本名) auch

207 Wilkinson 2000, S. 101, übersetzt durch den Verfasser.

208 Solche wörtlichen Bedeutungen von Namen können auch Gegenstand von Wortspielen sein. Vgl. z. B. Christian Soffel 2004: *Ein Universalgelehrter verarbeitet das Ende seiner Dynastie – Eine Analyse des Kunxue jiwen von Wang Yinglin*. Wiesbaden: Harrassowitz, S. 32.

209 Genauere Erläuterungen dazu siehe ab S. 100.

noch eine Vielzahl an alternativen Namen (*bieming* 別名) tragen können.[210] Dazu gehören unter anderem sogenannte Mannes- oder Großjährigkeitsnamen (*zi* 字), Literat:innennamen oder Pseudonyme (*[bie]hao* [別] 號) und postume Kanonnamen (*shi[hao]* 諡 [號]).[211] Als Extrembeispiel sei der Qing-zeitliche Gelehrte Liang Dingfen 梁鼎芬 (1859–1919) genannt, für den die *CBDB* 135 weitere Namen aufführt.[212]

Die Verwendung dieser teils ehrerbietigen Alternativnamen ist in schriftsprachlichen Texten als Referenz auf Personen durchaus üblich. Hinzu kommt, dass v. a. bei erneuter Nennung der Person häufig nur der Vorname (*ming*) genannt wird, oder allgemeinere Bezeichnungen, die Beruf, Amt oder sozialen Status widerspiegeln, sowie Höflichkeitsformen (*zuncheng* 尊稱).[213] Die Zuordnung dieser Referenzen zu biographischen Daten wird dadurch erschwert.

State-of-the-art NER für modernes Chinesisch basiert auf umfassenden Trainingsdaten, die für schriftsprachliche Texte so nicht vorliegen.[214] *NER*-Funktionalität wird zudem von einigen der in Kapitel 4.5 vorgestellten Tokenizer bereitgestellt, darunter *Jieba*, *CKIP Transformers*, *CKIP Tagger* und *ICTCLAS*.

Für schriftsprachliche Texte sei erneut die Plattform *MARKUS* erwähnt, die Orts- und Personennamen, Amtsbezeichnungen und temporale Ausdrücke in Texten, die über ein Webinterface hochgeladen werden, erkennt und hervorhebt (*Tagging*). Hierfür muss die Anwender:in die Epoche angeben, aus der der Text stammt. *MARKUS* verwendet regelbasierte Stringvergleiche und teilweise dynamisch erzeugte reguläre Ausdrücke.[215] Als Basis dafür dienen bestehende Datenbanken wie *CBDB* und die *DDBC Time Authority Database*.[216]

Da so gleichzeitig biographische Daten abgerufen werden können, die eine chronologische Einordnung erkannter *Named Entities* ermöglichen, wird im Rahmen dieser Arbeit ebenfalls eine datenbankgestützte Herangehensweise für die Erkennung von *Named Entities* gewählt. Diese eignet sich zugleich auch für den Abgleich mit *n*-Gramm Häufigkeiten. Im Folgenden wird auf die dafür verwendete *CBDB* eingegangen.

210 Eine Einführung in diese Thematik findet sich bei Wilkinson 2000, S. 98–103.

211 Die *China Biographical Database* (*CBDB*) unterscheidet – abgesehen von verschiedenen Familiennamen und Transliterationen – in der Tabelle `altname_codes` dreizehn Typen alternativer Namen, darunter z. B. *xiaoming* 小名 (Kindheitsname, auch *ruming* 乳名, „Milchname"), *shiming* 室名 („Studioname"), *faming* 法名 (Dharmaname) usw. Siehe *CBDB*.

212 Darunter z. B. die *zi* Xinhai 心海 und Bolie 伯烈, der *shiming* Buhui Shanmin 不回山民, der *shihao* 謚號 Wenzhong 文忠 uvwm. Siehe *CBDB*, Nr. 89173.

213 Ein kurzer Überblick findet sich z. B. in Wilkinson 2000, v. a. S. 104.

214 Siehe z. B. Zhang Yue und Yang Jie 2018: „Chinese NER Using Lattice LSTM". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne: Association for Computational Linguistics, S. 1554–1564. doi: 10.18653/v1/P18-1144, S. 1561–1562. Der entsprechende *Python*-Code ist bei *GitHub* verfügbar. Mittels eines Gittermodells (*Lattice LSTM*) können – je nach verwendeten Daten – *F*-Scores zwischen 58,59 und 94,46 bei der Erkennung chinesischer *Named Entities* erzielt werden. Abgesehen von der geringen Erfolgsaussicht scheitert die Anwendung auf Einträgen des *HYDCD* an im Modell unbekannten Zeichen. Eine manuelle Erstellung entsprechender Trainingsdaten wäre mit unverhältnismäßigem Aufwand verbunden. Eine noch rezentere Arbeit zu *NER* für Chinesisch mit *Python* ist Li Xiaonan et al. 2020: „FLAT: Chinese NER Using Flat-Lattice Transformer". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, S. 6836–6842, Die Erkennung von *bieming* wird in keiner dieser Veröffentlichungen thematisiert.

215 Ho und DeWeerdt. 2014–, Über die Erkennungsgenauigkeit werden keine Angaben gemacht. Um die Anzeige von *false positives* durch homonyme Personen zu reduzieren, muss in *MARKUS* zunächst die Epoche des Texts ausgewählt werden, für den das *Markup* generiert werden soll. Für undatiertes Material ist das selbstverständlich nicht denkbar.

216 Siehe *CBDB*; Marcus Bingenheimer et al. 2016: „Modelling East Asian Calendars in an Open Source Authority Database". In: *International Journal of Humanities and Arts Computing* 10.2, S. 127–144. doi: 10.3366/ijhac.2016.0164.

China Biographical Database Project (CBDB)

Mit der *China Biographical Database* steht eine frei nutzbare Datenbank mit biographischen Daten zu 366.588 Personen der chinesischen Geschichte zur Verfügung.[217] Zusätzlich zu den biographischen Daten sind auch bibliographische Daten zu Texten enthalten, die mit diesen Personen in Verbindung gebracht werden. Da die Datenbank in einem *SQL*-Dialekt frei heruntergeladen werden kann, kann sie in eigene Applikationen eingebunden werden.[218] Von den insgesamt 85 Datentabellen der verwendeten Version[219] der *CBDB* sind vor allem die folgenden für Datierungszwecke relevant:

— 1. `biog_main` enthält die eigentlichen biographischen Daten. Da diese nicht für alle aufgenommenen Personen vollständig und genau (bekannt) sind, sind viele Einträge nicht einheitlich bzw. unvollständig. Die Datenbankarchitekt:innen behelfen sich daher mit drei unterschiedlichen Arten von Jahresangaben:

— 1.1 `c_birthyear` und `c_deathyear` als Geburts- und Todesjahr. Nur bei knapp zehn Prozent der Datensätze ist diese genaue Angabe vollständig.[220]

— 1.2 Für deutlich mehr (253.969) Datensätze ist das sog. Indexjahr (`c_index_year`) gepflegt. Die Herausgeber:innen bezeichnen es als das Jahr, in dem eine Person vermeintlich etwa in ihrem sechzigsten Lebensjahr war. Für Personen, die früher gestorben sind, wird dann das (vermutete) Todesjahr angegeben.[221]

— 1.3 Die Dokumentation zu `c_fl_earliest_year` und `c_fl_latest_year`, der jeweils frühesten bzw. spätesten geschichtlich überlieferten Nennung einer Person ist leider weniger transparent.

Um unter Berücksichtigung dieser Angaben möglichst umfassende biographische Daten zu erhalten, können die Spalten priorisierend zusammengefasst werden.[222]

```
SELECT `c_personid`, `c_name_chn`,
  if(coalesce(`c_birthyear`,0) = 0, if(coalesce(`c_fl_earliest_year`,0) = 0, `c_index_year`, `
     c_fl_earliest_year`), `c_birthyear`) as startyear,
  if(coalesce(`c_deathyear`, 0) = 0, if(coalesce(`c_fl_latest_year`,0) = 0, `c_index_year`, `
     c_fl_latest_year`), `c_deathyear`) as endyear
FROM `biog_main`
HAVING startyear != 0 and endyear != 0 and endyear >= startyear
```

217 *CBDB*, Alle Angaben beziehen sich auf die Version vom April 2017. Für einen ausführlichen Einblick in die Geschichte und Entwicklung von *CBDB* siehe zudem `https://projects.iq.harvard.edu/cbdb/history-of-cbdb`.

218 *CBDB* liegt als *SQLite*-Datenbank vor. Zur Konvertierung von *SQLite* in *mySQL* kommt ein durch den Verfasser modifiziertes *Perl*-Script zum Einsatz: Shalmanese 2008: „Quick easy way to migrate SQLite3 to MySQL?“ In: *Stack Overflow*. URL: `http://stackoverflow.com/questions/18671/quick-easy-way-to-migrate-sqlite3-to-mysql` (besucht am 10. 07. 2016), Das Script liest alle Zeilen einer .sqlite-Datei und führt den Syntaxunterschieden zwischen den beiden *SQL*-Dialekten entsprechende Ersetzungen durch, ecke Klammern um *SQLite*-Spaltennamen werden z. B. durch *Backticks* (`) ersetzt.

219 In unterschiedlichen Ausgaben / Versionen der Datenbank unterscheidet sich der Aufbau teils marginal. Die Anzahl von 85 Tabellen bezieht sich auf hier verwendete Version vom 24. April 2017.

220 Ermittelt per `select (select count(c_personid) from biog_main where c_birthyear != 0 and c_deathyear! = 0) / (select count(c_personid) from biog_main)`.

221 Für die „Berechnung des Indexjahrs“ gibt es ein komplexes, statistisch und mathematisch fundiertes Regelwerk. Siehe China Biographical Database Project 2013: *Rules for Index Years*. URL: `https://projects.iq.harvard.edu/cbdb/supporting-documents` (besucht am 30. 11. 2017).

222 Dabei werden bevorzugt die Lebensdaten geladen. Wenn diese nicht zur Verfügung stehen, wird das früheste (späteste) Jahr der Nennung verwendet. Wenn dieses ebenfalls nicht zur Verfügung steht, wird auf das Indexjahr ausgewichen. Dabei ist zu beachten, dass die Werte `null` und `0` in der *CBDB* leider austauschbar verwendet werden. Datensätze ohne biographische Daten werden ausgeschlossen und eine minimale Plausibilitätsprüfung gemacht.

— 2. `altname_data` enthält alternative Namen der in `biog_main` geführten Personen. Die Art des *bieming* ist in der Spalte `c_alt_name_type_code` angegeben; `altname_codes` enthält die dazugehörige Liste der Arten alternativer Namen.

— 3. `text_data` enthält *n : m* Zuordnungen von Texten zu Personen, über die jeweiligen IDs `c_textid` und `c_personid`. In der Spalte `c_role_id` wird die Zuordnung einer Person zum Text klassifiziert, wodurch zwischen Autorschaft und Herausgeberschaft usw. unterschieden werden kann.[223]

— 4. `text_codes` enthält Informationen über Texte, u. a. den Titel in Langzeichen (`c_title_cn`), der teilweise auch die Anzahl der *juan* (z. B. „寧波府志: 三十六卷“) beinhaltet. Oft ist eine westliche Umschrift des Titels (`c_title`) angegeben, manchmal eine englischsprachige Übersetzung (`c_title_trans`). In `c_text_year` ist für 6.056 der aufgeführten 28.648 Texte (21,1 %) ein Jahr der Veröffentlichung angegeben.

— 5. `addresses` enthält Namen von administrativen Einheiten bzw. Ortsnamen von Städten, Provinzen und Ländern mit Angaben zur ältesten ermittelten Nennung (`c_firstyear`), die allerdings als unzuverlässig eingestuft werden müssen.[224]

Mit überschaubarem Aufwand lassen sich also Personen-, Text- und Ortsnamen mit chronologischen Daten aus der *CBDB* extrahieren. Wegen der bereits erwähnten Ambiguitäten chinesischer Namen ist ein kritischer Umgang mit den erhaltenen Daten geboten. Um ihre Verwendung für Datierungsaufgaben bewerten zu können, wird eine Statistik zur Länge von Namen erhoben und die erwähnte Problematik multipler Namensträger und lexikalisierter Namensbestandteile kurz beleuchtet.
Ein Großteil der 226.751 unterschiedlichen Namen in der Tabelle `biog_main`[225] hat 2–3 Zeichen (219.733 bzw. 96,9 %), davon bestehen 78.101 aus zwei, 141.632 aus drei Zeichen (Abb. 4.8).[226] 45.384 der betrachteten Namen kommen zweimal oder häufiger vor (20 %), wobei Namen aus zwei Zeichen mit 24.912 (31,9 %) einen deutlich höheren Anteil an Duplikaten aufweisen. Auch unter den Namen mit einer Länge von drei Zeichen kommt aber ein signifkanter Anteil mehrfach vor (20.116 bzw. 14,2 %).[227] Die verzeichneten *bieming* werden ebenfalls zu einem Viertel (25,5 %) von mehr als einer Person getragen.[228]

223 In der Tabelle `text_role_codes` werden die unterschiedlichen Rollen aufgeschlüsselt: 0 für *unknown*, 1 für *author*, 2 für *editor*, 3 für *compiler* usw. Die übrigen Rollen werden hier nicht berücksichtigt.

224 Siehe dazu Kapitel 6.2.2, ab S. 189.

225 Namen mit zusätzlichen Angaben in Klammern werden von dieser Erhebung ausgeschlossen, wie z. B. Zhou *shi* (Jiang Qing *mu*) 周氏 (姜清母) („Frau Zhou, Mutter von Jiang Qing“). Davon betroffen sind 42.686 unterschiedliche Namen in 42.966 Einträgen. Grundgesamtheit der Analyse sind die verbleibenden 323.610 Einträge.

226 Namen mit einer Länge von 5–17 Zeichen kommen hier zumeist durch Transliteration zu Stande, z. B. bei mandschurischen Namen wie Boerjijite E'erzheyitemuere'erkebabai 博爾濟吉特鄂爾哲伊特穆爾額爾克巴拜 (1747–1793), ein Enkel von Kaiser Qianlong 乾隆 (reg. 1736–1796), dessen Geburtsname Aixinjueluo Hongli 愛新覺羅弘曆 mit sechs Zeichen geschrieben wird.

227 Nur eines von zahllosen Beispielen: Neben dem song-zeitlichen Universalgelehrten (1223–1296) verzeichnet die *CBDB* noch zwei weitere Personen mit dem Namen Wang Yinglin 王應麟 (gest. 1515; 1545–1620). Vgl. *CBDB*, IDs 19.880, 313.052, 126.851.

228 Siehe *CBDB*, 18.228 von ingesamt 71.575 verzeichneten unterschiedlichen *bieming* sind Duplikate. (Eigene Berechnung).

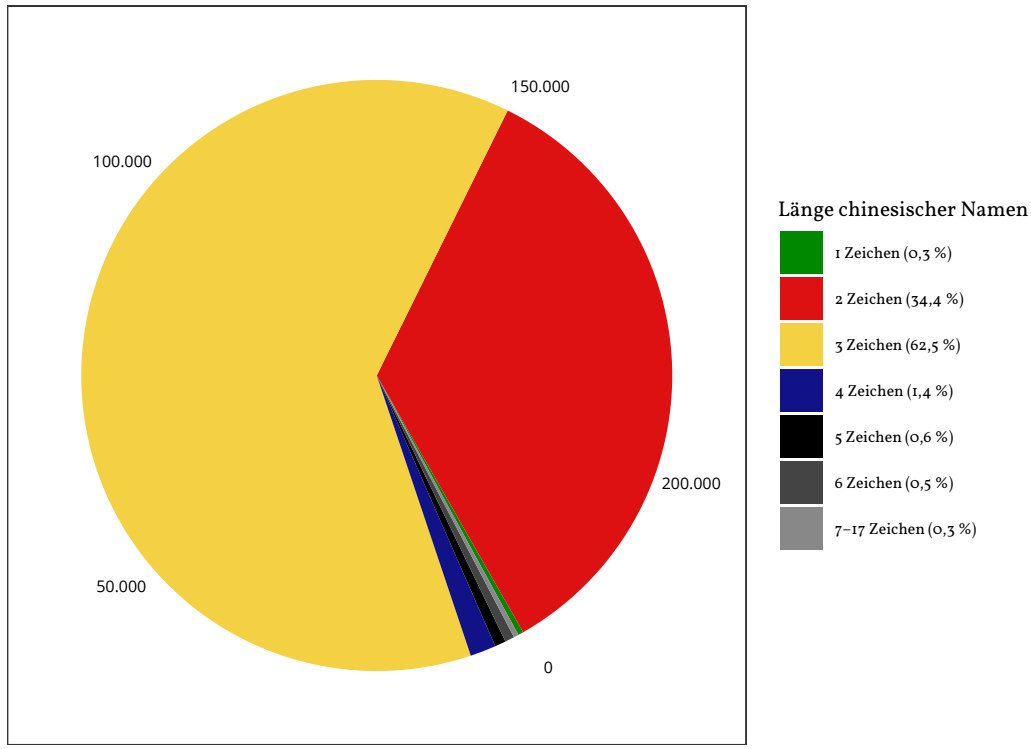


Abbildung 4.8 Länge unterschiedlicher Namen in der *CBDB* in Zeichen, anteilig

Besonders Namen mit zwei, aber auch Namen mit drei Zeichen können daher ambig sein und zu *false positives* führen. Dies kann anhand des *Shiji* 史記 (fertiggestellt 91 v. u. Z.) beispielhaft quantifiziert werden. Darin entsprechen 1.025 unterschiedliche Zeichenkombinationen ein-eindeutigen Namen, die in der *CBDB* verzeichnet sind.[229] Die Lebensdaten von 1.000 der zugehörigen Personen (97,6 %!) sind später als die Datierung des Textes, wobei 919 dieser *false positives* eine Länge von zwei Zeichen haben.

Neben einer höheren Wahrscheinlichkeit für mehrfache Namensträger gibt es einen weiteren wichtigen Grund für den hohen Anteil an *false* positives bei Namen mit einer Länge von zwei Zeichen, denn die Zeichen können auch als Wort oder Wortfolge im Text auftreten. Ein geringes Risiko solcher *false positives* ist auch bei dreisilbigen Namen vorhanden. Ein Beispiel aus *juan* 47 des *Shiji*: „[…] 武王在鎬 […]“, dort wörtlich „…König Wu 武 befindet sich in Hao 鎬,…“ – 王在鎬 Wang Zaigao ist zugleich der Name eines Qing-zeitlichen Autors (1724–1777). Es liegt daher nahe, nur wirklich „eindeutige“ Namen mit drei oder mehr Zeichen zum Zweck der Textdatierung einzusetzen. Der beschriebene Datensatz wird dadurch auf 83.680 Personen eingeschränkt, wobei das gerade umrissene Fehlerrisiko verbleibt.

229 Eigene Erhebung anhand des Sima Qian 司馬遷 2008 [91 v. u. Z.] und der *CBDB*.

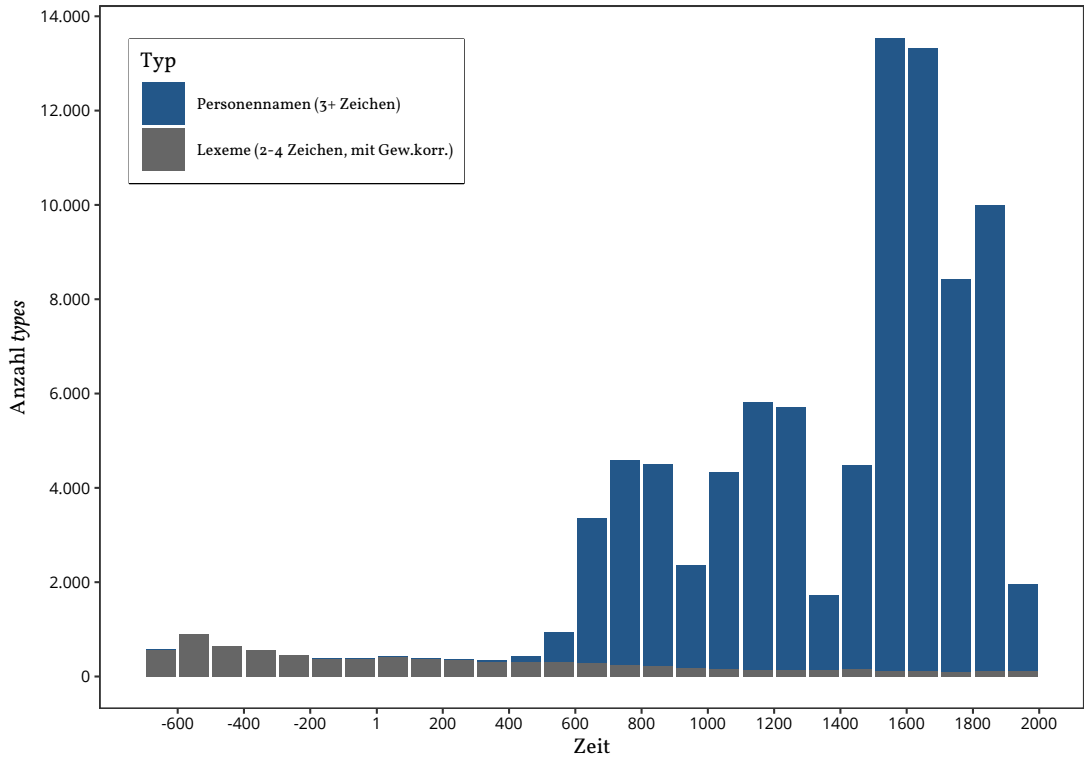


Abbildung 4.9 2–3 Gramme einzigartiger Namen in der *CBDB* nach Jahrhundert

Umgekehrt können natürlich auch Namen oder Teile von Namen fälschlich im Text als Wort erkannt werden, insbesondere wenn anstelle einer sauberen Tokenisierung eine *n*-Gramm Zerlegung der Texte durchgeführt wird, bzw. ein Teil der genutzten Korpora lediglich als *n*-Gramme zur Verfügung stehen.[230] In einer Liste von 129.581 einzigartigen *CBDB*-Namen mit 2–3 Zeichen finden sich Übereinstimmungen mit 9.066 Lexemen mit zwei und 23 Lexemen mit drei Zeichen.[231] Diese Zeichenkombinationen sind also potenzielle *false positives* für Lexeme, da sie auch als Name, bzw. als Bestandteil eines Namens auftreten können. Ihr Anteil und die Verteilung ihrer chronologischen Zuordnung wird auch in Abb. 4.9 deutlich. Die Graphik veranschaulicht gleichzeitig die chronologisch-inhaltliche Verteilung der *CBDB*.[232] Darin sind verhältnismäßig wenige Datensätze zur frühen Kaiserzeit enthalten. Erst ab dem 6. Jh. ist eine relevante Menge biographischer Daten verzeichnet. Der Schwerpunkt der Datenbank liegt eindeutig in der späten Kaiserzeit, während der Dynastien Ming und Qing (Abb. 4.9).
Die ein-eindeutigen Namen in der *CBDB* werden zum einen genutzt, um einen Teil der Belegstellen im *DHYDCD* zeitlich (genauer) einzuordnen.[233] Die so verdichteten Daten werden primär für die in Kapitel 6.2 und 6.3[234] vorgestellten Datierungsmethoden genutzt. Zur Erzeugung tem-

230 Siehe dazu Abschnitt 4.5.2, ab S. 91, bzw. 4.2, ab S. 62.
231 Berechnet als Abgleich der 2–3 Gramme der Namensliste mit den datierbaren Lexemen des *DHYDCD*. Siehe dazu auch Kapitel 5.5, ab S. 120.
232 Zum Aufbau der gewählten Darstellung siehe Kapitel 6.2, ab S. 179, insb. auch 6.2.2, ab S. 189.
233 Siehe dazu Kapitel 5.5.3, ab S. 132; vgl. auch 5.5.2, ab S. 128.
234 Siehe ab S. 179 bzw. ab S. 210.

poraler Textprofile[235] können auch die biographischen Daten aus der *CBDB* direkt verwendet werden. Unter Berücksichtigung der oben festgestellten Ambiguitäten werden jedoch nur eineindeutige Namen mit drei oder mehr Zeichen berücksichtigt, um den Anteil an *false positives* gering zu halten. Für die Datierung mithilfe statistischer Sprachmodelle können Personen- und Ortsnamen ebenfalls als *types* genutzt werden – eine entsprechende Restriktion ist hier nicht erforderlich.[236]

4.8 *Temporal Expressions* und die *Time Authority Database*

Absolute Zeitangaben in schriftsprachlichen Texten werden nicht im Format des gregorianischen Kalenders gemacht, sie lassen sich also nicht – wie z. B. Jahreszahlen – mit einfachen regulären Ausdrücken erkennen.[237] Jahresangaben findet man im Format *RegierungsdeviseJahr*, z. B. *Jiayou yi nian* 嘉祐一年.[238] Die Übertragung des ersten Jahres der Ära *Jiayou* („Gepriesener Schutz“) von Kaiser Renzong von Song (宋仁宗, reg. 1022–1063) in eine westliche Zeitangabe – das Jahr 1056[239] – erfordert Wissen über Dynastien, Herrscher und Regierungsdevisen (*nianhao* 年號). Für eine genauere Ermittlung von Monat oder Datum muss zudem der 60er-Zyklus des lunisolaren Kalenders (*tiangan dizhi* 天干地支-System) bemüht werden.[240] Während man hierfür lange auf Tabellen in Nachschlagewerken angewiesen war,[241] wird diese Arbeit mit dem online verfügbaren Umrechnungstool der Academia Sinica bedeutend erleichtert.[242] Eine darauf basierende, offene Datenbank, die in eigene Anwendungen integriert werden kann, wird von Marcus Bingenheimer et al. beschrieben.[243]

Dharma Drum Buddhist College Time Authority Database

Die *Dharma Drum Buddhist College Time Authority Database* (*DDBC*) enthält Tabellen mit Daten zu chinesischen Dynastien, Herrschern, Äranamen (Regierungsdevisen), Jahren und Monaten.[244] Für die westliche Entsprechung der Zeitangaben wird der julianische Tag (*Julian Day*) angegeben, da so präzise, einheitliche, tagesgenaue Angaben von Zeiträumen als Ganzzahlen gespeichert

235 Siehe dazu Kapitel 6.2.2, ab S. 189.
236 Siehe Kapitel 6.1, ab S. 156.
237 Siehe auch Kapitel 3.3, S. 46.
238 Siehe z. B. Shen Kuo 沈括 2008 [1088]: *Meng xi bi tan* 夢溪筆談 *(Pinselunterhaltungen am Traumbach)*. Project Gutenberg eBook. URL: http://www.gutenberg.net (besucht am 10. 09. 2018) (im Folgenden zit. als *MXBT*), *juan* 卷 25.
239 Siehe Buddhist Studies Authority Database Project 佛學規範資料庫, Hrsg. 2010–2020: *Dharma Drum Buddhist College Time Authority Database*. GitHub Repository. New Taipei City 新台北市. URL: https://github.com/DILA-edu/Authority-Databases/tree/master/authority_time (besucht am 17. 10. 2020) (im Folgenden zit. als *DDBC*), Nr. 26930.
240 Für eine ausführliche, gut verständliche Einführung in Zeitangaben in chinesischen Texten, siehe z. B. Wilkinson 2000, S. 175–184.
241 Vgl. z. B. Tung Tso-Pin 董作賓, Hrsg. 1960: *Chronological Tables of Chinese History*. Hong Kong 香港: Hong Kong University Press.
242 Siehe Academia Sinica, Center for Digital Cultures 中央研究院數位文化中心: *Liang qian nian zhong-xi li zhuanhuan* 兩千年中西曆轉換 *(Umwandlung zwischen chinesischem und westlichem Kalender für 2000 Jahre)*. Website. URL: http://sinocal.sinica.edu.tw/ (besucht am 08. 09. 2019).
243 *DDBC*.
244 Eine ausführliche Beschreibung findet sich in Bingenheimer et al. 2016.

werden können. Durch den Abzug von 1.721.424,5 Tagen kann der entsprechende Tag im gregorianischen Kalender ermittelt werden.[245]

Eine effiziente Implementierung, diese Daten mithilfe komplexer regulärer Ausdrücke zur Erkennung von Zeitangaben zu verwenden, findet sich in *MARKUS*.[246] Dabei werden für den abstrahierten regulären Ausdruck `<nianhao>((<number>)|(<period>)|(<season>)|(<tgdz>)){2,}` zunächst Listen von Ären, Ziffern, Zeitabschnitten, Jahreszeiten und Zykluszeichenkombinationen geladen und damit die in <> gerahmten Begriffe zur Laufzeit ersetzt, also z. B. `<number>` mit `[元正閏一二三四五六七八九十廿卅]{1,}` usw.[247] Damit erkennbare *temporal expressions* müssen also mit einem Äranamen beginnen und können dann beliebige Angaben der anderen Kategorien in der gegebenen Reihenfolge enthalten. Diese werden durch die Klammern in eigenen Gruppen erfasst und können dadurch in gefundenen Ausdrücken wieder separiert werden. Diese Herangehensweise lässt sich problemlos in *Python* adaptieren.

Das durch den regulären Ausdruck beschriebene Muster sei kurz an einem Beispiel veranschaulicht: „*Yingshun yuan nian si yue jiu ri jimao* 應順元年四月九日己卯“[248] („24. Mai 934, *jimao*“) wird zum einen insgesamt als *temporal expression* erkannt. Zudem können 應順 (Gruppe 1), 元, 四, 九 (Gruppe 3), 年, 月, 日 (Gruppe 4) und 己卯 (Gruppe 6) extrahiert werden. Dies ermöglicht nun eine gezielte Abfrage auf die *DDBC*, wobei zur Ermittlung des Jahres eine Einschränkung auf Jahr und Ära ausreicht:

```
select m.id, m.year, m.month_name, ceil((m.first-1721424.5)/365.25) as startyear,
   ceil((m.last-1721424.5)/365.25) as endyear, d.type, m.ganzhi, en.name as era_name,
   hm.name as emperor, dn.name as dynasty
from ddbc_time.t_month m
   left join ddbc_time.t_era e on e.id = m.era_id
   left join ddbc_time.t_era_names en on e.id = en.era_id
   left join ddbc_time.t_emperor h on e.emperor_id = h.id
   left join ddbc_time.t_emperor_names hm on h.id = hm.emperor_id
   left join ddbc_time.t_dynasty d on h.dynasty_id = d.id
   left join ddbc_time.t_dynasty_names dn on d.id = dn.dynasty_id
where en.name = '應順'  and year = 1 and type = 'chinese'
group by e.id, d.id
order by m.era_id, m.year, m.month;
```

Durch Gruppierung auf Herrschernamen und Dynastien werden Ergebnisse mit identischen Jahreszahlen in der Regel herausgefiltert. Im Beispiel des Jahres 934, in dem mehrere Herrscherwechsel stattgefunden und mehrere Herrscher, unter anderem von *Wuyue* 吳越 (907–978) und *Houtang* 後唐 (923–937) zeitgleich an der Macht waren, liefert die Abfrage mehrere Ergebniszeilen, die jedoch alle auf das Jahr 934 verweisen.[249]

Werden *temporal expressions* nicht aus vollständigen Texten extrahiert, sondern aus *n*-Gramm-Häufigkeiten mit limitiertem *n*, können *nianhao* immer noch problemlos erkannt werden. In vielen Fällen sind solche Angaben jedoch mehrdeutig. So gab es in der chinesischen Geschichte fünf unterschiedliche Ären mit dem euphemistischen Namen *Yongping* 永平 („Ewiger Frieden“), verteilt über einen Zeitraum zwischen 58 u. Z. bis 911 – die Angabe *Yongping yuan nian* 永平元年 („das erste Jahr der Ära *Yongping*“) könnte also gleichermaßen auf die Jahre 58, 291, 452, 508

245 Vgl. Nachum Dershowitz und Edward M. Reingold 2008: *Calendrical Calculations*. 3. Aufl. Cambridge & New York: Cambridge University Press, S. 16–17. Der so berechnete Tag lässt sich wiederum in Jahre konvertieren, indem durch die Dauer eines Jahres, 365,25 Tage, geteilt wird.

246 Die dortige Verwendung kann als *temporal tagging* bezeichnet werden, da die gefunden *temporal expressions* in den Eingabetexten markiert bzw. hervorgehoben werden.

247 Siehe Ho und DeWeerdt. 2014–, in den *JavaScript*-Methoden der `automarkup.html`, sowie `tagRegex.js`.

248 *MXBT*, *juan* 卷 1.

249 Angaben im Text folgen einer „traditionellen“ Zeitrechnung, Datenbankwerten wie „-100“ einer astronomischen Zählung, die ein zusätzliches Jahr 0 vorsieht. Das „Datenbankjahr“ 0 entspricht dabei also der Angabe 1 v. u. Z.

oder 911 verweisen.[250] Für Datierungszwecke müssen solche mehrfach verwendeten *nianhao* also ausgeschlossen werden – insbesondere bei der Verwendung von *n*-Gramm-Daten.

Während Zahlen aus vier arabischen Ziffern, wie z. B. 1999, nicht zwangsläufig Zeitangaben sein müssen, kann bei *nianhao* deutlich sicherer davon ausgegangen werden. Im Gegensatz zu eindeutig identifizierten realen Personen, deren namentliche Erwähnung weit vor ihrer Geburt nahezu ausgeschlossen werden kann, können zeitliche Referenzen mit Jahreszahlen überdies durchaus auch auf einen Zeitpunkt in der (fernen) Zukunft erfolgen, wie z. B. bei der Formulierung von Klimazielen. Bei Angaben historischer Regierungsdevisen ist dies nicht üblich. So erkannte *temporal expressions* geben also zuverlässig einen Zeitraum oder Zeitpunkt in der Vergangenheit, *vor* dem Verfassen des untersuchten Textes an. Damit geben sie nicht nur Aufschluss, über welche Zeit in dem Text geschrieben wird, sondern lassen auch Rückschlüsse über das maximale Alter des Textes selbst zu. Diese Erkenntnisse können insbesondere für die in Kapitel 6.2 vorgestellte Datierungsmethodik genutzt werden.[251] Die *nianhao* an sich können aber auch als *types* in statistischen Sprachmodellen verwendet werden.[252]

250 Siehe *DDBC*.
251 Siehe v. a. Kapitel 6.2.2, ab S. 189.
252 Siehe Kapitel 6.1, v. a. 6.1.1, ab S. 158.

5 Das *Hanyu da cidian* 漢語大詞典 als Datenquelle

„Was ist eines wörterbuchs zweck?
nach seiner umfassenden allgemeinheit kann ihm
nur ein groszes, weites ziel gesteckt sein.“[1]

Jacob Grimm

Das einsprachige *Hanyu da cidian* 漢語大詞典 (*Großes Wörterbuch der chinesischen Sprache*, *HYDCD*)[2] dient als wesentliche Datengrundlage für einen Teil der im Rahmen dieser Arbeit entwickelten Datierungsmethoden und deren Evaluation. Es enthält mit insgesamt etwa 370.000 Einträgen einen umfassenden Wortschatz, der die schriftlich überlieferten Anfänge der chinesischen Schriftsprache (*Shangshu* 尚書, *Shijing* 詩經) aus dem ersten Jahrtausend vor unserer Zeitrechnung bis hin zu Neologismen aus den 1990er Jahren abdeckt. Bei der Auswahl der Worteinträge lautete die Vorgabe, Altes und Neues solle gleichermaßen aufgenommen werden, und dabei [sprachlichem] Ursprung und Entwicklung dieselbe Bedeutung beigemessen werden.[3] Zusätzlich zu den unterschiedlichen Bedeutungen der enthaltenen Lexeme werden zumeist Belegstellen dafür aus Primärquellen zitiert, wobei die Herausgeber in der Regel *versuchen*, die jeweils früheste überlieferte Textstelle am Anfang einer Reihe solcher *attestations* anzugeben.[4] Aus bibliographischen Angaben zu diesen Zitaten lassen sich also chronologische Informationen über die Verwendung dieser Lexeme gewinnen. Das *HYDCD* ist zudem – ungeachtet zahlreicher Bemühungen, es als unzuverlässig darzustellen – das bisher umfangreichste *digital verfügbare* historische Wörterbuch der chinesischen Sprache.

Die digitale Ausgabe des *HYDCD*[5] kann also sowohl als Grundlage für die Erzeugung einer diachronen Lexemdatenbank (Kapitel 5.5, ab S. 120), als auch zur Erzeugung von *chronon*-Korpora auf Basis dieser Datenbank und den im *DHYDCD* als Belegstellen angegebenen Textzitaten (Kapitel 5.6, ab S. 137) verwendet werden.

Einige Sinolog:innen sehen im *HYDCD* eine Art kleineres Plagiat des ähnlich aufgebauten *Zhongwen da cidian* 中文大辭典 *(Großes Wörterbuch des Chinesischen)*[6] das selbst wiederum als Über-

1 Jacob und Wilhelm Grimm 1854: *Deutsches Wörterbuch*. Bd. I. A–Biermolke. Leipzig: S. Hirzel, S. XII.

2 Luo Zhufeng 羅竹風, Hrsg. 1986–1994: *Hanyu da cidian* 漢語大詞典 *(Großes Wörterbuch der chinesischen Sprache)*. Bd. 1-13. Shanghai 上海: Cishu chubanshe 辭書出版社.

3 „古今兼收，源流并重“ *HYDCD*, Bd. 1, S. 1; Yu Zhangrui 余章瑞 1988: „为伊消得人憔悴——记《汉语大词典》的编纂及为其辛勤工作的人们 (Zum Gedenken an Yi Xiao, Erinnerung an die Herausgabe und die Menschen, die hart am *HYDCD* gearbeitet haben)“. In: *Renmin ribao* 人民日报 06.23.

4 Siehe z. B. Henning Klöter 2013: „Chinese lexicography“. In: *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Hrsg. von Rufus H. Gouws et al. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: DeGruyter Mouton, S. 884–893, S. 887.

5 *DHYDCD*.

6 Zhang Qiyun 張其昀 et. al., Hrsg. 1973–1979: *Zhongwen da cidian* 中文大辭典 *(Großes Wörterbuch des Chinesischen)*. Bd. 1–10. Yangmingshan 陽明山: Zhongguo wenhua xueyuan 中國文化學院.

Veröffentlicht in: Tilman Schalmey: *Computerlinguistische Datierung schriftsprachlicher chinesischer Texte*. Heidelberg: HASP, 2023. DOI: https://doi.org/10.11588/hasp.1153

setzung des *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)*[7] betrachtet werden kann.[8] Beide früher erschienenen, konkurrierenden Wörterbücher enthalten deutlich mehr Zeichen- und Worteinträge als das *HYDCD*, obwohl letzterem ein „unprecedented scope" nachgesagt wird.[9] Eine entsprechende Diskussion[10] bleibt an dieser Stelle müßig: Trotz bekannter Schwächen (s. u.) bleibt das *HYDCD* als Datenquelle alternativlos, da vergleichbare Wörterbücher wie das *Zhongwen da cidian* und *Dai Kan-wa jiten* nicht in digitaler Fassung vorliegen.[11] Dem *HYDCD* wird ferner eine „greater awareness of historical principles operative in the development of language"[12] bescheinigt.

Ein Vorzug des *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)* dürfte in der Aufnahme von Personen- und Ortsnamen liegen. Durch eine Ergänzung der *DHYDCD*-Daten um Einträge aus der *China Biographical Database Project (CBDB)*[13] lässt sich dieser Nachteil abschwächen bzw. gänzlich eliminieren.

Die im *HYDCD* implizit enthaltenen Lexikalisierungsdaten auf diese Weise zu nutzen, macht es erforderlich, ein tieferes Verständnis dieser Daten zu gewinnen, um Vor- und Nachteile, sowie Einschränkungen der damit möglichen Ergebnisse sichtbar zu machen. Eine Analyse der erzeugten Daten (Kapitel 5.7, ab S. 138) kann dabei nicht nur weitere Einblicke in die Machart des *HYDCD* selbst geben, sondern auch einige kultur- und vor allem sprachgeschichtliche Entwicklungen sichtbar machen.

Obwohl das *HYDCD* für viele Sinolog:innen ein wichtiges Nachschlagewerk darstellt,[14] ist seine Entstehungsgeschichte bislang – zumal in westlichen Sprachen – kaum bearbeitet worden. Sowohl Wilkinson[15] als auch Hargett[16] und Mair[17] gehen zwar kurz darauf ein, geben aber im Wesentlichen die Informationen wieder, die auch im Vorwort des *HYDCD* selbst zu lesen sind.[18]

Wie die Herausgeber des *HYDCD* ihre Quellen für die sprachgeschichtlich so wesentlichen Belegstellen ausgewählt haben, bleibt darin relativ obskur. Hinzu kommt, dass eine Literaturliste der verwendeten Texte fehlt. Auch Hinweise auf die verwendete Ausgabe des jeweiligen Textes

7 Morohashi Tetsuji 諸橋轍次 1955–1960.
8 Siehe Wilkinson 2000, S. 73; siehe auch Victor H. Mair, Hrsg. 2003: *An Alphabetical Index to the Hanyu da cidian*. Honolulu: University of Hawai'i Press, S. 3.
9 Mair 2003, S. 3.
10 Eine vergleichende Analyse der Abdeckung und Qualität der Einträge findet sich z. B. in James M. Hargett 1990: „Review: Hanyu dacidian 漢語大詞典 by Luo Zhufeng 羅竹風". In: *Chinese Literature: Essays, Articles, Reviews (CLEAR)* 12, S. 138–143. DOI: 10.2307/495232, S. 140–142; Über Fachausdrücke etwa schreibt Hargett: „the glosses dealing with such words in the *Hanyu* were superior in any way to those in the *Zhongwen* and *Daikanwa*" Hargett 1990, S. 142.
11 Tatsächlich liegt das auch das *Dai Kan-wa jiten* seit April 2021 als digitale Ausgabe vor. Bei dieser handelt es sich jedoch lediglich um einen hochwertigen Scan, der nach den 51.110 *Kanji* 漢字 Einzelzeichen-Einträgen durchsucht werden kann. Ein digitaler *Plain*-Volltext existiert weiterhin nicht und eine Suchfunktion für mehrsilbige Lexeme fehlt. Vgl. Morohashi Tetsuji 諸橋轍次 (Komp.), Kamata Tadashi 鎌田正 und Yoneyama Torataro 米山寅太郎 (Rev.), Hrsg. 2021 [1955, 1990, 2000]: *Dai Kan-Wa jiten* 大漢和辞典 *Web* 版 *(Großes Chinesisch-Japanisches Wörterbuch, Online-Ausgabe)*. Bd. 1–13; 15. Tokyo 東京: Taishukan shoten 大修館書店, JapanKnowledge.
12 Mair 2003, S. 3.
13 *CBDB*, siehe Kapitel 5.5.3, ab S. 132, siehe auch Kapitel 4.7, ab S. 97.
14 Die Unverzichtbarkeit des *HYDCD* als sinologisches Hilfsmittel zeigt sich etwa darin, dass Mair einen aufwändig kompilierten *Pinyin*-Index zur Unterstützung des Nachschlageprozesses, den er als „agonizingly protracted" beschreibt, herausgegeben hat. Mair 2003, S. 4; siehe auch Wilkinson 2000, S. 69–73.
15 Siehe Wilkinson 2000, S. 69–71.
16 Hargett 1990, Siehe.
17 Siehe Mair 2003, S. 3–10.
18 Siehe *HYDCD*, Bd. 1, S. 1.

sucht man vergeblich, obwohl sich gerade bei älteren Texten wegen ihrer „fluid nature“[19] unterschiedliche Ausgaben massiv voneinander unterscheiden können.

Um ein präziseres Verständnis für die Datengrundlage zu schaffen, wird im Folgenden auch auf die Entstehungsgeschichte des *HYDCD*, sowie seinen Aufbau und Inhalt eingegangen (Abschnitt 5.3, ab S. 113). Um die Entstehung des *HYDCD* nachzuvollziehen, ist außerdem der Exkurs in die Geschichte eines wichtigen Vorbilds ‚[20] des *Oxford English Dictionary* (*OED*, Abschnitt 5.2, ab S. 111) aufschlussreich, dessen Genese sehr gut erforscht ist.

Da die Umformung der Inhalte zu einer relationalen Datenbank nicht auf Basis der gedruckten Ausgabe geschehen kann, muss für dieses Unterfangen eine digitale Ausgabe verwendet werden, die nicht exakt mit der gedruckten Version übereinstimmt. Ihr Inhalt und ihre Qualität werden deswegen im Stichprobenverfahren auf Übereinstimmung mit der Originalausgabe geprüft (Abschnitt 5.4, ab S. 115).

5.1 Eine kurze Geschichte des *HYDCD*

Das „historische Wörterbuch der chinesischen Sprache“[21] wurde zwischen 1986 und 1993 sukzessive in insgesamt zwölf Bänden veröffentlicht, hinzu kommt ein 1994 erschienener Indexband. Aus dem Wörterbuch selbst geht nur wenig über seine Entstehungsgeschichte und die Herangehensweise der Herausgeber und ihrer Mitarbeiter:innen hervor, sie lässt sich jedoch teilweise aus in der *Renmin ribao* 人民日報 (*RMRB*) erschienenen Zeitungsartikeln nachvollziehen. Das dort gezeichnete Bild ist mit Vorsicht zu genießen, da die Autor:innen der *RMRB* mit teils überzogen verherrlichenden, „patridiotischen“ Formulierungen dem Ruf der Zeitung als Organ der Kommunistischen Partei Chinas (KPCh) an vielen Stellen mehr als gerecht werden.[22] Der Leserschaft wird dabei eine Entstehungslegende vermittelt, die um den früheren Premierminister Zhou Enlai 周恩來 (1898–1976)[23] gewoben wird.[24] In einem Artikel aus dem Jahr 1997 wird sogar noch der gerade verstorbene „Genosse Xiaoping“ (Deng Xiaoping 鄧小平, 1904–1997) als einer der Auftraggeber für das *HYDCD* bemüht.[25] Übereinstimmenden Berichten der *RMRB* zufolge soll Zhou wiederholt in Situationen gekommen sein, in denen er beim Treffen mit anderen Staatsoberhäuptern eindrucksvolle, umfassende Wörterbücher aus dem jeweiligen Land geschenkt bekam, aber kein adäquates Gegengeschenk vorzuweisen hatte.[26] So auch beim Besuch eines Gesandten des Fürstentums Monaco, bei dem von chinesischer Seite als Staatsgeschenk lediglich das *Xinhua zidian* 新華字典 überreicht wurde. Zhou soll diesen Umstand mit den Worten „小國送大書，大國送小書“ („Kleines Land schenkt großes

19 Tao Hongyin 2015.

20 Mair bezeichnet das *HYDCD* sogar als „closest approximation to the *Oxford English Dictionary* that is available“. Mair 2003, S. 3.

21 „*lishixing de Hanyu yuwen cidian* , 历史性的汉语语文词典‘“ *HYDCD*, Bd. 1, S. 1.

22 Man sieht sich unbescheiden in der über 2.000-jährigen Tradition von *Erya* 爾雅, *Shuo wen jie zi* 說文解字 und *Kangxi zidian* 康熙字典, die Mitwirkenden werden für ihre Opferbereitschaft glorifiziert und dramatisch geschildert, wie „mit letzter Kraft“, „mit zitternder Hand“, im Krankenhaus usw. gearbeitet wurde. Siehe z. B. Yu Zhangrui 余章瑞 1988; Guan Xi 冠西 1997: „难忘罗老风范 (Schwer, das Gebaren des alten Luo [Zhufeng 羅竹風] zu vergessen)“. In: *Renmin ribao* 人民日报 11.10.

23 Kai Vogelsang 2012: *Geschichte Chinas*. Stuttgart: Reclam, S. 644.

24 Zhao Lanying 趙蘭英 1986: „《Hanyu da cidian》 bianzuan wancheng《汉语大词典》编纂完成 (Die Kompilation des *HYDCD* ist abgeschlossen)“. In: *Renmin ribao* 人民日报 01.11; Yu Zhangrui 余章瑞 1988; Li Hongbing 李泓冰 1994: „龙飞在天 —— 《汉语大词典》编纂前前后后 (Der Drache fliegt – Die ganze Geschichte hinter der Kompilation des HYDCD)“. In: *Renmin ribao* 人民日报 05.11.

25 Siehe Guan Xi 冠西 1997.

26 Yu Zhangrui 余章瑞 1988.

Buch, großes Land schenkt kleines Buch") kommentiert haben.[27] Vor diesem Hintergrund soll er dann – bereits schwer erkrankt – die nötigen Schritte in die Wege geleitet haben, ein Planungssymposium durchzuführen, aus dem ein entsprechendes Arbeitskomitee unter der Leitung von Chen Hanbo 陳翰伯 (1914–1988)[28] hervorging.[29] Einen entscheidenden Motivationsimpuls für die Politik gab allerdings sicherlich vor allem die Tatsache, dass in Japan, Südkorea und Taiwan bereits umfangreiche Chinesischwörterbücher herausgebracht worden waren.[30]

Ähnliche Planungen für ein umfassendes einsprachiges Wörterbuch der chinesischen Sprache gab es in Festlandchina allerdings bereits deutlich früher. Ab 1928 planten einige Linguisten um Li Jinxi 黎錦熙 (1890–1978)[31] und Chao Yuan Ren 趙元任 (1892–1982)[32] die Herausgabe eines *Zhongguo da cidian* 中國大辭典.[33] Das Projekt verzögerte sich jedoch durch den zweiten sino-japanischen Krieg (*Kang-Ri zhanzheng* 抗日戰爭, 1937–1945) und später durch weitere militärische Auseinandersetzungen und gesellschaftliche Umbrüche. In einem 12-Jahresplan für wissenschaftliche Forschung von 1956 wurde die Idee wieder aufgegriffen, der Sprachwissenschaftler Lü Shuxiang 呂淑湘 (1904–1998)[34] hatte darin bereits die Herausgabe des *HYDCD* vorgesehen. Erneut wurden die Pläne durch politische Kampagnen wie den „Großen Sprung nach Vorne" (*dayuejin* 大躍進, 1958–1962) und die Kulturrevolution (*wenhua da geming* 文化大革命, 1966–1976)[35] vereitelt.[36]

Nach dem tatsächlichen Projektstart in den 1970er Jahren erhielten bekannte Größen der chinesischen Linguistik wie Wang Li 王力 (1900–1986)[37] und Lü Shuxiang wichtige Beraterfunktionen und Luo Zhufeng 羅竹風 (1911–1996)[38] konnte als leitender Herausgeber gewonnen werden.[39] An den ersten beiden Bänden sollen 458 Personen aus mehreren Provinzen in 34 Gruppen gearbeitet haben,[40] andere Artikel sprechen sogar von über 1.000 Sprachwissenschaftler:innen und 43 *danwei*s 單位 („Arbeitseinheiten").[41],[42]

27 Li Hongbing 李泓冰 1994.
28 OCLC 2019: *oclc.org – Worldcat Identities*. Website. URL: https://www.worldcat.org/identities (besucht am 19. 05. 2019), 陈翰伯 1914-1988 (lccn-nr92017850).
29 Siehe Yu Zhangrui 余章瑞 1988; Li Hongbing 李泓冰 1994.
30 Vgl. auch Yu Zhangrui 余章瑞 1988.
31 OCLC 2019, 黎锦熙 1890–1978 (lccn-n82156948).
32 Ebd., Zhao Yuanren, 1892-1982 (lccn-n50036317).
33 Li Hongbing 李泓冰 1994.
34 OCLC 2019, 呂叔湘 (lccn-n79034713).
35 Die Dauer der Kulturrevolution wird von Historikern unterschiedlich bewertet – während sie offiziell schon 1969 (und erneut im August 1977 durch Hua Guofeng 華國鋒, 1921–2008) für beendet erklärt wurde, wird das eigentliche Ende der Bewegung eher auf den Tod von Lin Biao 林彪 (1907–1971) oder Mao Zedong 毛澤東 (1983–1976) datiert. Siehe z. B. Vogelsang 2012, S. 570–577.
36 Yu Zhangrui 余章瑞 1988.
37 OCLC 2019, 王力 1900-1986 (lccn-n81021999).
38 Siehe Guan Xi 冠西 1997.
39 Siehe Yu Zhangrui 余章瑞 1988.
40 Siehe ebd.
41 Der Begriff *danwei* bezeichnet im System der Volksrepublik China eine Art städtische Wohn- und v. a. Arbeitseinheit. In diesem Kontext dürften hier im weiteren Sinne Arbeitsstätten, z. B. Hochschulen gemeint sein. Siehe z. B. Vogelsang 2012, S. 644.
42 Zhao Lanying 趙蘭英 1986; He Jiazheng 何加正 und Li Hongbing 李泓冰 1994: „中华民族五千年文化的结晶中国辞书出版史上的壮举《汉语大词典》大功告成首都隆重举行庆功会江泽民李鹏等到会祝贺全书13卷，收词语37.5万余条，约5000万字，是千余专家学者18年艰苦努力的结果 (Die Quintessenz der 5.000-jährigen Kultur des chinesischen Volkes, die Höchstleistung der chinesischen Geschichte der Herausgabe von Wörterbüchern, das *HYDCD*, wurde endlich abgeschlossen und zu diesem Anlass in der Hauptstadt eine große Feier ausgerichtet. An der Veranstaltung nahmen Jiang Zemin, Li Peng und andere teil. Das Werk hat insgesamt 13 Bände, 375.000 Wörter wurden aufgenommen, etwa 50 Mio. Zeichen, das Ergebnis der harten 18 Jahre dauernden Arbeit von über 1.000 Spezialisten und Gelehrten)". In: *Renmin ribao* 人民日报 05.11.

Angespornt von den zuvor erschienenen *Dai Kan-Wa jiten* und *Zhongwen da cidian*,[43] und dank dem staatlich geförderten immensen personellen und finanziellen Aufwand benötigte man vom „Startschuss“ 1975 bis zur Fertigstellung des letzten Bandes 1994[44] weniger als 20 Jahre. In direkter Konkurrenz mit vergleichbaren Unternehmungen ist das ein verhältnismäßig kurzer Zeitraum. Die *RMRB* prahlt, dass man – gewissermaßen als Ausgleich für das späte Angehen des Projekts – deutlich schneller fertig geworden sei als die Engländer, die Deutschen oder die Russen.[45] Luo Zhufeng 羅竹風 wird zitiert, das *HYDCD* sei das Ergebnis einer groß angelegten sozialistischen Kooperation und verkörpere konkret die Überlegenheit des sozialistischen Systems.[46] Möglich, dass Luo diese Worte posthum in den Mund gelegt wurden – sie zeigen jedenfalls, dass das *HYDCD* eine gewisse Rolle für das festland-chinesische kulturelle Selbstbild spielt und als Vorzeigeprojekt staatlicher Kulturförderung und -propaganda gesehen werden kann. Hierbei scheut man sich auch nicht, die früher erschienene und doch umfangreichere[47] Konkurrenz aus Japan und Taiwan zu diskreditieren – nicht nur sei man viel schneller gewesen, viele Einträge im *Dai Kan-Wa jiten* oder dem *Zhongwen da cidian* seien gar keine richtigen Wörter (*„xuduo bucheng “ci” de ci* 許多不成“詞”的詞“, eine Aussage, die aus linguistischer Sicht sicherlich für das *HYDCD* ebenfalls zutrifft) und der Inhalt jener Werke sei „diffus“.[48]

Ein 2010 in der *RMRB* veröffentlichter Artikel schlägt wieder etwas differenziertere Töne an und betont, dass andere Nationen wie Frankreich, Deutschland, Russland und die Vereinigten Staaten mit vergleichbaren Wörterbuchprojekten durchschnittlich 50 Jahre früher fertig waren, außerdem habe das 100 Jahre früher erschienene *OED* deutlich mehr Einträge.[49]

5.2 Das Vorbild: *Oxford English Dictionary*

> *„...where every pains has been taken to ascertain the earliest occurrence of each word and of each signification...“*[50]
>
> Otto Jespersen

Wegen seines Vorbildcharakters für historische Wörterbücher sei an dieser Stelle auch auf das *Oxford English Dictionary* (*OED*) eingegangen. Seine Geschichte kann zwar schwerlich offene Fragen über die Entstehung des *HYDCD* beantworten, wegen der Ähnlichkeit beider Projekte kann sie aber zumindest zum Verständnis der Konzeption beitragen, denn im Gegensatz zum *HYDCD* ist die Entstehungsgeschichte des *OED* sehr gut dokumentiert und erforscht.[51] Die Planung für

43 Siehe Yu Zhangrui 余章瑞 1988.

44 *HYDCD*, Bd. 13, S. ii.

45 Li Hongbing 李泓冰 1994; Konkret bezieht man sich hier v. a. auf das *OED*, bei dem die Planung der ersten Ausgabe im Jahr 1858 begann, dessen letzter, zehnter Band aber erst 1928 fertiggestellt wurde. Siehe John Willinsky 1994: *Empire of Words: The Reign of the OED*. Princeton, New Jersey: Princeton University Press, S. 11.

46 „《汉语大词典》是社会主义大协作的产物，是社会主义制度优越性的具体体现.“ Siehe Guan Xi 冠西 1997.

47 Eine den Autoren der *RMRB* vorliegende Ausgabe von Morohashis Wörterbuch wird mit 550.000 Einträgen angegeben.

48 *Pang za* 龐雜, wörtlich etwa „riesig und divers“. Siehe Yu Zhangrui 余章瑞 1988.

49 Siehe Zhang Zhiyi 张志毅 2010: „, 辞书强国‘究竟有多远 (Wie weit es noch bis zum ‚starken Wörterbuchland‘ ist)“. In: *Renmin ribao* 人民日报 10.12.

50 Otto Jespersen 1912 [1905]: *Growth and Structure of the English Language*. Leipzig: B. G. Teubner, S. 222; zitiert in Willinsky 1994, S. 57.

51 Abgesehen von den im Folgenden zitierten Werken seien erwähnt: Katherine M. Elisabeth Murray 1977: *Caught in the Web of Words: James Murray and the Oxford English Dictionary*. New Haven & London: Yale University Press, eine Biographie über den Herausgeber James Murray; sowie ein Roman über die Beziehung zwischen einem der Bei-

dieses Mammutprojekt nahm ihren Anfang 1857 mit einem Vorschlag des Poeten und Bischofs von Dublin, Richard Trench auf einer Sitzung der Philological Society of London. Mehrere Jahrzehnte vergingen bis zur Veröffentlichung des ersten Faszikels, das 1884 noch unter dem Titel *A new English Dictionary* erschien, sowie der ersten vollständigen Ausgabe 1933.[52]

Das erklärte Ziel der Herausgeber des *OED* war es, „eine angemessene Darstellung der Bedeutung, des Ursprungs und der Geschichte der englischen Wörter zu liefern, die allgemein gebräuchlich sind oder bekanntermaßen zu irgendeinem Zeitpunkt während der letzten siebenhundert Jahre gebräuchlich waren“[53] – sehr ähnlich den Zielen der Urheber des *HYDCD*. Zur Legitimierung der Einträge wurden zusammen mit unzähligen Freiwilligen mehr als fünf Millionen Zitate gesammelt, von denen 1.827.306 in der Ausgabe von 1933 Verwendung fanden.[54]

Bei der Auswahl eben dieser Belegstellen liegt ein offensichtliches *Bias* vor: William Shakespeare (1564–1616) als am meistzitierter Autor wird in 14 Prozent aller Einträge herangezogen, insgesamt werden seine Werke 32.868 mal zitiert, was fast 2 Prozent aller Belegstellen entspricht.[55] Dieses *Bias*, das sich in ähnlicher Form auch im *HYDCD* beobachten lässt,[56] wurde früh bemerkt und kritisiert:[57]

> […] one is struck by the frequency with which Shakespeare's name is found affixed to the earliest quotation for words or meanings. […] this is no doubt due to the fact that Shakespeare's vocabulary has been registered with greater care in Concordances […] than that of any other author, so that his words cannot escape notice, while the same words may occur unnoticed in the pages of many an earlier author.[58]

Das Konsultieren von Konkordanzen vereinfacht natürlich die Identifikation von Belegstellen für die Kompilator:innen, bringt aber eine gewisse Unausgewogenheit mit sich. Wahrscheinlich ebenfalls im Vorhandensein entsprechender Konkordanzen begründet ist die Häufigkeit, mit der Übersetzungen europäischer Klassiker wie der *Aeneis*, der Bibel und anderer nicht originär englischsprachiger Texte herangezogen werden.[59] Darin liegt zugleich ein auffälliger Unterschied zwischen *OED* und *HYDCD*. Die Einträge in letzterem werden fast ausschließlich mit Texten belegt, die ursprünglich in chinesischer Sprache verfasst wurden.

---

tragenden und Murray. Simon Winchester 1998: *The Professor and the Madman: A Tale of Murder, Insanity, and the Making of The Oxford English Dictionary*. New York: HarperCollins, – Winchester ist ebenfalls bekannt für seinen Roman über den Sinologen, Biochemiker und Wissenschaftshistoriker Joseph Needham.

52 Trenchs ursprüngliche Idee der Bildung eines Komitees zur Sammlung unbekannter Lexeme mit dem Ziel der Erweiterung bestehender Wörterbücher erschien den Mitgliedern der *Philological Society* nach seiner Vorstellung systematischer Mängel eben jener vorhandenen Nachschlagewerke als nicht ausreichend. Daher wurde beschlossen, stattdessen die Arbeiten an einem *New English Dictionary* zu beginnen. Siehe Willinsky 1994, S. 3–16.

53 James A. H. Murray et al., Hrsg. 1913–1933: *Oxford English Dictionary*. Bd. 1–13. London: Oxford University Press (im Folgenden zit. als *OED*), Bd. 1, S. vi, übersetzt durch den Verfasser. zitiert in Willinsky 1994, S. 76.

54 Siehe Willinsky 1994, S. 3–4, S. 58.

55 Siehe ebd., S. 57–58. Im Anhang (ab S. 211) finden sich hier Tabellen mit genauen Zählungen der am häufigsten zitierten Autoren, Werke, Zeitschriften und Tageszeitungen unterschiedlicher Ausgaben.

56 Siehe Kapitel 5.7, ab S. 138.

57 Als Gegenargument sei an dieser Stelle aus Jacob Grimms Vorwort zum *Deutschen Wörterbuch* zitiert, das die große Zahl Shakespeare-Zitate in besserem Licht erscheinen lässt: „Hin und wieder wird man der belege zu viel angebracht meinen, namentlich aus Luther und Göthe. doch jenes einflusz auf die sprache, Göthes macht über sie müssen reich und anschaulich vorgeführt werden […]“ Grimm 1854, S. xxxvii.

58 Jespersen 1912 [1905], S. 222; zitiert in Willinsky 1994, S. 57.

59 Siehe Willinsky 1994, S. 115.

5.3 Aufbau und Inhalt des *HYDCD*

Eine kurze Beschreibung von Struktur, Aufbau und Machart des *HYDCD* soll über Inhalt, Umfang und Qualität der Daten Aufschluss geben, die für die Textdatierung nutzbar gemacht werden sollen. Die Einträge lassen sich in zwei Hauptkategorien unterteilen:

— 1. **Einträge für einzelne Schriftzeichen**, *dan zi tiaomu* 單字条目, sind gewissermaßen Haupt- oder Übereinträge.[60] Diese sind in 200 Radikale unterteilt,[61] sowie nach Radikal- und Zusatzstrichzahl (*residual strokes*) sortiert.[62] Diese Einträge bezeichne ich im Folgenden als „Zeicheneinträge". In der digitalen Ausgabe werden sie von einem Asterisk (*) eingeleitet. Bei Schriftzeichen, für die nicht nur mehrere Bedeutungen, sondern auch unterschiedliche Aussprachen angegeben sind (*duoyinzi* 多音字), werden die Einträge mit (in der gedruckten Ausgabe hochgestellten) arabischen Ziffern durchnummeriert, z. B. *zǐ* 仔 ¹, *zī* 仔 ² und *zǎi* 仔 ³.[63]

Im Unterschied zum *Hanyu da zidian* 漢語大字典 *(Großes Lexikon chinesischer Schriftzeichen)*[64] liegt der Fokus des *HYDCD* auf *ci* 詞, so dass deutlich weniger Einzelzeicheneinträge bestehen als Schriftzeichen bekannt sind. Das Auswahlkriterium für monosyllabische Einträge soll die fortwährende Verwendung der Zeichen gewesen sein, wohingegen im *Hanyu da zidian* auch *si zi* 死字, also „ausgestorbene" Zeichen, aufgenommen wurden.[65]

— 2. **Einträge mit mehreren Schriftzeichen**, *duo zi tiaomu* 多字条目 (wörtlich etwa: „Mehrzeicheneinträge")[66] sind den Zeicheneinträgen des jeweils ersten Zeichens untergeordnet. Im Folgenden bezeichne ich diese Einträge als „Unter-" oder „Worteinträge". Sie enthalten in erster Linie mehrsilbige „Wörter", d. h. *ci*, bzw. „strings of monosyllabic morphemes".[67] In diese Kategorie fallen zum Teil aber auch lexikalisierte Phrasen.[68]

Das *HYDCD* folgt damit der grundlegenden Terminologie der chinesischen Lexikographie, einer Dichotomie zwischen *zi* 字 und *ci* 詞.[69] Aus morphologischer Sicht können beides „Wörter" sein. Aufgrund der „inherent fuzziness" des in der Linguistik allgemein und für das Chinesische im Besonderen problematischen Wortbegriffs[70] wird hier für alle Zeichen- und Zeichenfolgen mit Einträgen im *HYDCD* der Begriff *Lexeme* verwendet.

In einem typischen Eintrag folgt auf das Stichwort eine Auflistung unterschiedlicher Bedeutungen. Ähnlich wie im *OED* und vergleichbaren Wörterbüchern wie *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)*, wird meist die Bedeutung erklärt, z. B. durch Angabe

60 Vgl. *HYDCD*, Bd. 1, S. 7.
61 Siehe *HYDCD*, Bd. 1, S. 12.
62 Vgl. *HYDCD*, Bd. 1, S. 7. Im *HYDCD* etwas umständlich: „bei gleichem Radikal nach Strichzahl (abzüglich der Strichzahl des Radikals selbst)" („部首相同的按画数(减去部首本身画数)").
63 *HYDCD*, Bd. 1, S. 7.
64 Xu Zhongshu 徐中舒, Hrsg. 1986–1990: *Hanyu da zidian* 漢語大字典 *(Großes Lexikon chinesischer Schriftzeichen)*. 3 Bde. Wuhan 武漢: Sichuan cishu chubanshe 四川辭書出版社, Hubei cishu chubanshe 湖北辭書出版社 (im Folgenden zit. als *HYDZD*).
65 Vgl. Yu Zhangrui 余章瑞 1988.
66 Vgl. *HYDCD*, Bd. 1, 7.
67 Norman 1988, S. 24.
68 Sprichwörter, sowie Phraseologismen wie *chengyu* 成語, *xiehouyu* 歇後語, und *suyu* 俗語. Siehe z. B. *HYDCD*, Bd. 1, S. 428, *bushi dongfeng yaliao xifeng, jiushi xifeng yaliao dongfeng* 【不是東風壓了西風，就是西風壓了東風】 od. Bd. 6, S. 1311, *zhao san mu si* 【朝三暮四】.
69 Eine aktuelle, ausführliche Diskussion dieser Problematik findet sich in Jiang Shaoyu 蔣绍愚 2015, S. 1–4; siehe auch Klöter 2013, S. 885.
70 Vgl. Lukáš Zádrapa 2015: „Word and Wordhood in Classical Chinese". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

von Synonymen, gefolgt von einer diachronen Reihe von Textbeispielen, die als Belegstellen zitiert werden – eine Struktur, die so erstmals im ab 1836 erschienenen *New Dictionary of the English Language* Verwendung findet.[71] In Wörterbüchern Bedeutungen mit illustrativen Textbeispielen zu belegen, ist keineswegs eine aus Großbritannien übernommene Innovation: in der chinesischen Lexikographie hat diese Praxis eine deutlich längere Tradition – bereits im *Erya* 爾雅[72] lassen sich solche Zitate aus Klassikern finden, auch wenn die Quellen nicht explizit angegeben werden.[73] Das 1710–1716 kompilierte *Kangxi Zidian* (*KXZD*) bietet bereits Belegreihen an, die aber kaum chronologisch geordnet sind und überwiegend frühe, kanonisierte Texte zitieren.[74] Aufgeführt werden dabei zudem Verwendungen einzelner Zeichen (*zi* 字), Zeichenverbindungen werden nicht systematisch berücksichtigt.[75]

Die Belegstellen im *HYDCD* stammen aus den unterschiedlichsten Textgattungen, darunter kanonisierte philosophische Klassiker, Geschichtstexte, Gedichte, Romane, Zeitungsartikel – bis hin zu Veröffentlichungen der kommunistischen Partei. Im Unterschied zum *OED* werden die Textbeispiele weder kontinuierlich für jedes Jahrhundert, in dem sich ein Wort nachweisen lässt, gegeben, noch die Erscheinungsjahre der Erstausgabe der zitierten Texte angegeben. Die Kompilator:innen des *HYDCD* begnügen sich mit Angabe von Dynastie und Autor:in, bei Kanontexten sogar mit dem Titel des zitierten Textes.[76] Trotz der Vorbildfunktion des *OED* sollte nicht vergessen werden, welches Sprachverständnis die Kompilator:innen des *HYDCD* hatten und vor allem, in welcher Tradition sie stehen. Die Selbstverständlichkeit, mit der Texte wie *Shangshu* oder auch die Dynastiegeschichten hier ohne Angabe von Autor:in, Ausgabe oder Dynastie zitiert werden, kann in diesem Licht auch mehr als Traditionsbewusstheit, denn als sprachwissenschaftliche Verfehlung erscheinen.

Wie im *OED* versucht man dabei, den *Locus classicus* ausfindig zu machen und als Beleg anzugeben.[77] Es liegt auf der Hand, dass dieses Unterfangen nicht immer gelingen kann. So hat sich in der chinesischen Lexikographie inzwischen gewissermaßen ein eigenes Aufsatzgenre etabliert, dessen Hauptinhalt die Ergänzung noch früherer Belegstellen (*ante-dating*) zu *HYDCD*-Einträgen ist.[78] Auch Li Shens 李申 Monographie *Hanyu da cidian yanjiu* 《汉语大词典》研究 *(A*

71 Siehe Charles Richardson 1836: *A New Dictionary of the English Language*. London: W. Pickering; erwähnt in Willinsky 1994, S. 21, S. 29, S. 94. Während Richardson in der ersten vollständigen Ausgabe von 1836 noch sehr spärlich mit Belegen umgeht, sind sie in der mehrbändigen, ab 1851 erschienenen Ausgabe bei einem Großteil der Einträge vorhanden. Vgl. Charles Richardson 1851: *A New Dictionary of the English Language. 2 vols.* Philadelphia: E. H. Butler & Co.; Eine sorgfältigere Zitierweise mit Jahres- und Seitenzahlen findet sich allerdings in diesem Kontext zuerst im *Deutschen Wörterbuch*, dessen erste Lieferung nur kurze Zeit später, 1852 erfolgte. Im Vorwort des ersten, vollständigen Bandes erläutert Jacob Grimm zu den Belegen: „[…] der name ihres urhebers reicht nicht aus, sie müssen aufgeschlagen werden können […]". Grimm 1854, S. xxxvi.

72 Der Titel wird im Englischen mit „approaching what is correct, proper, refined" wiedergegeben. South W. Coblin 1993: „Erh ya 爾雅". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 94–99, S. 94.

73 Siehe z. B. Yong Heming und Peng Jing 2008: *Chinese Lexicography: A History from 1046 BC to AD 1911*. Oxford: Oxford University Press, S. 90–91.

74 Nach-hanzeitliche Quellen werden eher vereinzelt zitiert, z. B. ein *ci* 詞 des tangzeitlichen Dichters Li Bai (701–762) und die 1343 veröffentlichte *Yuan shi* 元史 im Eintrag zu *jin* 金. Siehe Aisin-Gioro Xuanye 愛新覺羅 · 玄燁 (als Shengzu ren huangdi 聖祖仁皇帝) 1922 [1716], S. 1295.

75 Zum Aufbau des *KXZD* siehe z. B. auch Marc Winter 2015: „Kāngxī zìdiǎn 康熙字典". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

76 Zur Veranschaulichung von Inhalt und Struktur sei auf die vom Verfasser kommentierten Beispiele weiter unten (Abschnitt 5.5.2, ab S. 120) verwiesen.

77 In 2.450 Einträgen des *DHYDCD*, also bei einem Anteil von unter einem Prozent, findet sich explizit die Angabe *yuben* „ 语本…" („die Redewendung / das Wort hat seinen Ursprung bei / in…"). Da diese Markierung aber selten und nicht systematisch eingesetzt wird, widme ich ihr hier keine weitere Aufmerksamkeit. Vgl. *DHYDCD*, *passim*.

78 Vgl. z. B. Liu Bing 劉冰 2009: „《汉语大词典》书证迟后例补——以《先秦漢魏晉南北朝詩 (梁詩)》为例 (Ergänzungen für späte Belegstellen im *HYDCD* - anhand von Gedichten der Prä-Qin, Han, Wei, Jin und Nanbei-Zeit [Liang

study on Hanyu da cidian) widmet ein langes Kapitel der Ergänzung früherer Belegstellen.[79] Selbst wenn der wirkliche *Locus classicus* von den Herausgebern nicht gefunden wurde, lässt sich dennoch mit Sicherheit sagen: wenn eine Bedeutung z. B. mit einem Han-zeitlichen Text belegt wird, kann angenommen werden, dass das Lexem mindestens zu dieser Zeit so verwendet wurde. Die früheste Belegstelle impliziert damit ein Mindestwortalter.

Etliche Aufsätze befassen sich zudem mit der Identifizierung von Wörtern, die nicht als Einträge aufgenommen wurden, sowie der Ergänzung und Korrektur von Wortbedeutungen.[80] Bei aller berechtigten Kritik gilt das *HYDCD* dennoch als „most authoritative dictionary of the Chinese language."[81]

Eine herausfordernde Eigenheit des *HYDCD* ist die gemischte Verwendung von Lang- und Kurzzeichen. Die Herausgeber rechtfertigen diese Maßnahme mit der Historizität des Wörterbuchs. Die Worterklärungen sind grundsätzlich in Kurzzeichen (*jiantizi* 簡 [简] 體 [体] 字) verfasst, doch Zeichen- und Worteinträge selbst werden in Langzeichen (*fantizi* 繁體字) angegeben.[82] Vereinfachte Zeichenformen erhalten einen zusätzlichen Zeicheneintrag, der auf das entsprechende Langzeichen verweist.[83] Nicht nur Glossen, auch Namen von Autor:innen, Herausgeber:innen, sowie Titel von zitierten Werken und Dynastien, sind in Kurzzeichen gesetzt. Eine Ausnahme davon bilden Namen wie Gao Shi 高適, bei denen es so zu Mehrdeutigkeiten kommen kann – hier wird z. B. nicht 高适 geschrieben, da 适 *kuò* und 適 *shì* ein gemeinsames Kurzzeichen (适) teilen. Die Belege werden wiederum in Langzeichen gesetzt, sofern der zitierte Primärtext vor 1912 entstanden ist, oder selbst in Langzeichen verfasst wurde.[84]

5.4 Digitale Ausgaben des *HYDCD*

Voraussetzung für die Nutzung des *HYDCD* als Datengrundlage für Softwareprojekte ist eine digitale Ausgabe. Es kursieren drei unterschiedliche, zwischen 1997 und 2007 veröffentlichte CD-Rom Versionen, sowie mehrere Online-Veröffentlichungen.[85] Da keine der offiziellen Ausgaben als Volltext verfügbar ist, greife ich auf eine Online-Version (im Folgenden *DHYDCD*) zurück, die inhaltlich weitestgehend der gedruckten Ausgabe entspricht und in der eben-

Gedichte])". In: *Yuwen Xuekan* 语文学刊 *(Journal of language and literature studies)* 19, S. 72–73, S. 72–73; Zheng Xianzhang 郑贤章 2000: „《Hanyu da cidian》shuzheng chushi li shi bu《汉语大词典》书证初始例试补 (Supplementing some Earlier Citations to Hanyu da cidian)". In: *Gu Hanyu yanjiu* 古汉语研究 *(Research in Ancient Chinese Language)* 2, S. 94–96, S. 94–96; Siehe auch Jing-Schmidt und Hsieh 2019, S. 516 „Studies of the translated religious texts and the vernacular materials preserved at Dunhuang have enabled lexicographers to trace the attestations of many words to an earlier time point than indicated in the standard dictionary Hànyǔ Dàcídiǎn".

79 Li Shen 李申 und Wang Benling 王本靈 2015: *Hanyu da cidian yanjiu*《汉语大词典》研究 *(A study on Hanyu da cidian)*. Beijing 北京: Shangwu yinshuguan 商務印書館 (The Commercial Press), S. 96–172.

80 Siehe z. B. Hu Shaowen 胡绍文 2002: „The Shortages of Hanyu Da Cidian (汉语大词典) From the View of Yi Jian Zhi (夷坚志) – 从《夷坚志》看《汉语大词典》的若干阙失". In: *Research In Ancient Chinese Language* 古汉语研究 4, S. 87–89; Chai Hongmai 柴红梅 2005: „To Remedy Some Flaws of Hanyu da cidian (汉语大词典) – On the Basis of Entry C in Xiandai Hanyu cidian (现代汉语词典)《汉语大词典》瑕疵补正 —- 以《现代汉语词典》C 字条为例". In: *Research In Ancient Chinese Language* 古汉语研究 3.

81 Klöter 2013, S. 887.

82 Vgl. *HYDCD*, Bd. 1, S. 8.

83 Vgl. *HYDCD*, Bd. 1, S. 9.

84 Vgl. *HYDCD*, Bd. 1, S. 8; Siehe auch Hargett 1990, S. 139.

85 Für einen Überblick und genaueren Vergleich siehe Yang Lin 杨琳 2011: *"Hanyu da cidian"guangpan ban yu zhizhiban de qubie*《汉语大词典》光盘版与纸质版的区别 *(Unterschiede zwischen der CD-Rom und der Papierausgabe des Hanyu da cidian)*. URL: http://www.guoxue.com/?p=4453 (besucht am 22. 07. 2018), *passim*. Der Autor rät von einer leichtgläubigen Verwendung zumindest der CD-Rom Versionen zu wissenschaftlichen Zwecken ab („*bu neng qing xin guangpan ban* 不能轻信光盘版"), wohingegen die Online-Ausgabe 2.0 (汉语大词典》网络版 2.0) gelobt wird.

falls Kurz- und Langzeichen gemischt verwendet werden.[86] Sie enthält insgesamt 365.102 Wort- und Zeicheneinträgen (davon 22.327 Zeicheneinträge mit insgesamt 16.361 graphisch unterschiedlichen Schriftzeichen-*types*).

Diese Version entspricht keiner der genannten CD-Rom Versionen: Die 1997 erschienene **Version 1.0** kommt mit nur 18.000 Zeichen- und 336.000 Worteinträgen nicht in Frage, auch scheinen in dieser Ausgabe etliche Belegstellen zu fehlen, was bei der verwendeten Ausgabe kaum der Fall ist.[87] Die zuerst 2003 veröffentlichte **Version 2.0** verwendet zwar, wie die vorliegende Version, eine Mischung von Lang- und Kurzzeichen, es wurden jedoch Einträge hinzugefügt, die in der gedruckten Ausgabe fehlen, z. B. die Zeicheneinträge zu *qiao* / *jiao* 釙, *wang* 亟 und *fa* 泼. Alle drei fehlen in der verwendeten Version. Die aus Kompatibilitätsgründen schwieriger zu betreibende **Version 3.0** aus dem Jahr 2007 kommt mit 336.706 Worteinträgen ebenfalls nicht infrage.[88]

Auch wenn die tatsächliche Provenienz der verwendeten Daten schwer feststellbar ist, entspricht sie mit hoher Wahrscheinlichkeit im Wesentlichen der Online-Ausgabe 2.0 des *HYDCD*.[89] Die Tatsache, dass die neueste zitierte Primärquelle aus dem Jahr 1992 stammt,[90] lässt zudem darauf schließen, dass keine rezenteren Lexikalisierungen hinzugekommen sind bzw. keine Einträge vorhanden sind, die in der gedruckten Version, deren 12. und letzter Inhaltsband 1993 erschien, fehlen.

5.4.1 Qualitätssicherung: Abgleich mit der gedruckten Ausgabe

Um die Verlässlichkeit der verwendeten Daten sicherzustellen und das Ausmaß von Abweichungen in angemessenem Umfang zu prüfen, wird ein vereinfachtes Stichprobenverfahren angewandt.[91]

Dabei soll die Übereinstimmung der Einträge mit der gedruckten Ausgabe als Merkmal untersucht werden, Merkmalsträger für die Stichprobe sind die Wort- und Zeicheneinträge *mit Belegstellen*, mit einer Grundgesamtheit von 323.321 Einträgen.[92] Wesentlich für das hier verfolgte Unterfangen sind die erste bzw. älteste darin angegebene Primärquelle. Wie bei solchen Verfahren üblich soll eine Sicherheit von 95 % erreicht werden, woraus sich ein Wert für das Quantil der Standardnormalverteilung z von 1.96 ergibt.[93] Bei einer ersten Sichtung von 36 Einträgen weisen nur zwei Einträge (ca. 6 %) minimale Abweichungen auf, der Anteilswert P liegt

86 *DHYDCD*.
87 Siehe S. 117.
88 Zu den Angaben siehe den Vergleich von Yang Lin 杨琳 2011, *passim*.
89 Luo Zhufeng 羅竹風, Hrsg. 2005: *Hanyu da cidian* 漢語大詞典 *UTF-8 (Großes Wörterbuch der chinesischen Sprache, Unicode-Version)*. Shanghai 上海. URL: `http://bbs.gxsd.com.cn/forum.php?mod=viewthread&tid=498015` (besucht am 13. 01. 2013).
90 z. B. *Renmin ribao* 人民日报 1992. Siehe *HYDCD*, Bd. 12, S. 421, 體壇; bzw. *DHYDCD*, 體壇.
91 Die Methodik ist den Prognosen nach Bundes- oder Landtagswahlen entlehnt, die auf Befragungen der Wähler:innen basieren. Der Vorteil einer solchen Herangehensweise gegenüber einer Vollerhebung liegt nicht nur im deutlich reduzierten Aufwand, sondern auch in der Vermeidung von Erhebungsfehlern, die bei letzterer „bedingt durch den hohen Aufwand häufig" auftreten. Siehe Göran Kauermann und Helmut Küchenhoff 2010: *Stichproben – Methoden und praktische Umsetzung mit R*. Berlin & Heidelberg: Springer. DOI: 10.1007/978-3-642-12318-4, S. 1, S. 6.
92 Der Begriff der Grundgesamtheit beschreibt die „Menge aller Individuen oder Objekte, über die eine Aussage getroffen werden soll". Merkmalsträger sind „die Einheiten oder Objekte, an denen Untersuchungen, Messungen oder Beobachtungen vorgenommen werden"; Merkmale „die Eigenschaften […], die untersucht, […] werden sollen." ebd., S. 5, vgl. auch S. 29.
93 Siehe ebd., S. 28.

daher bei 10 % oder weniger.[94] Strebt man eine Genauigkeit e von wenigstens 0,05 an, ergibt sich folgende Berechnung der benötigten Stichprobengröße:[95]

$$n \geq \frac{P(1-P)}{e^2/z^2 + P(1-P)/N}$$

$$\frac{0,1 \times 0,9}{0,05^2/1,96^2 + 0,1 \times 0,9/323321} \approx 138,24$$

Eine Stichprobe von 139 Einträgen reicht also aus, um mit 95 %iger Sicherheit den Anteil der von der gedruckten Ausgabe abweichenden Einträge mit einer Genauigkeit von 5 % bestimmen zu können.[96] Die zur Durchführung verglichenen Einträge werden zufällig ausgewählt.[97] Der manuelle Abgleich wird anhand der folgenden Einträge durchgeführt (hier sortiert nach ihrem Vorkommen im *DHYDCD*):

> 一箭道、不期、不論、乾精、亂坟崗、伐棠、休光、佚怠、保佐、偏惰、偏比、傾顛、僞戾、六神無主、出入人罪、剛鍱、剪剪、劍鐔、南沃沮、吹火、周盈、噦吐、圓紗、外轉、大朝覲、奭發、好勇、季王、宸注、尊便、小半、尼軻、層、幔城、庵、弓勢、彝典、怒譴、思如湧泉、怨憎、恃怙、悲筑、慕容、掩滅、捷毒、收煞、收身、收過、放大炮、斃倫、舅、昌羊、春風面、昭飾、智將、曲致、本情、杜口裹足、杯水候、柮、桃葉女、械索、梳雲、椎擊、榆莢錢、機勇、機暇、款、此以、殷劉、每事問、毛胚、氣苦、混一、清狂、溪蓀、無人問津、煙册、牆頭馬上、王鳩、異謀、瞠惑、矚、積秀、突地吼、第恐、笳管、絶腸、綱梢、老是、肉裏錢、胡顏、膽悸、茁茂、茹古涵今、荒率、虞曹、蟻垤、蠢然、要實、覆地翻天、規約、親妮、訓言、講切、警俊、資、賠餉、賡歌、跳鱗、踏蹬、輪轉椅、辟摽、過動、選序、遺寇、配匹、酸溜溜、金衣丹、鉅萬、閔隔、陵誑、雅奏、集裝箱、難分難捨、雷火、靈湖、青粱、鞴靼、食官、餔糜、餘印、馬隊、驍銳、鶴瘦、鶻、鷙膺、麗象、龍馭。

Ergebnisse der Stichprobenanalyse

Insgesamt weisen 8 Einträge aus der Stichprobe (5,76 %) nennenswerte inhaltliche Abweichungen auf, meist in Form von fehlenden Belegstellen. Bedeutsam für die chronologische Einordnung ist dabei lediglich ein einziges fehlendes Zitat im Eintrag 資 ² (*zi*),[98] da hierdurch die älteste Belegstelle abweicht (siehe unten). Die relevante Abweichung liegt also unter einem Prozent. Die verwendete Ausgabe kann damit als hinreichend verlässlich angesehen werden.

Tabelle 5.1 gibt einen Überblick über das Ergebnis der Stichprobenanalyse. In der Spalte *Hochrechnung* ist dabei die theoretische, hochgerechnete Gesamtanzahl der Einträge angegeben, auf die das jeweilige Merkmal zutrifft, wenn man von 323.321 relevanten Einträgen ausgeht.
Im Verlauf der Analyse können im Detail folgende formale bzw. typographische und inhaltliche Unterschiede zwischen *HYDCD* und *DHYDCD* festgestellt werden:

94 Die Bestimmung des Anteilswerts mittels einer Pilotstichprobe ist in der Statistik nicht unüblich. Siehe dazu ebd., S. 39.
95 Ebd., S. 41.
96 Die Repräsentativität „typischer" Einträge kann hier außer Acht gelassen werden, da keine Anzeichen vorliegen, dass bestimmte Typen von Einträgen sich stärker als andere zwischen den verglichenen Ausgaben unterscheiden. Vgl dazu. ebd., S. 8f.
97 Für die Auswahl wurden Einträge aus der *SQL*-Datenbank (siehe dazu Kapitel 5.5, ab S. 120) selektiert und zufällig „sortiert" (`[…] order by rand() limit 139`).
98 Zur Nummerierung von Einträgen zu Zeichen mit unterschiedlichen Aussprachen siehe auch den Abschnitt zur Struktur (5.5.1, ab S. 121).

Tabelle 5.1 Qualität der digitalen Ausgabe – Ergebnisse der Stichprobenanalyse

Merkmal	Einträge	Anteil (von 139)	Hochrechnung
Älteste Belegstelle weicht ab	1	0,72 %	2.326
Fehlende Belegstellen	7	5,03 %	16.282
Inhaltliche Abweichung	8	5,76 %	18.608

— 1. **Belegstellen**. Wie bereits angedeutet fehlen in der digitalen Ausgabe einzelne Quellenzitate. Dies trifft gleichermaßen auf Zeichen-[99] und Worteinträge zu.[100] Auch inhaltliche Unterschiede in zitierten Belegstellen können festgestellt werden.[101]

Lediglich in einem einzigen untersuchten Eintrag weicht – wie bereits erwähnt – die älteste Quellenangabe ab: Im Eintrag von *zì* 資 2 gibt die digitale Ausgabe lediglich eine Belegstelle aus dem Gedicht *Huashan nü* 華山女 von Han Yu 韩愈 (768–824) an,[102] während die gedruckte Ausgabe noch zwei deutlich ältere Belegstellen aus dem *Han shu* 漢書 und dem *Shiji* 史記 enthält.[103] Allerdings finden sich identische, ältere Belegstellen im Eintrag *zì* 資 1,[104] so dass dieser konkrete Fall [zufälligerweise] keine Auswirkungen auf die erzeugten Daten hätte, da graphisch gleiche Zeichen in *Plain Text*-Daten nicht unterschieden werden können.

— 2. **Typographische Markierungen**. In der gedruckten Ausgabe sind Personen-, Dynastie- und Ortsnamen unterstrichen. Durch Unterbrechungen können Dynastie- und Personennamen hier klar unterschieden werden, z. B. 宋 穆休.[105] Da *Plain-Text* keine Formatierungen enthalten kann, fehlen diese Informationen in der hier verwendeten digitalen Ausgabe vollständig, während sie in der offiziellen CD-Rom Version vorhanden sind. Dadurch wird an einigen Stellen beim Parsen der Daten die Unterscheidung, ob eine Quellenangabe im Format *DynastieNachnameVorname*, oder lediglich *NachnameVorname* vorliegt, erschwert.[106]

— 3. **Zitierweise**. Wird mehrmals in Folge dieselbe Quelle zitiert, wird das Zitat in der gedruckten Ausgabe bei den Folgeangaben mit *you* 又 („erneut“) eingeführt, etwa im Eintrag zu *yidai* 佚怠, in welchem zwei Stellen aus dem *Yanzi Chunqiu* 晏子春秋 zitiert werden.[107] Die digitale Ausgabe wiederholt die vollständige Quellenangabe, was die Extraktion dieser Daten erleichtert.

— 4. **Gruppierung und Nummerierung von Untereinträgen**. In der gedruckten Ausgabe werden unterschiedliche Wortbedeutungen, sofern zutreffend, nach syntaktischen Kategorien, z. B. *lianci* 連詞 (Konjunktion), gruppiert. Die Kategorien werden dabei mit eingekreister Nummerierung (❶, ❷, ❸…) markiert, die Unter-untereinträge mit einfachen Klammern (3).

99 z. B. zu *rui* 枘 gibt die gedruckte Ausgabe zur zweiten Bedeutung ein *Tang*-zeitliches Zitat an, das in der digitalen Ausgabe fehlt. *HYDCD*, Bd. 4, S. 854; *DHYDCD*, 枘. Ebenfalls nur in der Papierversion enthält der Eintrag *fu* 甫 ein *Qing*-zeitliches Zitat. *HYDCD*, Bd. 6, S. 1599; *DHYDCD*, 甫.

100 Im digitalen Eintrag zu *gengge* 賡歌 fehlen zwei Quellenzitate. *DHYDCD*, 賡歌; *HYDCD*, Bd. 10, S. 275.

101 In beiden Ausgaben wird im Eintrag zu *longyu* 龍馭 eine Stelle aus dem Gedicht *Yu dong jun* 喻東軍 von Wei Zhuang 韋莊 (*ca.* 836–910) in jeweils unterschiedlicher Fassung wiedergegeben. „四年龍馭守峨眉，到此躊躇不能去“ lautet in der digitalen Ausgabe „四年龍馭守峨嵋，鐵馬西來步步遲“ – vermutlich handelt es sich um eine Korrektur in der neueren Ausgabe. *HYDCD*, Bd. 12, S. 1481; *DHYDCD*, 龍馭.

102 *DHYDCD*, 資 2.

103 *HYDCD*, Bd. 10, S. 200.

104 *DHYDCD*, 資 1; *HYDCD*, Bd. 10, S. 199.

105 Siehe *HYDCD*, Bd. 2, S. 1224.

106 Auf diese Problematik wird in Abschnitt 5.5.2, S. 130, genauer eingegangen.

107 Siehe *HYDCD*, Bd. 1, S. 1244.

Die digitale Ausgabe nimmt nur eine einstufige Nummerierung vor, die dadurch häufig abweicht.[108] Bei manchen Einträgen entfällt in der gedruckten Ausgabe die Nummerierung und die zweite bzw. weitere Bedeutungen werden mit *yi zhi* 亦指…(„deutet auch auf…“) oder *yinshen wei* 引申为 (etwa: „eine erweiterte Bedeutung ist…“) eingeleitet.[109] In der digitalen Ausgabe wird in allen Fällen konsequent mit „1., 2., 3.…“ nummeriert, so dass die Abschnitte mit den unterschiedlichen Bedeutungen einfach zu segmentieren sind.

— 5. **Strichzahl des zweiten Zeichens**. In der gedruckten Ausgabe wird diese bei jeder Erhöhung durch eine hochgestellte Zahl ausgewiesen, also z. B. [12]【伐棠】,[110] wobei 12 die Anzahl der Striche von *tang* 棠 angibt. In der digitalen Ausgabe fehlen solche Angaben vollständig.

— 6. **Querverweise** auf andere Worteinträge zeigen in der gedruckten Ausgabe stets auf den relevanten Untereintrag.[111] Auch kleine inhaltliche Unterschiede in den Querverweisen kommen vor – dabei scheint aber nicht eine der beiden Ausgaben genauer zu sein, sondern schlicht beide minimal unterschiedlich.[112]

— 7. In der gedruckten Ausgabe wird bei nicht eindeutiger **Aussprache** des zweiten (dritten, usw.) Zeichens eines Worteintrages die Lesung dieser Zeichen explizit in *Hanyu Pinyin* 漢語拼音 angegeben.[113] In der digitalen Ausgabe fehlen solche Angaben leider.

— 8. Im *DHYDCD* wird ***Zhuyin Fuhao*** 注音符號 („*Bopomofo*“) zur Angabe der Aussprache zusätzlich angebeben, in der gedruckten Ausgabe lediglich das festlandchinesische *Hanyu Pinyin*.

Durch die einfachere, konsequentere Struktur der Untereinträge und die stets vollständigen Quellenangaben ist der Text der digitalen Ausgabe insgesamt leichter maschinenlesbar und damit sogar besser für die Analyse geeignet als der ursprüngliche Text. Für eine Extraktion der Wortklassen aus den Kategorien (siehe 3.) – als Möglichkeit zur Gewinnung von Daten zum *Part-of-Speech Tagging* – wären auch die Angaben in der gedruckten Ausgabe zu unvollständig und unsystematisch.

Zur Veranschaulichung sei an dieser Stelle ein Beispiel aus der gedruckten Ausgabe wiedergegeben:[114]

108 Vgl. z. B. die Einträge zu *bulun* 不論 *HYDCD*, Bd. 1, S. 468; sowie *ceng* 層 *HYDCD*, Bd. 4, S. 60.

109 Siehe z. B. in den Einträgen zu *jianyin* 劍鐔 und *kuan* 款. *HYDCD*, Bd. 2, S. 753, Bd. 6, S. 1444.

110 *HYDCD*, Bd. 1, S. 1190, *fatang* 伐棠.

111 Siehe z. B. im Eintrag *yujiaqian* 榆莢錢: „参见“榆莢❷”“ („siehe *yujia* ❷“) – in der digitalen Ausgabe wird lediglich auf 榆莢 verwiesen. Vgl. *HYDCD*, Bd. 4, S. 1188.

112 Nur die digitale Ausgabe weist im Eintrag zu *ceng* 層 darauf hin, dass dieses Zeichen mit dem homophonen *ceng* 增 austauschbar verwendet werden kann. Siehe *DHYDCD*, 層; *HYDCD*, Bd. 4, S. 60. Umgekehrt verweisen gedruckte wie digitale Ausgabe im zweiten Eintrag zu 增 auf 層. Siehe *DHYDCD*, 增; *HYDCD*, Bd. 2, S. 1222. Während die gedruckte Ausgabe im Eintrag zu *zi* 資 auf das „gleiche“ Zeichen 恣 verweist, fehlt diese Information in der digitalen Ausgabe. Vgl. *HYDCD*, Bd. 19, S. 200; *DHYDCD*, 資.

113 Siehe z. B. im Eintrag zu *zūnbiàn* 尊便 die Angabe „– biàn“ vor der Angabe der Bedeutung. *HYDCD*, Bd. 2, S. 1283. 便 kann, je nach Kontext bzw. Bedeutung auch *pián* oder *biān* gelesen werden.

114 *HYDCD*, Bd. 7, S. 986.

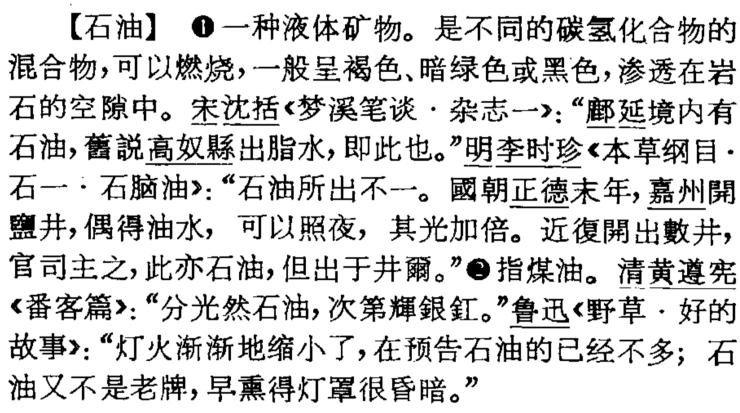

【石油】 ❶一种液体矿物。是不同的碳氢化合物的混合物，可以燃烧，一般呈褐色、暗绿色或黑色，渗透在岩石的空隙中。宋沈括《梦溪笔谈·杂志一》："鄜延境内有石油，舊説高奴縣出脂水，即此也。"明李时珍《本草纲目·石一·石脑油》："石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。近復開出數井，官司主之，此亦石油，但出于井爾。"❷指煤油。清黄遵宪《番客篇》："分光然石油，次第輝銀釭。"鲁迅《野草·好的故事》："灯火渐渐地缩小了，在预告石油的已经不多；石油又不是老牌，早熏得灯罩很昏暗。"

Abbildung 5.1 Eintrag *shiyou* 石油 („Steinöl", Erdöl) in der Originalausgabe des *HYDCD*.

In der digitalen Version ist der gleiche Eintrag enthalten – wobei ein Teil der oben beschriebenen typographischen Vereinfachungen sichtbar wird:

> 【石油】1. 一种液体矿物。是不同的碳氢化合物的混合物，可以燃烧，一般呈褐色、暗绿色或黑色，渗透在岩石的空隙中。宋沈括《梦溪笔谈·杂志一》："鄜延境内有石油，舊説高奴縣出脂水，即此也。"明李时珍《本草纲目·石一·石脑油》："石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。近復開出數井，官司主之，此亦石油，但出于井爾。"2. 指煤油。清黄遵宪《番客篇》："分光然石油，次第輝銀釭。"鲁迅《野草·好的故事》："灯火渐渐地缩小了，在预告石油的已经不多；石油又不是老牌，早熏得灯罩很昏暗。"[115]

Eine offensichtliche Schwäche der hier durchgeführten Stichprobe ist ihre Auswahl aus den Einträgen des *DHYDCD*, denn darin fehlende Zeichen- und Worteinträge bleiben unbemerkt. Das trifft z. B. auf die Zeichen *bing* 丙 und *mei* 美 und die zugehörigen Worteinträge zu.[116] Ob noch weitere Einträge fehlen, die in der gedruckten Ausgabe vorhanden sind, ist nur mit unverhältnismäßigem Aufwand feststellbar.[117] Auf die inhaltliche Qualität der verbleibenden – und davon abgesehen augenscheinlich auch vollständigen – Daten hat dies jedoch keinen Einfluss.

5.5 Erzeugung einer diachronen Lexemdatenbank

Um die Lexikalisierungsdaten aus dem *DHYDCD* nutzbar zu machen, wird dieses in eine *SQL*-Datenbank umgeformt und anschließend mit weiteren Informationen angereichert. Die Strukturierung als relationale Datenbank ermöglicht es, die Daten bei minimaler Redundanz und guter Nachvollziehbarkeit zu erweitern und später zielgenau effizient abzufragen. Dafür werden die Quellen der Belegstellen extrahiert und – soweit ermittelbar – zur chronologischen Einordnung der Lexeme der Entstehungszeitraum bzw. -zeitpunkt des ältesten

115 *DHYDCD*, *shiyou* 石油.

116 Siehe *HYDCD*, Bd. 1, 丙, S. 509–510 , Bd. 9, 美, S. 158–164. 13 Worteinträge zu *bing*, sowie 137, die mit *mei* beginnen, fehlen ebenfalls. Vgl. *DHYDCD*.

117 Das Fehlen der Einträge zu *mei* 美 und *bing* 丙 folgt keiner erkennbaren Logik. Die vorherigen und nachfolgenden Einträge zu *qie/ju/zu/cu* 且 und *qiu* 丘 bzw. *da* 夆 und *qiang* 羌 sind in beiden Ausgaben vorhanden. Eine Möglichkeit, Kandidaten für im *DHYDCD* fehlende Einträge systematisch aufzuspüren ist es, Zeichen zu ermitteln, die zwar in Worteinträgen verwendet werden, aber keinen eigenen Zeicheneintrag haben. Das trifft auf insgesamt 216 Zeichen zu, unter denen sich aber etliche Varianten befinden, z. B. *kuai* 凷 (für 塊) und *chu* 出 (für 出).

zitierten Texts verwendet. Zur strukturierten Extraktion der Inhalte kommen dabei überwiegend **Reguläre Ausdrücke** (*Regular Expressions*, kurz *RegEx*) zum Einsatz. Sie sind ein in vielen Programmiersprachen verbreitetes syntaktisches Konzept zur Beschreibung von Mustern und werden eingesetzt, um bestimmte Informationen aus Texten zu extrahieren, zu suchen oder zu ersetzen.[118] Aufbau und Erstellung dieser historischen Lexemdatenbank sind im Folgenden dokumentiert.

1. Segmentierung der Rohdaten in Zeichen- und Worteinträge.
2. Erkennen der Belegstellen in diesen Einträgen und Interpretation der zugehörigen Metadaten, d. h. Titel, Autor und Entstehungszeit des zitierten Textes.
3. Die oft unvollständigen oder ungenauen Metadaten werden mit externen Datenquellen verdichtet.

5.5.1 Datenstruktur

In der vorliegenden digitalen Ausgabe wird jeder Zeicheneintrag (*dan zi tiaomu* 單字條目) mit einem Asterisk eingeleitet (z. B. „* 漢“, so dass die Einträge mithilfe des regulären Ausdrucks `(\*\p{IsHan})` segmentiert werden können. Die so getrennten Haupteinträge lassen sich in zwei Arten von Untereinträgen unterteilen: mehrsilibige Lexemeinträge, sowie die unterschiedlichen Lesungen der Zeicheneinträge bei *duoyinzi* 多音字. Mehrsilbige Lexemeinträge sind stets an „gefüllten quadratischen Klammern“ 【 】 (*shixin fangtou kuohao* 實心方頭括號) erkennbar.

> 【且 2 末】汉代西域国名。《汉书 · 西域传上 · 且末国》: “且末國，王治且末城，去長安六千八百二十里。”地在今新疆且末县。[119]
> 【且並】并且。清和邦额《夜谭随录 · 诡黄》: “驚惶間已失鞋，且並脱去一襪。”[120]

Die Untereinträge für *duoyinzi* lassen sich an den eckigen Klammern ［ ］erkennen, in welchen die Aussprache angegeben wird, z. B.

> 且 2［jū ㄐㄩ］［《廣韻》子魚切，平魚，精。］1. 多貌。《诗 · 大雅 · 韩奕》: “籩豆有且，侯氏燕胥。” …[121]

Der reguläre Ausdruck `(【[^】]+】|\p{IsHan} [0-9]［.+］)` [122] beschreibt die somit möglichen Markierungen von Untereinträgen.

Da für spätere Verarbeitungsschritte etliche Besonderheiten berücksichtigt werden müssen, werden die Daten mit den obigen regulären Ausdrücken in *Python* segmentiert und direkt in die benötigte Datenbankstruktur geschrieben.[123] In diesem Rahmen finden zusätzliche Verarbeitungsschritte statt:

---

118 *RegEx* finden häufig auch im kommerziellen Kontext Verwendung, z. B. für Formularvalidierungen. Soll zum Beispiel eine Kundin in einem Webshop die IBAN-Nummer einer deutschen Bankverbindung angeben, kann der Anbieter prüfen, ob die Eingabe dem Muster `^DE(?:[ ]?[0-9]){20}$` entspricht. Eine Zeichenkette, die mit „DE“ beginnt, gefolgt von zwanzig Ziffern von 0 bis 9, zwischen denen einzelne Leerzeichen zugelassen sind. Die Existenz oder gar Deckung des Kontos lässt sich damit sicherlich nicht absichern, wohl aber, ob die Kundin passenden Inhalt in das vorgesehene Feld eingibt.

119 *DHYDCD*, 且 2 末. Farbliche Markierungen gemäß Übereinstimmung mit dem Muster der verwendeten regulären Ausdrücke.

120 *DHYDCD*, 且並.

121 *DHYDCD*, 且 2.

122 【, gefolgt von allen Zeichen, die nicht „】“ sind, bis 】 *oder* alternativ: Ein einzelnes *Hanzi* 漢字, eine Ziffer, gefolgt von mind. einem beliebigen Zeichen in eckigen Klammern ［ ］.

123 Ein *Python*-Script verarbeitet die 365.102 Einträge in etwa 90 Sekunden (knapp 4.000 Einträge pro Sekunde).

1. Alle Zeilenumbrüche und Leerzeichen[124] werden entfernt, da sie innerhalb von Einträgen lediglich das Ende einer Zeile von max. 50 Zeichen bzw. etwa 44 chinesischen Schriftzeichen markieren und für die Inhaltsanalyse hinderlich sind.
2. Extraktion der Aussprache- und Reiminformationen aus den Zeicheneinträgen, ebenfalls mit regulären Ausdrücken.
   a) *Pinyin*-Aussprache: `aa\u0101\u00E0\u00E1\u01CEe\u0113\u00E9\u011B\u00E8i\u012B\u00ED\u01D0\u00ECo\u014D\u00F3\u01D2\u00F2uū\u016B\u00FA\u01D4\u00F9v\u00FC\u01D6\u01D8\u01DA\u01DCbcdfggkhjlmmm̄m̀ńňǹpqrstwxyz`.[125]
   b) *Zhuyin*-Aussprache: `([ㄅㄆㄇㄈㄉㄊㄋㄌㄍㄎㄏㄐㄑㄒㄓㄔㄕㄖㄗㄘㄙㄚㄛㄜㄝㄞㄟㄠㄡㄢㄣㄤㄥㄦㄧㄨㄩ市\u02C7\u02CA\u02CB\u02D9]+)`.[126]
   c) Angaben über Reime wie „`［《廣韻》子魚切，平魚，精。］`“[127] entsprechen stets dem regulären Ausdruck `(［《.+》.+］)`.

Segmentierung des *Plain Text*-Wörterbuchs in Datenbanktabellen

Der Aufbau der Datenbank spiegelt die oben beschriebene Struktur des *HYDCD* mit Zeichen- und Worteinträgen wider, die zunächst in zwei entsprechende Tabellen geschrieben werden:

— 1. `hydcd_entries` – Ablage der Zeicheneinträge (Haupteinträge) mit ihrem vollständigen Inhalt (inkl. aller untergeordneten Worteinträge).

- `id` – Eindeutige ID des Zeicheneintrags, nach Radikal und Zusatzstrichzahl sortiert. Der erste Eintrag, *yī* 一 erhält die ID 1, der letzte, *guī* 龜, die ID 16.358.
- `char` – Das Schriftzeichen, auf das sich der Inhalt des Eintrags bezieht.
- `entry` – Der vollständige Inhalt des Eintrags, inklusive aller Untereinträge (Wort- und Zeicheneinträge).

— 2. `hydcd_words` – Eine Datenbankzeile pro Zeichen- *und* Lexemeintrag, d. h. unterschiedliche Lesungen bei Zeicheneinträgen (*duoyinzi* 多音字, Heteronyme) werden gesondert gelistet. So kommen z. B. auf einen Eintrag für *qiě* 且 (ID 30 in der Tabelle `hydcd_entries`) die vier Einträge *qiě* 且, *jū* 且, *zǔ* 且 und *cú* 且, sowie 35 polysyllabische Lexeme, die mit 且 beginnen.

- `id` – Eindeutige ID des Worteintrages.
- `id_internal` – Interner Zähler für Untereinträge des Zeicheneintrages.
- `char_id` – ID des übergeordeten Zeicheneintrags in der Tabelle `hydcd_entries`.
- `word` – gibt das Lemma im ursprünglichen Format in Klammern 【】 mit dem Verweis auf die Lesung des führenden Zeichens an, also z. B. 【且 2 末】, wobei die „2“ signalisiert, dass 且 hier *jū* gelesen wird.
- `cleanword` – das Schlagwort ohne Zusatzinformationen, z. B. 且末.
- `nakedword` – das Schlagwort ohne eventuell vorhandene Interpunktionszeichen.

124 D. h. `\n, \t, \f, \r, \v` und das Leerzeichen.
125 Erweiterung des regulären Ausdrucks aus Thomas Roten 2017: *zhon - Constants used in Chinese text processing*. URL: `https://github.com/tsroten/zhon` (besucht am 10. 11. 2017), `pinyin.py`. Zusätzlich zu den im *Pinyin*-Standard vorgesehenen Zeichen sind hier die im *DHYDCD* verwendeten Sonderformen berücksichtigt.
126 Alle verfügbaren *Zhuyin*-Zeichen und die Unicode-Repräsentationen der verwendeten Diakritika für Töne. Der erste Ton wird in *Zhuyin* nicht graphisch wiedergegeben. Basierend auf ebd., `zhuyin.py`.
127 *DHYDCD*, 且 2.

- `zhuyin`, `pinyin` – gibt die Aussprache jeweils in *Zhuyin* 注音 (z. B. ㄐㄩㄇㄛˋ) und *Hanyu Pinyin* 漢語拼音 (z. B. *jūmò*) an.[128]
- `rhyme` – gibt, sofern vorhanden, für Zeicheneinträge Reim-, Ton und *Fanqie* 反切 Informationen oder Schriftzeichen mit gleicher Lesung, sowie die Quelle der jeweiligen Angabe an, z. B. ［《廣韻》子魚切，平魚，精。］.[129]
- `entry` – Der Inhalt des Worteintrags.
- `entrytype` – *C* für einzelne Zeichen (monosyllabische Wörter), *W* für Worteinträge (polysyllabische Wörter).

Tabellen der diachronen Lexemdatenbank

Aus den so strukturierten Daten werden nun die Lexikalisierungsdaten der Worteinträge extrahiert. Im Beispieleintrag zu *shiyou* 石油[130] werden zu zwei Bedeutungen (unten markiert in rot) Erklärungen gegeben (hier ausgegraut). Zu jeder Bedeutung werden zudem entsprechende Belege aus der Literatur (schwarz) in Anführungsstrichen “” zitiert. Diese werden stets mit einer vereinfachten bibliographischen Angabe eingeführt, bestenfalls im Format DynastieAutor 《Werk · Kapitel》 und sind in der Regel chronologisch sortiert. Da Unterstreichungen im *DHYDCD* fehlen, sind sie zu Illustrationszwecken aus der Originalausgabe übernommen:

> 【石油】
> 1. 一种液体矿物。是不同的碳氢化合物的混合物，可以燃烧，一般呈褐色、暗绿色或黑色，渗透在岩石的空隙中。 宋 沈括《梦溪笔谈 · 杂志一》:“鄜延境内有石油，舊説高奴縣出脂水，即此也。” 明 李时珍《本草纲目 · 石一 · 石脑油》:“石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。近復開出數井，官司主之，此亦石油，但出于井爾。”
> 2. 指煤油。 清 黄遵宪《番客篇》:“分光然石油，次第輝銀釭。” 鲁迅《野草 · 好的故事》:“灯火渐渐地缩小了，在预告石油的已经不多；石油又不是老牌，早熏得灯罩很昏暗。”[131]

Der Begriff *shiyou* 石油 mit der Bedeutung „eine Art flüssiges Mineral“ (*„yi zhong yeti kuangwu* 一种液体矿物“) ist also spätestens in der Song 宋-Zeit (960–1279) belegt und im *Meng xi bi tan* 夢溪筆談 („Pinselunterhaltung am Traumbach“[132]) von Shen Kuo 沈括 (1031–1095) zu verorten. Im Optimalfall handelt es sich bei dieser Angabe um den *Locus classicus*, was aber nicht letztgültig geklärt werden kann. Unabhängig davon, ob Shen Kuo den Begriff wirklich geprägt oder sogar erfunden hat, ist gesichert, dass *spätestens* zu dieser Zeit der Begriff bereits verwendet wurde.

Auch aus dem Ming 明-zeitlichen (1368–1644) *Ben cao gangmu* 本草綱目 („*Materia Medica, Arranged according to Drug Descriptions and Technical Aspects*“[133]) ist ein Zitat angegeben. Die zweite Bedeutung, „Lampenöl“ (Petroleum, *meiyou* 煤油), ist hier erst für die Qing 清-Zeit (1644–1911)

128 Die Aussprachеinformationen für mehrsilbige Einträge werden aus den Angaben in den Zeicheneinträgen des *DHYDCD* zusammengesetzt. Alternative Zeichenlesungen können dabei leider nur für das jeweils erste Zeichen eines mehrsilbigen Ausdrucks berücksichtigt werden, da für die nachfolgenden Zeichen im Gegensatz zur gedruckten Ausgabe kein Hinweis auf die Aussprache gegeben wird (siehe auch Abschnitt 5.4.1, S. 117). Für die „hinteren“ Zeichen wird hier daher immer die erste Lesung angenommen.

129 *Fanqie* ist eine traditionelle Methode zur phonetischen Analyse – die Lesung eines Zeichens wird zu diesem Zweck mit zwei weiteren Zeichen repräsentiert, von denen das erste Zeichen den An-, das zweite den Auslaut angibt. Im Beispiel werden also *zi* 子 und *yu* 魚 zu *ju* „geschnitten“ (*qie* 切). Die Kompilator:innen haben hier zumeist Informationen aus den song-zeitlichen Reimwörterbüchern *Guangyun* 廣韻 bzw. *Jiyun* 集韻 angegeben. Vgl. *HYDCD*, *passim*.

130 Siehe auch S. 120.

131 *DHYDCD*, 石油. Unterstreichung und farbliche Hervorhebungen durch den Verfasser.

132 Shen Kuo 沈括 1997 [1088]: *Pinselunterhaltungen am Traumbach. übs. von Konrad Herrmann.* München: Diederichs.

133 Paul Ulrich Unschuld 1986: *Medicine in China: A History of Pharmaceutics*. Comparative Studies of Health. Berkeley & Los Angeles: University of California Press, S. 145.

nachgewiesen, im *Fanke pian* 番客篇 („The Foreign Guest“[134]) von Huang Zunxian 黃遵憲 (1848–1905)[135], sowie später in Lu Xuns 魯迅 (1881–1936)[136] Prosagedichtsammlung *Ye Cao* 野草 („Wildes Gras“). Für die vorgesehene Anwendung der Datenbank ist die früheste Angabe zur Lexikalisierung über das *Meng xi bi tan* am wichtigsten.

Das Beispiel zeigt, dass chronologische Informationen im *HYDCD* sich zumeist (wenn überhaupt) auf die Angabe der Dynastie beschränken und damit vage bzw. implizit sind. Eine Liste mit Angaben zu den zitierten Werken bzw. der verwendeten Ausgaben fehlt im *HYDCD* zudem völlig, so dass solche Daten aus externen Quellen ergänzt werden müssen.

Zunächst werden die vorhandenen Lexikalisierungsdaten in drei weitere Datenbanktabellen strukturiert, die eine Verknüpfung von Lexem, Belegstellen, sowie zitierten Werken und damit eine (implizite) chronologische Einordnung der Lexeme ohne urheberrechtlich geschützte Inhalte enthalten.

Die Tabelle `the_words` soll alle Schlagworte enthalten, zu denen ein Textbeispiel angegeben ist,[137] sowie den Verweis auf die älteste angegebene Belegstelle. In `the_books` sind die verfügbaren Metadaten zu allen unterscheidbaren, im *DHYDCD* zitierten Texten enthalten und `the_citations` ermöglicht *n : m*-Verknüpfungen für *alle* in den Wörterbucheinträgen zitierten Quellen. Letzteres kann für die diachrone Betrachtung der Wortnutzung und die Datenabdeckung bzw. die Erzeugung eines diachronen Arbeitskorpus herangezogen werden.[138] Die wichtigsten Spalten der genannten Datenbanktabellen werden im Folgenden dokumentiert.

— 1. `the_words` – Alle Zeichen- und Worteinträge, die eine Belegstelle mit Quellenangabe aufweisen, mit Verknüpfung zur Quelle des ältesten zitierten Belegs.

- `id` – ID des Eintrags in der Tabelle `hydcd_words`.[139]
- `cleanword` – Das Lexem bzw. Schlagwort.
- `pinyin` – Die Aussprache in *Hanyu Pinyin* 漢語拼音.
- `firstentry` – Die erste, in dem Eintrag zitierte Primärquelle, inklusive möglicher, mit einem · abgetrennter, Kapitelangaben (z. B. „庄子 · 齐物论“).
- `unordered` – Markiert Einträge, bei in denen eine nicht chronologische Reihenfolge der zitierten Quellen vermutet wird (siehe unten).
- `indirectsource` – Markiert Einträge, in denen bei der ersten Bedeutung keine Quellenangabe gefunden wurde.
- `book` – Der Titel der frühesten zitierten Primärquelle, ohne Kapitelangaben.
- `book_id` – ID der zitierten Quelle in der Tabelle `the_books`.
- `earliest_evidence_id` – `book_id` derjenigen Quelle, die den ältesten in Korpusdaten gefundenen Beleg für die Zeichenkombination in `cleanword` enthält – unabhängig von der Angabe im *DHYDCD*. Diese Spalte wird genutzt, falls ältere Belegstellen gefunden werden können, als im *DHYDCD* angegeben sind.

134 Yang Zhiyi 楊治宜 2015: „The Modernity of the Ancient-Style Verse“. In: *Frontiers of Literary Studies in China* 9.4, S. 551–580, S. 554.
135 Raoul David Findeisen 2004: „Literatur im 20. Jahrhundert“. In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 288–395, S. 293.
136 Reinhard Emmerich 2004: „östliche Han bis Tang“. In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 88–186, S. 136.
137 Damit die erzeugten Daten ohne lizenzrechtliche Bedenken veröffentlicht werden können, sind die Einträge selbst nicht enthalten.
138 Siehe Kapitel 5.7, ab S. 138 bzw. 5.6, ab S. 137.
139 Vgl. Abschnitt 5.5.1, S. 122.

— 2. `the_books` – Alle unterscheidbaren im *DHYDCD* aufgeführten Primärquellen der Zitate aus den Einträgen in `the_words`. Unterscheidbar bedeutet dabei, dass sich eines der Kriterien Titel, Autor, Dynastie *oder* Erscheinungsjahr bzw. einer der Werte in den Datenspalten `clearbook`, `author`, `dynasty` oder `startyear` unterscheidet. Nur wenn *alle* diese Angaben identisch sind, wird davon ausgegangen, dass aus derselben Quelle zitiert wird.[140]

Datenmodell: $1:n$ – jede Primärquelle kann in n Worteinträgen die erste bzw. älteste Quellenangabe (*Locus classicus*) sein. Aufgrund der uneinheitlichen Zitierweise im *HYDCD* ist es bei mehrfach zitierten Quellen teilweise unvermeidlich, dass unter einer ID in `the_books` nicht alle Nennungen zusammengeführt werden, die sich *de facto* auf dasselbe Werk beziehen.[141]

- `id` – Eindeutige ID der Primärquelle, Reihenfolge wie im *DHYDCD*.
- `clearbook` – Der Titel der zitierten Primärquelle. Hier sind aufgrund der uneinheitlichen Zitierweise Duplikate möglich.
- `cbdb_text_id` – ID des Textes in der *CBDB*, falls ermittelbar.[142]
- `title_py` – Die Aussprache des Titels in *Hanyu Pinyin*.
- `title_western` – Englische Übersetzung des Titels, falls vorhanden.
- `startyear` – Frühestmögliches ermitteltes Entstehungsjahr des Textes.
- `endyear` – Spätestmögliches ermitteltes Entstehungsjahr des Textes.
- `dynasty` – Name der Dynastie, während der das Werk entstanden ist.
- `estimate` – Ungenaue Angaben zu Jahr bzw. Zeitraum der Veröffentlichung, bzw. Schätzungen werden mit 1 gekennzeichnet.
- `usecount` – Zähler, wie häufig der Text insgesamt im *DHYDCD* zitiert wurde.
- `useinfirstcount` – Zähler, wie häufig der Text als ältester Beleg im *DHYDCD* angegeben wurde.
- `source` – Quelle der ermittelten Metadaten in den Feldern `startyear`, `endyear` und `author`.[143]
- `author` – Autor:in des Textes, sofern angegeben bzw. ermittelbar.

— 3. `the_citations` – Verortung *aller* Belegstellen im *DHYDCD*. Datenmodell: $n:m$ – zu jedem Schlagwort in `the_words` können n Zitate mit Quellenangabe vorhanden sein. Jedes dabei zitierte Werk aus `the_books` kann in m Einträgen verwendet werden. Mithilfe dieser Tabelle können sowohl alle in einem Wörterbucheintrag zitierten Texte selektiert werden, als auch alle Einträge gefunden werden, in denen bestimmte Texte zitiert werden.

- `id` – Eindeutige ID des Quellenzitats, Reihenfolge entspricht dem *DHYDCD*.

140 Es wird z. B. aus Gedichten mit dem naheliegenden Titel *Qiu ye* 秋夜 („Herbstnacht“) von insgesamt zwölf Urheber:innen zitiert, die während zehn unterschiedlichen Dynastien über einen Zeitraum von insgesamt 1.646 Jahren gelebt haben, von der Jin 晋 bis zur Qing 清-Zeit. Dies ist zwar ein eher extremes Beispiel, aber es finden sich insgesamt über 3.000 Quellen, bei denen zwei Texte gleichen Namens in unterschiedlichen Dynastien verfasst wurden und unbedingt unterschieden werden müssen. In `the_books` nicht unterschieden werden Quellenangaben, bei denen aus demselben Werk, aber aus unterschiedlichen Kapiteln bzw. Abschnitten oder *juan* 卷 zitiert wird. So werden etwa die Referenzen zum *Liexian zhuan* 列仙传 („Biographien beispielhafter Heiliger“) „ 汉刘向《列仙传 · 骑龙鸣》“ und „ 汉刘向《列仙传 · 方回》“ als eine einzige Primärquelle mit mehreren Verwendungen betrachtet. Vgl. *DHYDCD*, * 一, 一丸泥. Unterscheiden sich hingegen zwei Angaben weder im Titel noch in der Dynastie, aber im Autor, werden zwei separate Quellen registriert. So haben z. B. in der Jin 晋-Zeit zwei Historiker, Sun Chuo 孙绰 und Guo Yuanzu 郭元祖 die Biographien aus dem *Liexian zhuan* kommentiert: „ 晋孙绰《列仙传赞 · 老子》“ Siehe *DHYDCD*, 頤生; „ 晋郭元祖《列仙传赞 · 方回》“ *DHYDCD*, 冥神.

141 Siehe dazu Abschnitt 5.5.2, ab S. 128.

142 Siehe dazu Abschnitt 5.5.3, ab S. 132.

143 Ebd.

- `word_id` – ID des Worteintrags, in dem das Zitat gefunden wurde – referenziert auf `the_words.id`.
- `sub_id` – interne ID des Untereintrags, bei mehreren gelisteten Bedeutungen.
- `book_id` – ID der Primärquelle, aus der das Quellenzitat stammt – referenziert auf `the_books.id`.

Zur Veranschaulichung sind die wichtigsten Datenspalten des oben beschriebenen Modells am Beispiel des Lexemeintrags zu *shiyou* 石油 (ID Nr. 208.670) illustriert (Abb. 5.2). Die Spalte `the_words.book_id` referenziert auf das Werk mit der frühesten Belegstelle: *Meng xi bi tan* (ID Nr. 96) – eine tatsächlich frühere Belegstelle (`earliest_evidence_book_id`) konnte nicht ermittelt werden. In der Tabelle `the_books` finden sich unter der referenzierten ID die entsprechenden Metadaten, inklusive dem Jahr (bzw. Zeitraum) der Veröffentlichung und, sofern vorhanden, der Autor, hier Shen Kuo 沈括. Die Daten konnten in diesem Fall aus der *CBDB* ergänzt werden, die IDs der entsprechenden Einträge stehen in `the_books.cbdb_text_id` bzw. `the_books.cbdb_author_id`.[144] Bei insgesamt 414 Lexemen ist die älteste angegebene Belegstelle ein Zitat aus dem *Meng xi bi tan* (`useinfirstcount`).

Die Tabelle `the_citations` gibt zudem die *n : m* Beziehung zwischen Wörterbucheinträgen und zitierten Texten an, so dass nicht nur das älteste, sondern *alle* in dem Eintrag angegebenen Zitate verortet werden können. Zusätzlich zur *Locus classicus*-Angabe werden noch Stellen aus dem *Ben cao gangmu*, *Fan ke pian* und *Ye cao* zitiert.[145]

the_words

id	208670
cleanword	石油
pinyin	*shíyóu*
book_id	96
earliest_evidence_book_id	*NULL*

the_books

id	96
clearbook	梦溪笔谈
cbdb_text_id	2206
startyear	1095
endyear	1095
dynasty	宋
useinfirstcount	414
source	CBDB
author	沉括
cbdb_author_id	1450

the_citations

word_id	book_id	the_books.clearbook
208670	96	梦溪笔谈
208670	1793	本草纲目
208670	45960	番客篇
208670	67	野草

Abbildung 5.2 Beispielzeilen aus den Tabellen `the_words`, `the_books`, `the_citations`

Bei einer geringen Anzahl von Wort- oder Zeicheneinträgen,[146] bei denen unterschiedliche Bedeutungen angegeben sind, wird im ersten Untereintrag kein Beleg angegeben.[147]

144 Siehe dazu Abschnitt 5.5.3, ab S. 132.
145 Siehe auch S. 123.
146 Betroffen sind ca. 1.400 bzw. 0,3 % der Einträge. (Eigene Berechnung / Zählung mit dem regulärem Ausdruck `(?<!([0-9\.]))1\.[^《]+2\..+《.+)`
147 Ein Beispiel hierfür ist der Eintrag zu *daqing* 大青. Unter „1." ist angegeben, es handele sich um eine Art 1–2 Meter hohen Strauch, dessen äußere Erscheinung dann beschrieben wird – jedoch ohne Angabe einer Belegstelle. Erst unter „2." ist *daqing* als Bezeichnung für eine Art Farbpigmentstein, *bianqing* 扁青, angegeben, wofür dann das Ming-

Zudem kommt es vereinzelt vor, dass die älteste *belegte* Bedeutung nicht zuerst angegeben wird, wodurch solche Lexeme zunächst als „zu neu“ eingestuft würden.[148] Um dem entgegenzuwirken wird für Einträge, bei denen mehrere Bedeutungen angegeben sind und – sofern vorhanden – die chronologische Reihenfolge der verwendeten Dynastieangaben von der tatsächlichen Dynastiefolge abweicht, diejenige Primärquelle als früheste Belegstelle angenommen, die die früheste Dynastieangabe aufweist.[149] Dieses Vorgehen sei anhand des Eintrags zu *muguang* 目光 kurz geschildert.

【目光】
1. 眼睛的光芒。明 高启《猛虎行》：“目光燀燀當路坐，將軍一見弧矢墮。”[…]
2.识见；见解。宋 梅尧臣《梦睹》：“既非由目光，所見定何稟。”[…][150]

Bei der unter „1.“ angegebenen Bedeutung, *„yanjing de guangmang* 眼睛的光芒“ (etwa „[klarer] Blick, [klare] Sicht“ usw.) wird ein Ming-zeitliches Werk, *Meng hu xing* 猛虎行 von Gao Qi 高启 (ca. 1336–1374) als Beleg zitiert, bei der zweiten Bedeutung („Erfahrung, Einsicht, Verstehen“) dann zuerst das Gedicht *Meng du* 夢睹 („Traumbeobachtung“) des Song 宋-zeitlichen (960–1279) Dichters Mei Yaochen 梅堯臣 (1002–1060). Durch die Dynastieangaben ist die nicht-chronologische Anordnung der Glossen erkennbar und die Song-zeitliche Textstelle kann als früheste enthaltene Belegstelle erkannt werden.

Bei einigen Einträgen wird die älteste Bedeutung nicht zuerst aufgeführt bzw. die zuerst genannte Belegstelle ist nicht gleichzeitig die älteste ist. Wenn möglich findet deshalb bei der Extraktion der Lexikalisierungsdaten eine chronologische Reihenfolgenprüfung auf Basis des Dynastiemodells statt. Bei etwa 1.800 Einträgen wird daher nicht die erste im *DHYDCD* genannte Quelle als Zeitpunkt der Lexikalisierung verwendet, sondern stattdessen diejenige mit der frühesten Dynastieangabe.[151]

5.5.2 Verwendung der Metadaten aus dem *DHYDCD*

Zu vielen der Quellenangaben lassen sich, wie oben erklärt, Angaben zu Dynastie und Autor:in direkt aus dem *DHYDCD* ermitteln. Für einige wenige Texte, v. a. Zeitungsartikel, ist der Jahrgang angegeben. Da die Position der *Named Entities* klar definiert ist und ein Abgleich mit der *CBDB* oder die Verwendung von *NER*-Methoden eine unnötige Limitation der erkennbaren Namen zur Folge hätte, wird hier ein regelbasiertes Vorgehen gewählt.[152] Die wesentlichen Herausforderungen, die bei der Extraktion der Daten beachtet werden müssen, sind im Folgenden dokumentiert:

---

zeitliche *Ben cao gangmu* 本草綱目 als Beleg herangezogen wird. Siehe *DHYDCD*, 大青. Tatsächlich ist *daqing* aber in seiner ersten Bedeutung bereits im deutlich älteren *Baopu zi* 抱朴子 zu finden. siehe Ge Hong 葛洪 2020 [Anfang 4. Jh.] *Baopuzi* 抱朴子. Digitalisierte Version der *Sibu congkan*-Ausgabe von *Baopuzi nei wai pian* 《四部叢刊初編》本《抱朴子內外篇》. URL: https://ctext.org/baopuzi (besucht am 20. 09. 2020), *neipian* 內篇, *zhi li* 至理.

148 Davon betroffen sind etwas weniger als 0,5 %, bzw. etwas mehr als 1.800 aller Einträge des *DHYDCD*.

149 Zur Erkennung solcher Einträge wird ein einfacher Sortieralgorithmus genutzt: Beginn und Ende der angegebenen Dynastien werden nachgeladen und die Bedeutungen so nach dem ersten Jahr der jeweils angegebenen Dynastie (*startyear*) sortiert. Weicht die entstandene Sortierung von der ursprünglichen Sortierung im Eintrag ab, wird das Lexem in der Tabelle `the_words` als `unordered` markiert. Nur wenn alle Bedeutungen geeignete Belege haben, kann dieser Eingriff vorgenommen werden.

150 *DHYDCD*, 目光. Unterstreichung und farbliche Hervorhebungen durch den Verfasser.

151 Die betroffenen Einträge werden in der Spalte `the_words.unordered` markiert.

152 Vgl. dazu Kapitel 4.7, ab S. 97. Versuche mit *CKIP Tagger* und *CKIP Transformers* haben zudem Probleme bei DynastieAutor:in-Zeichenfolgen gezeigt.

Grundsätzlich lassen sich die Titel von zitierten Werken mit einem einfachen regulären Ausdruck erkennen: `《[^》]+》`, d. h. beliebig viele beliebige Zeichen innerhalb von *double angle brackets* 《》. Vor und nach den Klammern können jedoch ebenfalls relevante Informationen über Erscheinungsjahr[153] und Textgattung[154] angegeben sein. Diese Informationen helfen zudem, zahlreiche Werke gleichen Titels voneinander zu unterscheiden.

Angaben wie *zhu* 注 vor Werktiteln werden dem Titel des Werkes zugerechnet, wie z. B. bei Zheng Xuans 鄭玄 Kommentar zum *Lunyu* 論語 („东汉末郑玄注《论语》“[155]) Dies ist nicht unproblematisch, da z. B. das Zeichen *zhu* 注 auch in Namen vorkommen kann.[156]

Es ergibt sich folgender regulärer Ausdruck:

```
ur'(?:(?:释文引|注)[^)(*.“’””’。，；？！：、》>…】 () [] ]{0,9})?(?:《[^》]+》)(?:\d{4}|
    注|词|诗|曲|套曲)?(?![^“]*”)'
```

Extrahiert werden also zunächst die erwähnten, den Titel ergänzenden Angaben, sowie bis zu neun Zeichen vor Werksnamen, bis das vorhergehende Satzzeichen erreicht wird. Innerhalb dieser Stelle finden sich, sofern vorhanden, unterschiedliche Angaben zum zitierten Werk, die bestenfalls Dynastie und Autor:in enthalten.

Zitierweise und Dynastiesystem

Eine möglichst vollständige Erkennung der Dynastieangaben ist wichtig, da sie oft die einzige chronologische Angabe an den Quellenzitaten darstellen. Dabei verwendet das *HYDCD* – wie bereits angesprochen – ein eigenwilliges System, das im Sprachverständnis der Herausgeber begründet sein dürfte: Antike Werke, wie etwa das *Yijing* 易經,[157] das *Shijing* 詩經,[158] das *Shujing* 書經 bzw. *Shangshu* 尚書[159] oder *Zhuangzi* 莊子[160] werden fast grundsätzlich ohne Angabe von Dynastie und Autor:in zitiert, wenn man sich nicht auf eine bestimmte Ausgabe bezieht. Eine Ausnahme davon sind Werke, die Song Yu 宋玉 (ca. 319–298 v. u. Z.)[161] zugeschrieben werden.[162]

Bei Texten deren Autor:in unbekannt ist, die von einem Autor:innenkollektiv stammen oder unter chinesischen Gelehrten allgemein bekannt sind, wird die Angabe von Dynastie und Autor üblicherweise ebenfalls weggelassen. Ein typisches Beispiel dafür sind die offiziellen Dynastiegeschichten (*zhengshi* 正史). So wird etwa die *Neue Geschichte der Fünf Dynastien* (*Xin Wudai shi* 新五代史) konsequent nur als Werktitel zitiert, obwohl es ebenso unproblematisch wäre, den Text Ouyang Xiu 歐陽修 (1007–1072)[163] zuzuschreiben und in die Song 宋-Zeit zu datieren.[164]

153 z. B. im Format „《人民日报》1982.3.14“; „《人民日报》1957.10.29“. *DHYDCD*, 交售, 交議.
154 Beispiele dafür sind Angaben wie „李善注引《广雅》“ und „陆德明释文引《广雅》“. Ebenfalls nachgestellt finden sich Hinweise auf Kommentare (*zhu* 注), sowie Gedichte und Lieder (*ci* 詞, *shi* 詩, *qu* 曲, *taoqu* 套曲). *DHYDCD*, 壘 2 壘, 廉劌.
155 *DHYDCD*, 張侯論.
156 Der Name von genau 50 Personen aus der *CBDB* endet mit *zhu* 注.
157 Siehe z. B. *DHYDCD*, 左右.
158 Siehe z. B. *DHYDCD*, 左右.
159 Siehe z. B. *DHYDCD*, 一心. Das *Buch der Urkunden* wird stets als 《*Shu*》《书》 zitiert.
160 Siehe z. B. *DHYDCD*, 朝三暮四.
161 Shih Hsiang-lin 施祥林 und David R. Knechtges 2014: „Song Yu 宋玉“. In: *Ancient and Early Medieval Chinese Literature. A Reference Guide. Part Two*. Hrsg. von David R. Knechtges und Chang Taiping 張泰平. Handbook of Oriental Studies. Leiden: Brill, S. 1007–1022, S. 1007.
162 Hier wird *Zhanguo Chu* 战国楚, *Zhanguo shi* 战国时 („die Zeit der Streitenden Reiche“), oder manchmal nur *Zhanguo* 战国 als Dynastie angegeben. Der Grund für diese Ausnahme erschließt sich nicht, vermutlich ist sie auf eine Ungenauigkeit im ursprünglichen Karteikartensystem der *HYDCD*-Herausgeber zurückzuführen. Siehe *DHYDCD*, 對問, 大王風, 更唱迭和.
163 Wilkinson 2000, S. 504.
164 Siehe z. B. *DHYDCD*, 三十六英雄.

Erst bei Werken ab der Han 漢-Zeit finden sich regelmäßig Angaben im Stil DynastieAutor:in《Werksname》.[165] Wie schon die aus dem Werk Song Yus 宋玉 zitierten Stellen zeigen, sind die Dynastienennungen nicht konsequent einheitlich gehalten. Eine Quelle aus der östlichen Han (*Dong Han* 東漢, 25–220) kann etwa mit der Angabe *Dong Han* 东汉, oder lediglich *Han* 汉 versehen sein.[166]

Eine weitere Kuriosität in der Zitierweise stellt der Umgang mit Texten dar, die nach Ende der Qing 清-Zeit (1644–1911) erschienen sind. Bei republikzeitlichen Werken oder Werken aus der Volksrepublik wird grundsätzlich nur der Autor genannt.[167] Da – wie oben erläutert – auch bei ganz frühen oder als allgemein bekannt geltenden Werken die Zeitangabe fehlt, lässt sich diese Erkenntnis leider nicht ohne Weiteres für die Verortung solcher Werke ins 20. Jh nutzen.

Insgesamt sind die chronologischen Angaben im *DHYDCD* für den Zeitraum vom Beginn der Han-Dynastie im Jahr 206 v. u. Z. bis 1911 am vollständigsten. Die unzähligen Quellen aus der Zeit davor und danach, sowie weitere allgemein bekannte Texte, lassen sich nur mit zusätzlichen Daten einordnen.

Aus den teils unorthodoxen oder inkonsequenten Angaben des *DHYDCD* ergibt sich folgendes Dynastiesystem (Tabelle 5.2),[168] das zur Erkennung der Angaben bzw. zeitlichen Einordnung der zitierten Werke genutzt wird.[169] Im *DHYDCD* uneinheitliche Angaben, z. B. *Han* 漢 und *Han dai* 漢代 bzw. *Nan Qi* 南齊 und *Nanchao Qi* 南朝齊 führen darin zu entsprechenden Mehrfacheinträgen.

Tabelle 5.2 Ergänztes Dynastiesystem des *HYDCD*, chronologisch nach Anfangsjahr

Dynastie	正體	*DHYDCD* 简体	von	bis	# zit. Werke
Streitende Reiche[170]	戰國	战国	-1030	-223	354
Chu [Streitende Reiche] (*Zhanguo Chu*)	戰國楚	战国楚	-1030	-223	19
Yan [Streitende Reiche] (*Zhanguo Yan*)	戰國燕	战国燕	-1030	-223	12
Qin	秦	秦	-221	-206	57
Han	漢	汉	-206	220	1.522
= Han (*Han dai*)	漢代	汉代	-206	220	45
Östliche Han (*Dong Han*)	東漢	东汉	25	220	84
Wei [Drei Reiche] (*Sanguo Wei*)	三國魏	三国魏	220	265	927
Shu [Drei Reiche] (*Sanguo Shu*)	三國蜀	三国蜀	221	263	70
Wu [Drei Reiche] (*Sanguo Wu*)	三國吳	三国吴	222	280	58
Jin	晉	晋	265	420	2.035
Frühere Qin [16 Reiche] (*Qian Qin*)	前秦	前秦	350	394	10
Nördliche Wei (*Bei Wei*)	北魏	北魏	386	534	138
Nördliche Liang [16 Reiche] (*Bei Liang*)	北涼	北凉	401	439	35
Song [Südliche Dynastien] (*Nanchao Song*)	南朝宋	南朝宋	420	479	1.117

165 So z. B. im Eintrag zu *Mao Nü* 毛女 („Haarfrau"; eine besonders behaarte Heilige beschrieben, die am *Huàshan* 華山 beheimatet sein soll). „传说中得道于华山的仙女。汉刘向《列仙传·毛女》：“毛女者[…]” *DHYDCD*, 毛女.

166 Xu Shens 許慎 *Shuo wen jie zi* 說文解字 etwa wird einmal im Eintrag zu *xuxue* 鄦學 in der östlichen Hanzeit verortet, im Eintrag zu *shuo wen* 說文 lediglich allgemeiner in der Hanzeit. Doch damit nicht genug: vereinzelt wird auch noch die Angabe *Han Dai* 汉代 („Han-Dynastie") verwendet, wie im Eintrag zu *xingfa zhi*: „汉代班固的《汉书》“. *DHYDCD*, 鄦學, 說文, 刑法志.

167 Eine Unterscheidung in Republik und Volksrepublik wurde möglicherweise aus politischen Gründen vermieden, da eine Zuordnung von Werken aus der Zeit von 1912–1949 zur danach in Taiwan 台灣 weitergeführten Republik (*Minguo* 民國) als Anerkennung ihrer Legitimität gedeutet werden könnte.

168 Zeitangaben übernommen aus Vogelsang 2012, S. 24. Die letzte Spalte gibt an, bei wie vielen unterschiedenen zitierten Werken die jeweilige Dynastieangabe verwendet wurde.

169 Zwar enthält der Indexband des *HYDCD* selbst eine Dynastietabelle, die dort verwendeten Bezeichnungen stimmen aber nicht zuverlässig mit der tatsächlich verwendeten (inkonsequenten) Zitierweise überein. Siehe *HYDCD*, Bd. 13, S. 3–7.

Tabelle 5.2 (Fortsetzung)

Dynastie	正體	*DHYDCD* 简体	von	bis	# zit. Werke
Qi [Südliche Dynastien] (*Nan Qi*)	南齊	南齐	479	502	2
= Qi [Südliche Dynastien] (*Nanchao Qi*)	南朝齊	南朝齐	479	502	523
Liang [Südliche Dynastien] (*Nanchao Liang*)	南朝梁	南朝梁	502	587	2.954
Qi [Nördliche Dynastien] (*Bei Qi*)	北齊	北齐	550	578	110
Zhou [Nördliche Dynastien] (*Bei Zhou*)	北周	北周	557	581	515
Chen [Südliche Dynastien] (*Nanchao Chen*)	南朝陳	南朝陈	557	589	448
Sui	隋	隋	581	618	462
Tang	唐	唐	618	907	31.921
Frühere Shu [Zehn Reiche] (*Qian Shu*)	前蜀	前蜀	903	925	1.095
Fünf Dynastien *Wudai*	五代	五代	907	960	647
Liao	遼	辽	947	1115	33
Song	宋	宋	960	1279	31.678
Jin	金	金	1115	1234	1.962
Yuan	元	元	1234	1367	6.949
Ming	明	明	1368	1644	14.991
Qing	清	清	1644	1911	22.028
Taiping Tianguo[171]	太平天國	太平天国	1851	1864	195
Republik (*Minguo*)[172]	民國	民国	1912	1992	[702]

In Anbetracht dieser Erkenntnisse werden folgende möglichen Zitierweisen berücksichtigt. Es wird dabei immer jeweils ein zitiertes Werk unterschieden, wenn sich Angabe von Dynastie, Autor:in *oder* Titel unterscheiden.

— 1. **DynastieAutor:in**, z. B. „宋司马光《乞罢免役钱状》“.[173] In diesem Fall soll Song 宋 als Dynastie, Sima Guang 司马光 als Autor und *Qi ba mianyi qian zhuang* 乞罢免役钱状 als Titel der Primärquelle extrahiert werden. Da kein zuverlässiges Muster für das Erkennen chinesischer Namen existiert, ist die Aufgabe, Dynastie und Autor:in zu trennen nicht trivial.[174] Die meisten Fälle lassen sich mit folgenden Annahmen abdecken:

— 1.1 **Personennamen** haben mindestens zwei und maximal sieben Zeichen: Vornamen bestehen in aller Regel aus 1–2 Zeichen, Nachnamen haben mindestens ein Zeichen und können nicht länger als vier Zeichen sein. Im Extremfall des letzten Kaisers der Qing 清-Dynastie Aixinjueluo Puyi 愛新覺羅 · 溥儀 kommen wir zusammen mit dem sonst im *HYDCD* kaum in Namen verwendeten Mittelpunkt („ · “) auf die Maximallänge von sieben Zeichen. Für jeden gefundenen *String*, der ***keinen*** Dynastienamen aus Tabelle 5.2 enthält und doch länger als sieben Zeichen ist, kann davon ausgegangen werden, dass es sich nicht (oder zumindest nicht nur) um einen Namen handelt.

170 Da im *HYDCD* nicht zwischen westlicher Zhou (*Xi Zhou* 西周, 11. Jh. –771 v. u. Z.), Frühlings- und Herbstzeit (*Chunqiu* 春秋, 722–482 v. u. Z.) und der Zeit der streitenden Reiche (*Zhanguo* 戰國, 453–221 v. u. Z.) unterschieden wird, ist hier der gesamte Zeitraum der *Zhou*-Dynastie(n) angegeben.

171 Das „Himmlische Reich des höchsten Friedens“ (*Taiping Tianguo* 太平天國) wird von Historiker:innen weniger als legitime Unterbrechung der Qing 清-Herrschaft, denn als Aufstand (*luan* 亂) gewertet.

172 Die Angabe *Minguo* 民國 (Republik, 1912–) wird im *(D)HYDCD* nicht explizit gemacht, genauso wenig wie *Renmin gongheguo* 人民共和國 (Volksrepublik, 1949–), obwohl durchaus Belege aus Texten angegeben werden, die nach 1912 verfasst wurden. Als Enddatum ist aus praktischen Gründen das Jahr 1992 angegeben, da keine neueren Belege im *DHYDCD* vorhanden sind.

173 *DHYDCD*, 朝三暮四.

174 Siehe dazu auch 4.7, ab S. 97.

— 1.2 Alle **Dynastieangaben** entsprechen der Liste in Tabelle 5.2. Dass die Bezeichnungen teilweise überlappen, kann durch absteigende Sortierung nach Länge umgangen werden.[175]

Einige Dynastiebezeichnungen treten auch als Familienname auf. Bei den *xing* 性 Qing 清, Ming 明, Yuan 元, Jin 金, Song 宋, Tang 唐, Sui 隋, Jin 晉, Han 漢 Qin 秦 ist daher besondere Vorsicht geboten. Es ist dabei etwas wahrscheinlicher, dass jemand den Namen einer vorangegangenen Dynastie als Nachnamen trägt als den Namen einer zukünftigen Dynastie.[176] Zur Minimierung dieser Problematik werden Dynastien mit gleicher Anzahl Zeichen also aufsteigend chronologisch sortiert. Bei Dynastien mit einer Zeichenlänge > 1 ist gesichert, dass es sich nicht um den Familiennamen einer Autor:in handelt.
Wie kann jedoch unterschieden werden, ob es sich bei einer Zeichenfolge wie 元麻革 um den Namen einer Person (Yuan Mage) oder um eine *Yuan* 元-zeitliche Person namens Ma Ge handelt? Absolute Sicherheit kann nur die Recherche des zitierten Werks bzw. der möglichen Namen geben. Mittels einer **Liste bekannter Familiennamen** können zumindest mögliche Familiennamen erkannt werden.[177] Dass Ma 麻 als *xing* nachgewiesen ist, macht es zumindest wahrscheinlich, dass es sich um eine Angabe im Format DynastieAutor:in handelt.

— 1.3 **Aufeinanderfolgende Dynastienamen** wie bei Tang Song Jing 唐宋璟[178] lassen zudem darauf schließen, dass es sich beim zweiten Zeichen um den Familiennamen des Autors handelt. Eine Sicherheit besteht dennoch auch hier nicht, wie der Name Song Qinghai 宋清海 (geb. 1947) beweist.[179]

— 2. **Autor:in** – nur der Name der Autorin oder des Autors.[180]

— 3. **Dynastie** – bloße Angabe des Dynastienamens, z. B. „北魏《元灵耀墓志》：“少傾乾蔭，孤苦自立。”“.[181] Diese Zitierweise kommt selten vor.

— 4. Zeitangabe als **Datum oder Jahreszahl**: „《花城》1981 年第 3 期：“恰好她的父亲是个热心研制中草药的老人[…]“.[182] Solche Angaben finden sich bei Zeitschriften oder Tageszeitungen in der Regel direkt hinter dem Titel der Primärquelle. In selteneren Fällen finden sich Jahreszahlen auch im Titel, z. B. „《<1958 年儿童文学选>序言》“.[183] Ebenfalls möglich sind Jahreszahlen in vollbreiten Unicode-Ziffern (z. B. „１９５７“).[184] Chinesische Jahreszahlen wie „一九二九年“

175 D. h. *Zhanguo Chu* 战国楚 (drei Zeichen) vor *Zhanguo* 战国 (zwei Zeichen).
176 Diese Annahme ist nicht statistisch erforscht. Es finden sich aber in den aus dem *DHYDCD* extrahierten Metadaten z. B. etliche Qing 清-zeitliche Autor:innen mit dem Nachnamen Song 宋 (Song Weipan 宋維藩, Song Qianxu 宋潛虛, Song Yongyue 宋永岳, Song Xiangfeng 宋翔鳳, Song Xuezhu 宋學洙, Song Wan 宋琬, Song Dazun 宋大樽, Song Xian 宋銑 usw.), jedoch nur ein einziger Song-zeitlicher Autor, dessen Name – überdies ein Pseudonym – mit dem Zeichen *qing* 清 beginnt: *Qingyuan Zhenjun* 清源真君, der „wahren Fürst der klaren Quelle“. Allgemeingültig ist diese Annahme jedoch nicht, da sich bereits zur Tang 唐-Zeit Personen mit dem Nachnamen Song 宋 – bereits vor der Tang-Zeit der Name einer Dynaste – finden. Vgl. *DHYDCD*, 一枕, 一線, 井井有法, 屯 2 田, 中 2 率, 升堂拜母, 脩辭, 俯燭.
177 Eine zum Abgleich mit den Inhalten des *DHYDCD* in Kurzzeichen erzeugte Liste aller 1.549 ersten Zeichen von Familiennamen wurde zu diesem Zweck aus der *CBDB* erzeugt. Die Liste wurde mit weiteren Erkenntnisen aus dem *DHYDCD* angereichert und die einsilbigen Dynastienamen (s. o.) entfernt.
178 *DHYDCD*, 養老.
179 *DHYDCD*, 論 2.
180 Siehe dazu auch das Beispiel auf S. 123.
181 *DHYDCD*, 乾 2 蔭.
182 *DHYDCD*, 中草藥.
183 *DHYDCD*, 恰切.
184 „【踏察】勘察，探测。《１９５７散文特写选》序：“你可以认识从北大荒的踏察人员到海南岛的盐业工人[…]“ *DHYDCD*, 踏察.

(1929) werden hingegen als Bestandteil des Titels betrachtet.[185]

— 5. **Sonstige** häufiger angegebene Informationen, z. B. „ 中国近代史资料丛刊“ („Collectanea of Materials on Modern Chinese History“) als Titel einer Buchreihe[186] oder „ 马王堆汉墓帛书甲本“[187] als Hinweis auf einen bestimmten Band bzw. Version des danach angegebenen Werkes. In beiden genannten Fällen kann eine Dynastieangabe aus einer Liste hinzugefügt werden.

— 6. **Keine Metadaten**, wie in der Regel z. B. bei Klassikerzitaten.[188] Hier kann zunächst lediglich der Titel extrahiert werden.

Weitere, seltenere Fälle werden zur Vermeidung von *Over-Engineering* nicht berücksichtigt.[189] Selbst durch ausgefeiltes Parsing lässt sich bei der Extraktion der Metadaten keine Perfektion erreichen, vor allem die Trennung von Dynastie und Autor:in ist problematisch. Angaben wie „ 宋宋祁《授龙图阁谢恩表》[…]“[190] können durch das doppelte Auftreten von Song 宋 als Dynastie- und Familienname verhältnismäßig gut erkannt werden. Bei Angaben wie Tang Wuke 唐无可 ohne zusätzliche Informationen zu erkennen, ob Tang hier tatsächlich der Familienname des Autors ist, bleibt unmöglich.[191] Ein noch komplexeres Regelwerk würde die Nachvollziehbarkeit der erzeugten Daten zudem immer weiter verschlechtern.

5.5.3 Gewinnung von Daten aus der *China Biographical Database*

> *„Dowerjai, no prowerjai! Доверяй, но проверяй!*
> *Vertraue, aber prüfe nach!“*
>
> Russisches Sprichwort

Die *CBDB*[192] eignet sich in zweierlei Hinsicht, um die bereits aus dem *DHYDCD* gewonnenen Metadaten zu verdichten. Aus der Tabelle `text_codes` lässt sich das Erscheinungsjahr für einige der zitierten Primärtexte ermitteln. Wenn dies nicht gelingt, oder der Text nicht verzeichnet ist, können gegebenenfalls die Lebensdaten von Autor:innen ergänzt werden: Deren Lebensspanne ist in der Regel deutlich kürzer, als die – wenn überhaupt – im *DHYDCD* angegebenen Dynastien. In „ungünstigen“ Fällen wie Tang 唐, Song 宋, Ming 明 oder Qing 清 beschreiben sie einen

185 z. B. „[…]殷夫《一九二九年的五月一日》诗:“我们总同盟罢业,[…]“. *DHYDCD*, 罷業. Zitiert wird hier ein Gedicht mit dem Titel „Der 1. Mai 1929“ von Xu Xiaojie 徐孝杰 (1909–1931), der unter dem Pseudonym Yin Fu 殷夫 veröffentlichte. Vgl. OCLC 2019, `1ccn-n82080968` 殷夫 1910-1931. Ob das Gedicht wirklich 1929 verfasst wurde, geht aus dem Titel keineswegs hervor.

186 Siehe z. B. *DHYDCD*, 三點會, „【三點會】天地会的别名。中国近代史资料丛刊《辛亥革命·兴中会革命史要》:“本来在广州的客籍人,多半加入三点会。“ *Zhongguo Jindai Shi Ziliao Congkan* 中国近代史资料丛刊 ist eine 1951–1961 veröffentlichte Buchreihe über moderne chinesische Geschichte. Im hier gezeigten Beispiel wird aus dem Band über die Xinhai-Revolution (*Xinhai Geming* 辛亥革命) zitiert.

187 „Erstes Buch der Manuskripte aus den Han-Gräbern von Mawangdui 馬王堆“. Siehe z. B. *DHYDCD*, 才 2, „[…]通“哉1”。语气词。马王堆汉墓帛书甲本《老子·德经》:“以正之國,以畸用兵,以無事取天下,吾何以知其然也才。” […]“.

188 z. B. „《庄子·齐物论》:“狙公賦芧,曰[…]“ *DHYDCD*, 朝三暮四.

189 Ein Beispiel dafür wären die inkonsequenten Angaben einer Autorenmehrheit, etwa „ 夏丏尊叶圣陶《文心》“ *DHYDCD*, 規約. Zwar wird das *Wenxin* meistens mit genau diesen beiden Autoren, Xia Mianzun und Ye Shengtao, zitiert, vereinzelt wird aber auch die Reihenfolge umgekehrt oder nur Xia wird mit *deng* 等 („et al.“) genannt.

190 *DHYDCD*, 膚屑.

191 Siehe z. B. *DHYDCD*, 朦朧.

192 Siehe dazu auch Kapitel 4.7, ab S. 97.

Zeitraum von etwa 300 Jahren, während die durchschnittliche Lebensspanne einer in der *CBDB* katalogisierten Person 60 Jahre beträgt.[193]

Die Datenübernahme erfolgt in zwei Schritten:

— 1. Das **Erscheinungsjahr der Texte** wird – sofern verfügbar – ermittelt und zugewiesen. Wenn möglich werden auch Lebensdaten von Autor:innen bzw. Herausgeber:innen ergänzt, sofern sie eindeutig zugeordnet werden können. Als Treffer wird dabei die Übereinstimmung von Titel *und* Autor oder Titel *und* Epoche gewertet. Ist beides im *DHYDCD* nicht angegeben, so werden nur Informationen für eineindeutige Werktitel übernommen.[194]

— 2. Wenn in der *CBDB* keine geeigneten Informationen über den Text vorliegen, aber bereits die Autor:in ermittelt werden konnte,[195] werden die **biographischen Daten** zur genaueren chronologischen Einordnung der im *DHYDCD* zitierten Texte verwendet. Da auch Personennamen in beiden Datenbanken mehrfach vorkommen können,[196] werden die Daten nur dann übernommen, wenn eine ein-eindeutige Übereinstimmung besteht, oder wenn die Lebensdaten der Person zu einer bereits aus dem *DHYDCD* extrahierten Dynastieangabe passen.

Die beschriebenen Einschränkungen reduzieren das Potenzial der Datengewinnung aus der *CBDB* zwar, stellen aber sicher, dass deutlich weniger *false positives* übernommen werden. Für mehr als 100.000 der in `the_books` unterschiedenen Quellen können so Metadaten aus der *CBDB* ergänzt bzw. präzisiert werden.[197]

Da die Metadaten hier jedoch „unüberwacht“ gewonnen wurden, kann die Datenübernahme in einzelnen Fällen auch zu falschen bzw. zu späten Lexemdatierungen führen. So wird z. B. das *Xiaojing* 孝經 (*Klassiker der kindlichen Pietät*), ein Text aus dem konfuzianischen Kanon, der ziemlich sicher in die vorchristliche Zeit datiert werden kann,[198] durch einen in der *CBDB* gelisteten, gleichnamigen songzeitlichen Text mehr als 1.000 Jahre zu spät in das Jahr 1098 datiert. Das wirkt sich wiederum auf die zeitliche Einordnung aller 62 *DHYDCD*-Einträge aus, die das *Xiaojing* als ältesten Beleg zitieren.

Ergänzende Recherche von Metadaten

Da einige frühe bzw. kanonisierte Texte im *HYDCD* ohne Angabe von Dynastie oder Autor zitiert werden, kann durch die in den Abschnitten 5.5.2 und 5.5.3 beschriebenen Maßnahmen gerade für einige sehr häufig zitierte Texte keine chronologische Einordnung der zugehörigen Lexeme vorgenommen werden. Für die am häufigsten als *Locus classicus* zitierten Werke lohnt es sich daher – sofern möglich – das Erscheinungsjahr oder zumindest den ungefähren Zeitraum „von Hand“ zu

193 Errechnet aus 33.200 Datensätzen, für die Geburts- und Todesjahr zur Verfügung standen, per `select avg(c_deathyear - c_birthyear) from biog_main where c_birthyear != 0 and c_deathyear != 0 and c_deathyear > c_birthyear and (c_deathyear - c_birthyear) < 200`. (Da z. B. das Todesjahr des ming-zeitlichen Beamten Dong Fan 董璠 (1444–1526) in der *CBDB* mit 4526 angegeben ist (Siehe *CBDB*, 32020, 董璠.) – ist eine einfache Plausibilitätsprüfung angezeigt.)

194 So ist z. B. dem Qing-zeitlichen Kalligraphen Jin Renrui 金人瑞 (1610–1661) ein Werk mit dem Titel *Shiji* 史記 zugeordnet, siehe *CBDB*, 65871, 金人瑞. Im *DHYDCD* ist mit *Shiji* 史记 hingegen in der Regel das Sima Qian 司馬遷 (ca. 145–90 v. u. Z.) zugeschriebene Geschichtswerk gemeint.

195 Siehe dazu Abschnitt 5.5.2, ab S. 128.

196 Siehe dazu Kapitel 4.7, ab S. 97.

197 Siehe dazu auch die Daten und Visualisierungen zur Datenbank in Kapitel 5.7, ab S. 138.

198 Siehe William G. Boltz 1993a: „Hsiao ching 孝經“. In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 141–153, S. 143.

recherchieren und die ermittelten Daten zu ergänzen.[199] Für insgesamt 116.862 Lexeme liegen dadurch ergänzte oder genauere chronologische Daten vor.

5.5.4 Ergänzung um frühere Belegstellen

Eine immer wieder angeführte Kritik am *HYDCD* ist die Unzuverlässigkeit bei der Angabe der frühesten Belegstellen (*Locus classicus*).[200] Diese Problematik kann reduziert werden, indem die Belegstellen um Vorkommen in digital verfügbaren Texten ergänzt werden. Durch Verwendung vorhandener elektronischer Textsammlungen ist dies mit überschaubarem Aufwand möglich. Ein dafür geschriebenes *Python*-Skript verarbeitet die *types* der jeweiligen Texte und ergänzt Belegstellen in der Datenbank. Hierfür werden das Loewe-Korpus,[201] sowie die Volltexte der *zhengshi* 正史 genutzt.[202] Für Lexeme mit einer Länge von 1–3 Zeichen werden zusätzlich Daten des *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌)[203] für 1.000 zufällig ausgewählte Texte[204] herangezogen.

— 1. Für jeden der Korpustexte wird die „beste“ datierte `id` aus der Tabelle `the_books` nachgeschlagen.[205] Texte, die nicht zugeordnet werden können, etwa, weil sie nicht im *DHYDCD* verortet oder nicht datiert sind, werden übersprungen. Für die *Difangzhi* 地方誌 *n*-Gramm-Daten werden Einträge in der Tabelle `the_books` aus den Metadaten des Datensatzes ergänzt.

— 2. Alle 1–4- bzw. 1–3-Gramme der betrachteten Texte werden mit der Liste der Lexeme im *DHYDCD* abgeglichen, für die bereits Belegstellen bekannt sind (Tabelle `the_words`). Die Interpunktion der Korpustexte bleibt dabei unverändert.[206]

— 3. Ist der gerade betrachtete Korpus-Text älter als derjenige, der für ein Lexem im *DHYDCD* als *Locus classicus* angegeben ist, so wird seine in Schritt 1 ermittelte bzw. angelegte ID in die Spalte `earliest_evidence_id` bzw. `earliest_evidence_dfz_id` der Tabelle `the_words` geschrieben.[207] Falls im Verlauf eine noch ältere Belegstelle gefunden wird, wird die `earliest_evidence_id` überschrieben, bis die älteste im Korpus vorhandene Belegstelle dokumentiert ist. Die ursprüngliche Angabe aus dem *DHYDCD* bleibt durch die Verwendung der zusätzlichen Spalten erhalten. Die Verwendung der früheren Belegstellen bleibt so optional und die Nachvollziehbarkeit gewährleistet.

199 Siehe dazu die Auswertungen in Abschnitt 5.7.4, S. 150.
200 Siehe dazu Kapitel 5.3, ab S. 113.
201 Siehe T. Schalmey 2009, S. 104–106, einige Texte dieses Korpus sind nicht genau datierbar, siehe auch Kapitel 4.2, S. 66.
202 Siehe dazu auch Kapitel 2.3, ab S. 20.
203 *DFZ*.
204 Unter Ausschluss der Texte, die in Kapitel 6.1.1 (ab S. 158) bzw. 6.2.5 (ab S. 197) als Testdaten verwendet werden.
205 Falls mehrere Einträge desselben Titels bestehen, wird der mit der frühesten Datierung bevorzugt. Falls auch hier mehrere Einträge bestehen, wird derjenige Text verwendet, der am häufigsten als *Locus classicus* zitiert wurde. In SQL ausgedrückt: `where startyear is not null order by startyear asc, useinfirstcount desc limit 1`. Dass gerade bei häufig zitierten Werken zahlreiche Duplikate in `the_books` vorhanden sind, ist der ungenauen Zitierweise des *DHYDCD* geschuldet. Das Liu Xiang 劉向 zugeschriebene *Liexian zhuan* 列仙傳 („Biographien von Unsterblichen“) wird z. B. auf insgesamt acht unterschiedliche Weisen, teils indirekt (d. h. innerhalb eines Zitats aus einer anderen Quelle), zitiert. Siehe dazu auch Abschnitt 5.5.1, S. 123.
206 D. h. enthält ein Text z. B. die Zeichenfolge „人。人“ oder „人，人“, wird dies nicht als Belegstelle für das Lexem *renren* 人人 gezählt.
207 Um die Wirkung dieser Maßnahmen bewerten zu können (siehe S. 183), werden diese Belege in gesonderte Datenspalten gespeichert (siehe dazu auch Abschnitt 5.5.1, ab S. 123).

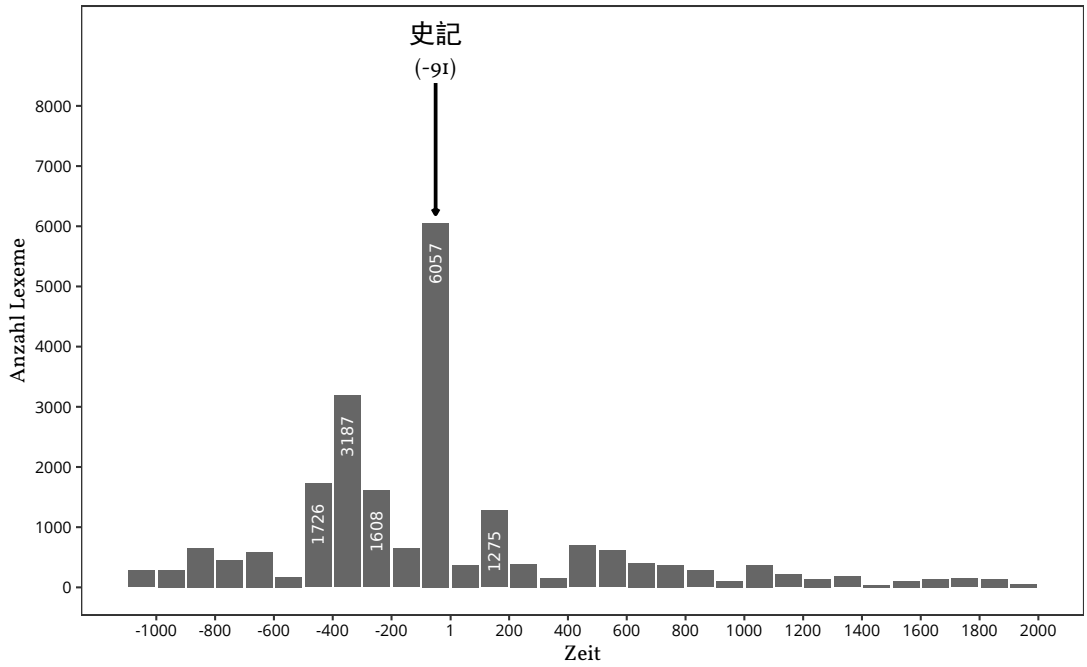


Abbildung 5.3 Neologismusprofil des *Shiji* 史記, ohne Korpusbelegstellen

Für 80.547 (etwa ein Viertel) der im *DHYDCD* lexikalisierten Zeichenkombinationen können so aus den Loewe und *zhengshi*-Korpora frühere Belegstellen ergänzt werden. Aus den *Difangzhi*-Daten werden für 10.325 Zeichenkombinationen frühere Verwendungen aus 731 Texten hinzugefügt.[208] Am Beispiel des *Shiji* 史記-Neologismusprofils (Abb. 5.3, 5.4)[209] wird der Effekt dieser Maßnahme deutlich sichtbar.

Da der Text selbst in den Trainingsdaten enthalten ist, weist die zweite Abb. fast ausschließlich Zeichenkombinationen auf, die *vor* oder auf das 1. Jh. v. u. Z. datiert sind. Trotz der offensichtlich sehr intensiven Rezeption des Texts durch die Herausgeber:innen des *DHYDCD*[210] lassen sich allein im *Shiji* für über 3.000 2–4 Zeichen-Kombinationen frühere Belegstellen finden. Dies ist allerdings nicht ausschließlich auf die Nachlässigkeit der Herausgeber:innen zurückzuführen. Zeichenkombinationen, die in einem der Korpustexte in einer abweichenden Bedeutung auftreten, sollten als *false positives* angesehen werden.[211] Mit 7.468 Lexemen, bei denen bereits im *DHYDCD* das *Shiji* als *Locus classicus* angegeben ist,[212] machen die ergänzten Stellen also 28,7 % der insgesamt mit dem *Shiji* belegbaren Lexeme aus.[213]

Wie viel später die frühesten Belege im *HYDCD* sein können und wie (un-)zutreffend die älteren Belegstellen sind, sei anhand zweier Beispiele veranschaulicht. Die Zeichenfolge *sudi* 宿地,

208 Zu den verwendeten Korpora siehe Kapitel 4.2, ab S. 64.
209 Die hier verwendete Darstellung wird in Kapitel 6.2 (ab S. 179) ausführlich erläutert.
210 Siehe auch Abschnitt 5.7.4, ab S. 150.
211 Im Gegensatz zu den hier ohne jegliche semantische Analyse verglichenen Zeichenfolgen, beziehen sich die Belege im *DHYDCD* auf konkrete Bedeutungen. Entsprechende Beispiele finden sich in Kapitel 6.2.3, ab S. 190.
212 Siehe dazu Tabelle 5.3, S. 150.
213 Der tatsächliche Anteil an *false positives* ist dabei schwer feststellbar. Da in den nachgelagerten Analysen (v. a. Kapitel 6.2, ab S. 179) Zeichenfolgen ebenfalls zunächst ohne jede semantische Analyse verglichen werden, lohnt sich der Aufwand einer genaueren, manuellen Analyse an dieser Stelle nicht.

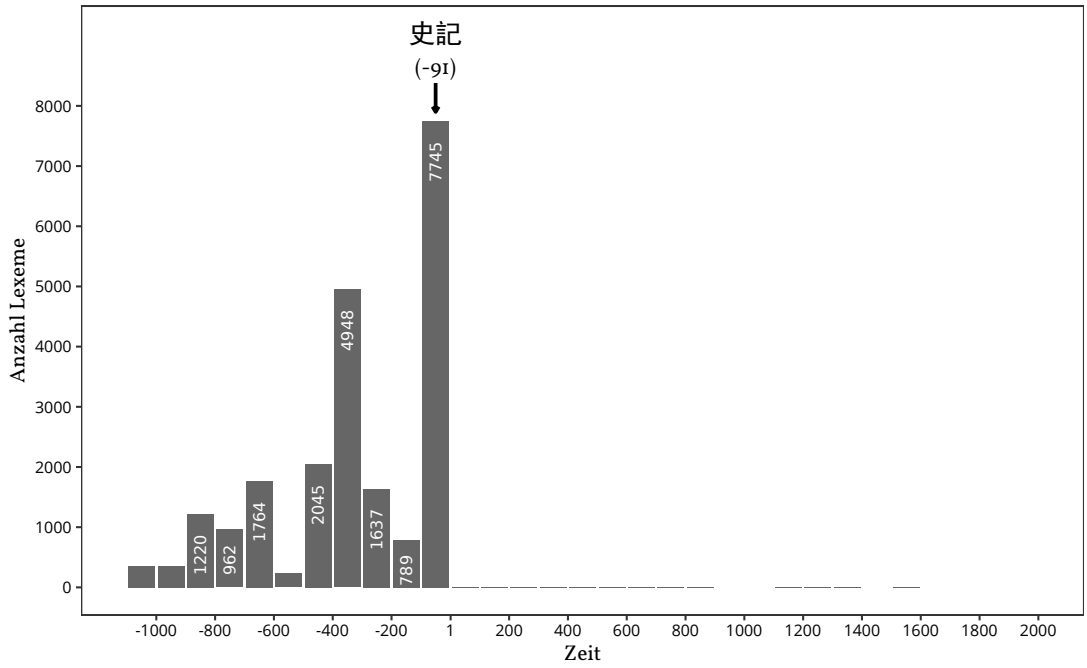


Abbildung 5.4 Neologismusprofil des *Shiji* 史記, mit Korpusbelegstellen

im *HYDCD* erklärt als *zhusu de difang* 住宿的地方, „ein Ort zum Übernachten" und erst mit einer Ausgabe der *Xinmin Evening News* (*Xinmin Wanbao* 新民晚報) vom 29. März 1987 belegt,[214] findet sich im *Shiji* bereits in ähnlicher Bedeutung.[215]

Als *Wan dan jun* 萬石君 („Zehntausend-*dan*-Fürst") bezeichnet Sima Qian 司馬遷 den hanzeitlichen Beamten Shi Fen 石奮 (gest. 124 v. u. Z.), dem im *Shiji* eine Biographie gewidmet ist.[216] Im gleichlautenden *HYDCD*-Eintrag wird zwar auf ihn verwiesen, die früheste Belegstelle für eine angelehnte Bedeutung stammt aber aus dem 1772 veröffentlichten *Gai yu cong kao* 陔餘叢考 von Zhao Yi 趙翼 (1727–1814), obwohl *wan dan* 萬石 bereits für die Han-Zeit belegt wird.[217]

Auch für die Daten aus der *CBDB* (siehe Kapitel 4.7, S. 97) können zeitlich frühere Belege für 12.541 Personennamen und 707 Ortsnamen gefunden werden. Hierzu werden einzigartige Personennamen mit einer Länge von drei Zeichen[218] aus der `biog_main`-Tabelle geladen. Ein Text wird als frühere Belegstelle gewertet, wenn sein `lastyear` früher ist als das angegebene Geburts- bzw. Indexjahr der gefundenen Person. Eindeutige Ortsnamen werden aus der `addresses`-Tabelle geladen und ein Text als frühere Belegstelle gewertet, wenn sein `lastyear` früher ist als die früheste Nennung des Ortsnamens.

214 Siehe *DHYDCD*, 宿地.

215 Vgl. z. B. Sima Qian 司馬遷 1959 [91 v. u. Z.] *Shiji* 史記 *(Records of the Grand Historian)*. Beijing 北京: Zhonghua shuju 中華書局, S. 476: „[…] 古者天子五載一巡狩, 用事泰山, 諸侯有朝宿地 […]". „In alten Zeiten, wenn der Kaiser alle fünf Jahre eine Inspektionsreise machte und am Taishan [Opfer]dienste ausführte, hatten alle Fürsten, die ihm [dort] die Ehre erwiesen einen Ort zum Übernachten […]".

216 Siehe ebd., S. 2763–2768.

217 Siehe *HYDCD*, Bd. 9, S. 462, 萬石, 萬石君. Vgl. auch Soffel 2004, S. 173.

218 Namen mit zwei Zeichen Länge weisen ein sehr hohes Ambiguitätspotenzial auf. Ausführlicher dazu siehe Kapitel 4.7, ab S. 97 und 6.2.2, S. 189.

Theoretisch ließe sich der beschriebene Vorgang mit beliebig vielen Texten wiederholen, um für jede Zeichenkombination die tatsächlich älteste überlieferte Belegstelle zu finden, um Belege für im *HYDCD* unbelegte Lexeme zu finden, oder um Vorkommen von Zeichenkombinationen diachron zu dokumentieren, die nicht als Lexem im *HYDCD* gelistet sind. Letzteres würde den Umfang der erzeugten Daten allerdings dramatisch vergrößern – wahrscheinlich ohne entscheidenden Mehrwert für den Anwendungszweck.[219]

Die nun zur Verfügung stehende diachrone Lexemdatenbank dient als Grundlage für die in Kapitel 6.2 und 6.3 (ab S. 179) behandelten Textdatierungsmethoden. Durch statistische Auswertung dieser Daten können zudem weitere Rückschlüsse auf die Machart des *HYDCD* gezogen werden.[220]

5.6 Das *DHYDCD* als diachrones Behelfskorpus

Die in Kapitel 3.3 vorgestellten Datierungsmethoden werden anhand diachroner Korpora evaluiert.[221] In Ermangelung eines umfangreichen, diachronen Korpus, welches den gesamten Zeitraum der schriftsprachlichen Texttradition abdeckt,[222] wird aus den Belegen im *DHYDCD* ein Behelfskorpus erzeugt.[223] Auch die Entstehung der offiziellen Dynastiegeschichten (*zhengshi* 正史) erstreckt sich zwar über einen großen Zeitraum, mit ingesamt nur 25 Texten bzw. in der Regel einem Text pro Dynastie eignen sie sich jedoch nicht für die Erzeugung von statistischen *chronon*-Sprachmodellen. Wünschenswert ist zudem eine ausgewogene Mischung relevanter Textgattungen.

Die Verwendung eines aus solchen Einzelsätzen erzeugten Korpus wurde bereits am Beispiel des *Oxford English Dictionary* beschrieben.[224] Obwohl es nicht als „vollständig ausgewogen und repräsentativ“[225] gelten kann, stellt es eine umfangreiche Sammlung natürlicher Sprache dar, wobei „die erfasste Zeitspanne von keiner anderen digitalisierten Quelle übertroffen wird.“[226] Dies gilt umso mehr für die insgesamt 919.280 Belegstellen aus dem *DHYDCD*, von denen 612.639 aus etwa 41.436 unterscheidbaren Texten zeitlich eingeordnet werden können.[227]

Um die Methodik der geringen Genauigkeit der zeitlichen Zuordnung der *attestations* anzupassen, wird eine grobe Einteilung in Zeiträume von 100 Jahren (*chronons*) mit einer Überlappung von jeweils 50 Jahren zum nächsten Subkorpus verwendet. Inhaltliche Überschneidungen werden dabei zugelassen. Ein Zeitraum von 100 Jahren erscheint sinnvoll, da im *HYDCD* – im Gegensatz zum *OED* – nicht das Jahr des Erscheinens der zitierten Texte angeben ist, sondern lediglich die Dynastie. Für einen Teil der Texte konnten durch Hinzuziehen externer

219 Zu Unterschieden bei der Nutzung von *n*-Gramm- und wort- bzw. lexembasierten Sprachmodellen siehe v. a. auch Kapitel 6.1, ab S. 156.
220 Siehe Kapitel 5.7, ab S. 138.
221 Siehe Kapitel 3.3, S. 55.
222 Siehe Kapitel 4.2, ab S. 62.
223 Zur Verwendung dieses Korpus siehe Kapitel 6.1.3, ab S. 171.
224 Siehe Kathryn Allan 2012: „Using OED data as evidence“. In: *Current Methods in Historical Semantics*. Hrsg. von Kathryn Allan und Justyna A. Robinson. Topics in English Linguistics. Berlin & Boston: Walter de Gruyter, S. 17–39, S. 19; siehe auch Hoffmann 2004, Hoffmanns Belegstellen aus dem *OED* ergeben ein diachrones Korpus des Englischen von etwa 2,4 Mio. Sätzen bzw. 33–35 Mio. Wörtern aus dem Zeitraum vom 11. bis zum 20. Jh, wobei erst ab dem 15. Jh. eine nennenswerte Menge an Textmaterial vorliegt.
225 Hoffmann 2004, S. 26.
226 Ebd., S. 26, übersetzt durch den Verfasser.
227 Zu Einschränkungen bei der Differenzierung von im *DHYDCD* zitierten Quelltexten siehe Kapitel 5.5.2, S. 127. Siehe auch Kapitel 5.5.3, S. 132.

Quellen wie der *CBDB* die Lebensdaten der Autor:innen oder sogar das Jahr der Veröffentlichung ergänzt werden.[228] Insgesamt wird dadurch eine durchschnittliche Genauigkeit von 76 Jahren erreicht.[229] Durch die Verwendung einer *chronon*-Länge von 100 Jahren ist gleichzeitig sichergestellt, dass für jeden Zeitraum eine für die Erkennung sprachgeschichtlicher Trends ausreichende Menge an Textmaterial extrahiert wird.[230]

Für jeden Zeitraum werden zunächst die relevanten Primärquellen aus der Tabelle `the_books` geladen. Anschließend werden aus der Tabelle `the_citations` die *DHYDCD*-Einträge mit Zitaten aus diesen Werken ermittelt.[231] Daraus werden nun die entsprechenden Belegstellen extrahiert und als chaotisches Pseudotext-Potpourri aneinander gereiht. Zur Veranschaulichung sei hier ein Auszug aus dem Subkorpus für den Zeitraum 1000–1100 gegeben, das sich aus insgesamt 11.191 Belegstellen aus 10.169 Texten zusammensetzt.

> […] 鄜延境内有石油……余疑其煙可用，試掃其煤以爲墨，黑光如漆，松墨不及也。[232]細看落墨皆松瘦，想見掀髯正鶴孤。[233] < 蔚州 > 土貢：熊鞹、豹尾、松實。[234] 罪出其身，不使廢松檟之奉。[235] 爛文章之糾纏，驚節解而流膏……收薄用於桑榆，製中山之松醪。[236]撥置千憂並百慮，且醉一斛松醪春。[237] […]

Aus den so entstandenen 53 Subkorpora lassen sich nun grobe temporale Sprachmodelle berechnen. Dabei kann ein Zeitraum von 700 v. u. Z. bis zum 20. Jh. abgedeckt werden.[238]

5.7 *HYDCD-Data Science*: Erkenntnisse aus der Datenbank

> *„But we are in greater darkness*
> *if we go still further back […]“*[239]
>
> Mario Alinei

Die in Kapitel 5.5 erzeugte Lexemdatenbank erlaubt einige Einblicke in die Machart des *HYDCD* bzw. des *DHYDCD*, sowie in die Entwicklung des chinesischen Wortschatzes.[240] Ein tiefergehendes Verständnis der erzeugten Daten ist zudem für die Entwicklung von Datierungs- bzw. Fälschungserkennungssoftware auf dieser Basis nützlich.

228 Siehe Abschnitt 5.5.3, S. 132
229 Siehe auch Abschnitt 5.7.1, ab S. 139.
230 Vgl. auch Hoffmann 2004, S. 17, siehe auch S. 24.
231 Siehe Kapitel 5.5.1, S. 123. Um den Anteil bei der Extraktion entstandener *false positives* im Korpus zu minimieren, werden nur Primärquellen mit zwei oder mehr *attestations* in Betracht gezogen.
232 *DHYDCD*, 松煙墨. Belegstelle aus dem *Meng xi bi tan* 夢溪筆談 von Shen Kuo 沈括 (1031–1095).
233 *DHYDCD*, 松瘦. Aus *Ciyun* Liu *Jingwen jian ji* 次韻劉景文見寄 von Su Shi 蘇軾 (1031–1101).
234 *DHYDCD*, 松實. Aus dem 1060 fertiggestellten *Xin Tangshu* 新唐書.
235 *DHYDCD*, 松檟. Aus *Xie zhe shou Xiuzhou tuanlian fushi biao* 謝謫授秀州團練副使表 von Shen Kuo.
236 *DHYDCD*, 松醪. Aus *Zhong shan song lao fu* 中山松醪賦 von Su Shi.
237 *DHYDCD*, 松醪春. Aus *Wang Baishuishan ci Hejiang lou yun* 望白水山次合江樓韻 von Li Gang 李綱 (1083–1140). Durch die Art der Datengewinnung aus der *CBDB* wird das Zitat ausgewertet, obwohl Li Gang das Gedicht vermutlich erst nach 1100, also nach *chronon*-Ende verfasst hat, da für diese Quelle nur die biographischen Daten des Autors vorliegen. Erläuterungen dazu siehe Kapitel 5.5.3, S. 132.
238 Siehe dazu Kapitel 6.1.3, ab S. 171.
240 Für eine sprach- und kulturhistorische Betrachtung der hier vorgestellten Daten siehe auch Tilman Schalmey 2020: „Das *Hanyu Da Cidian* 漢語大詞典 als Sprachgedächtnis“. In: *Erinnern und Erinnerung, Gedächtnis und Gedenken*. Hrsg. von Maria Khayutina und Sebastian Eicher. Jahrbuch der Deutschen Vereinigung für Chinastudien. Wiesbaden: Harrassowitz, S. 73–90, *passim*.

5.7.1 Genauigkeit der gewonnenen Daten

Bedingt durch die Zitierweise im *DHYDCD*[241] und die teilweise sehr ungenaue Datierbarkeit älterer Primärquellen[242] können einige Lexeme keinem genauen Jahr zugeordnet werden, sondern unterschiedlich langen Zeiträumen. Abb. 5.5 stellt die Genauigkeit der Datierung aller so eingeordneten Lexeme dar. In Abb. 5.5a wird der Datenstand ohne zusätzliche Belegstellen gezeigt, in 5.5b sind diese berücksichtigt. Die durchschnittliche Genauigkeit der Datierung $\bar{u}$ beträgt 86, bei Berücksichtigung der ergänzten Belege 76 Jahre. Die Darstellung zeigt auch, dass bereits ab der Han-Zeit der überwiegende Anteil der Primärquellen sehr genau bzw. mit einer Genauigkeit von weniger als 100 Jahren datiert werden kann. In Abb. 5.6a zeigt, wie viele Bedeutungen in den datierten Einträgen unterschieden werden. In knapp 80 % der Einträge wird nur eine Bedeutung angegeben und belegt. Mit zunehmender Anzahl an Bedeutungen nimmt dann die Anzahl entsprechender Einträge logarithmisch ab. In einem durchschnittlichen Eintrag werden 1,45 Bedeutungen oder Konnotationen unterschieden und belegt. Ein offensichtlicher Zusammenhang besteht zwischen Äquivokation[243] und der Zeichenlänge der Einträge. Längere, mehrsilibige Wörter bzw. Phrasen weisen tendenziell eine geringere Anzahl an Bedeutungen auf (Abb. 5.6b).

Für einzelne Zeichen können – im Extremfall des Eintrags zu *fa* 發 – bis zu 81 Bedeutungen unterschieden werden.[244] In solchen Fällen werden aber zahlreiche streitbare Konnotationen betrachtet, deren semantischer Unterschied aus den angegebenen Textbelegen oft nicht klar wird.[245] Davon abgesehen lassen die Anteile an mehrdeutigen Lexemen erahnen, dass die Unterscheidung von Wortbedeutungen (*word sense disambiguation*) einen Mehrwert für Datierungsaufgaben bringen würde,[246] gleichzeitig aber insbesondere für klassische Texte eine immense Herausforderung darstellt.[247]

Im Kontext der Mehrdeutigkeit chinesischer Zeichen sei nochmals auf *duoyinzi* 多音字, Zeichen mit unterschiedlichen Aussprachen, eingegangen. Von den 16.361 graphisch unterschiedlichen Schriftzeichen, denen im *DHYDCD* Einträge gewidmet sind, ist für etwas mehr als 80 % nur eine einzige Lesung angegeben, für die restlichen sind zwei oder drei, in einzelnen Fällen sogar bis zu sieben Lesungen bekannt (Abb. 5.7).[248] Fast immer werden unterschiedliche Lesungen mit abweichenden Bedeutungen, häufig auch mit anderen grammatikalischen Kategorien in Verbindung gebracht.

---

240 Alinei 2004, S. 215.

241 Siehe Abschnitt 5.5.2, ab S. 127.

242 Siehe Loewe 1993, S. xi.

243 Da hier nur die graphische Gestalt der Lexeme und die Anzahl der angegebenen Erklärungen untersucht werden kann, ist eine Unterscheidung zwischen Homographie (bei unterschiedlicher Aussprache), Homophonie (unterschiedliche Bedeutung bei gleicher Aussprache) und Polysemie (unterschiedliche, verwandte Bedeutungen) hier nicht abbildbar.

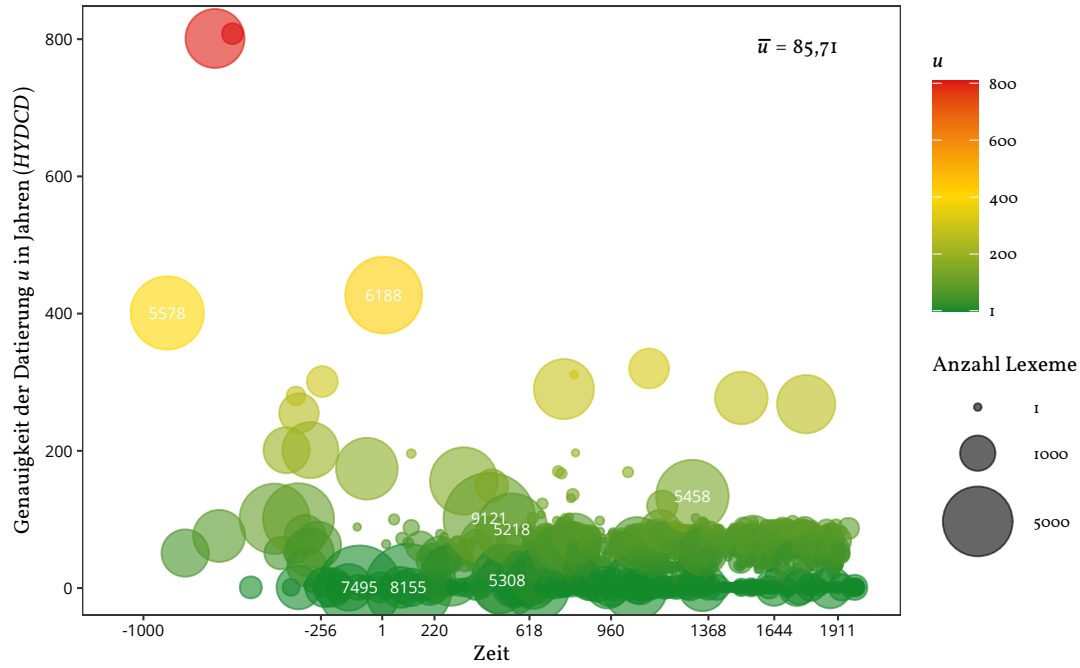
244 Siehe *DHYDCD*, *fa* 發.

245 Im Eintrag zu *zhi* 之 etwa wird zwischen mehreren Fällen unterschieden, in denen *zhi* stets als subordinierende Partikel zwischen zwei Satzgliedern funktioniert, der grammatikalische Unterschied wirkt eher konstruiert. Ähnliches gilt für die Verwendung als Pronomen. Siehe *DHYDCD*, *zhi* 之.

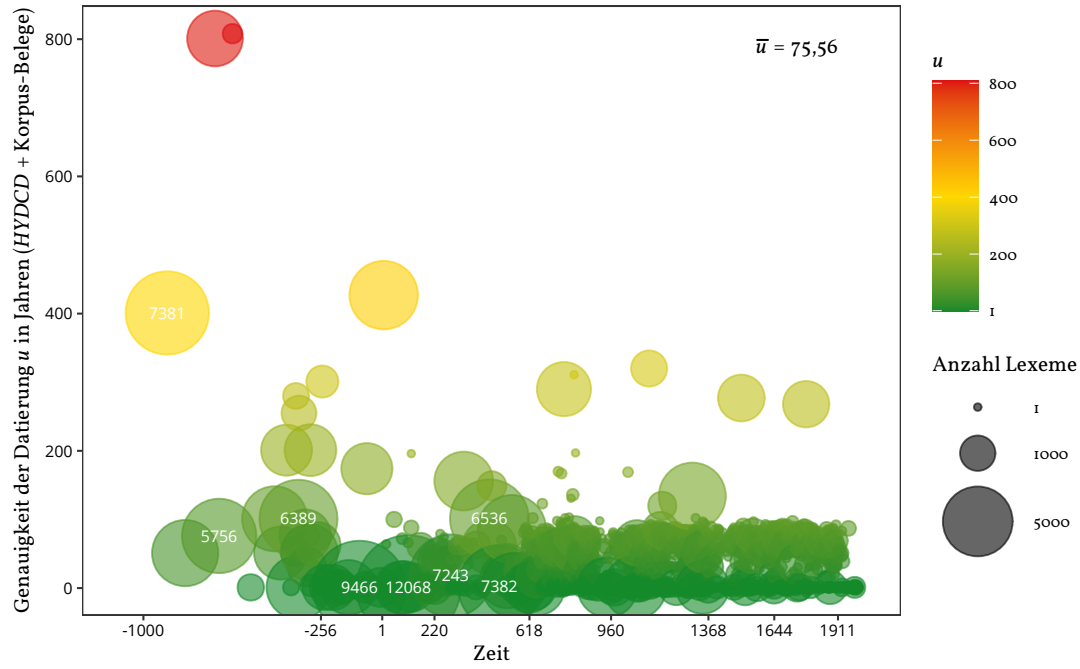
246 Vgl. Kanhabua und Nørvåg 2008, S. 361.

247 Dass eine automatisierte, kontextuelle Erkennung unterschiedlicher Wortbedeutungen und deren Veränderung grundsätzlich möglich ist, wird in Tahmasebi, Borin und Jatowt 2019, erörtert. Siehe S. 1–2. Eine darin vorgestellte Studie über semantische Veränderungen chinesischer Wörter ist Tang Xuri, Qu Weiguang und Chen Xiaohe 2015: „Semantic Change Compuation: A Successive Approach". In: *World Wide Web* 19.3, S. 375–415. doi: 10.1007/s11280-014-0316-y, Die Implementierung vergleichbarer Techniken würde den Rahmen dieser Dissertation sprengen. Für *n*-Gramm-Daten sind sie zudem nicht anwendbar.

248 Für das Zeichen 繆 werden folgende Lesungen gegeben: *móu*, *jiū*, *miù*, *mù*, *miào*, *liáo* und *lù*. Siehe *DHYDCD*, 繆/繆 1 – 繆 7.

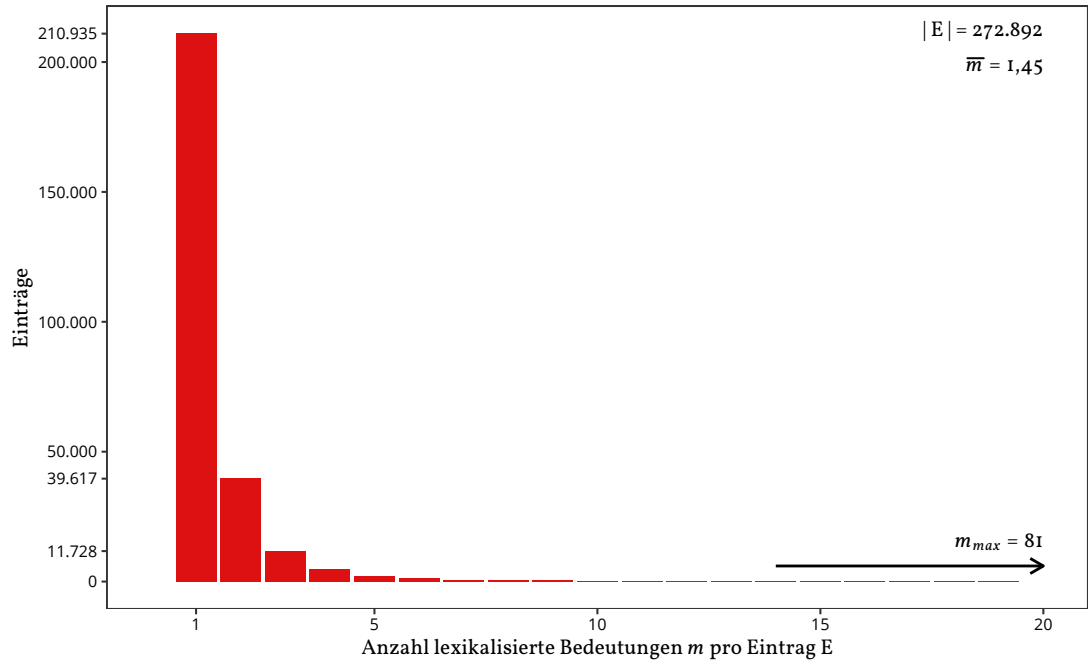


(a) *HYDCD*

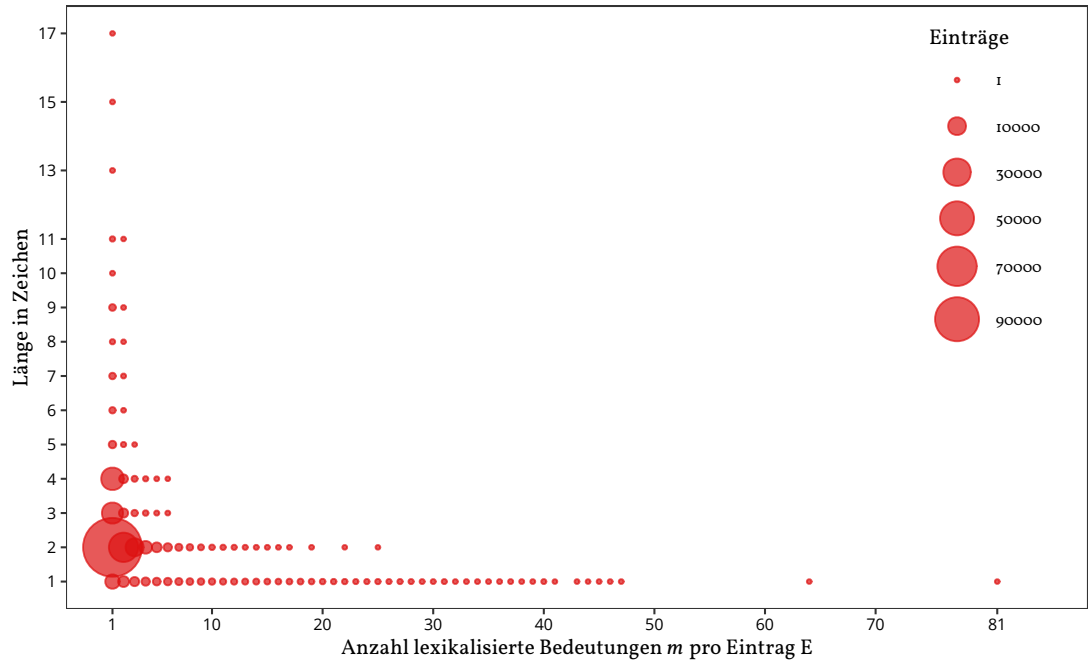


(b) + Korpora-Belege aus Loewe, *zhengshi* 正史

Abbildung 5.5 Genauigkeit der Lexemdatierung



(a) Anzahl angegebener Bedeutungen *vs.* Anzahl Einträge



(b) Anzahl angegebener Bedeutungen *vs.* Länge in Zeichen

Abbildung 5.6 „Unterschiedliche“ Bedeutungen in *HYDCD*-Einträgen

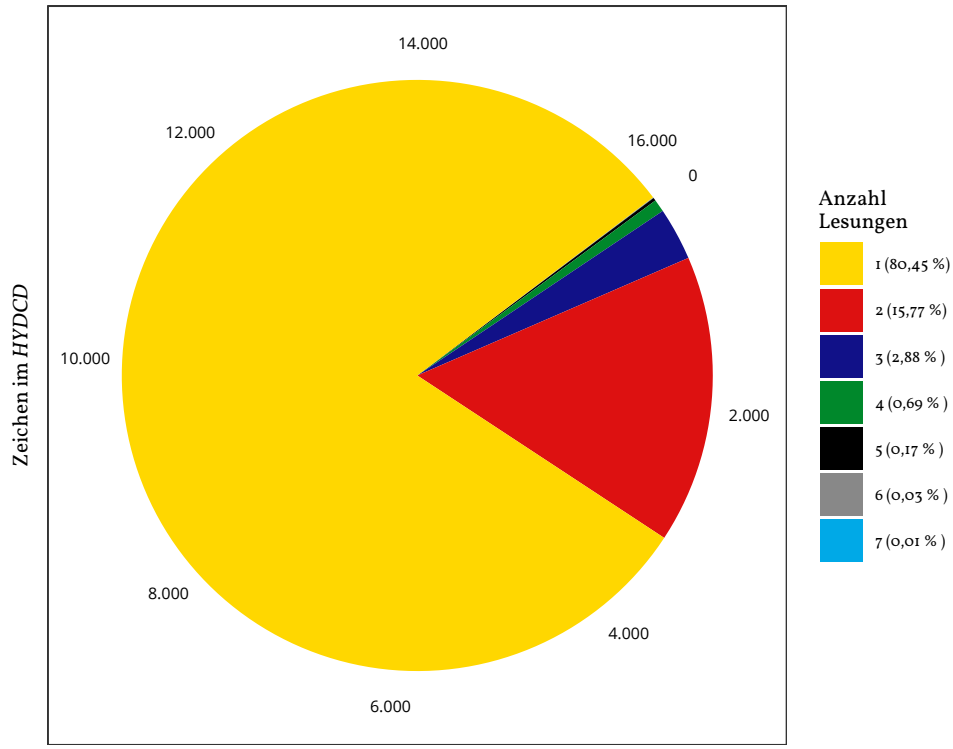


Abbildung 5.7 Lexikalisierte Zeichen im *DHYDCD* nach Anzahl ihrer Lesungen

5.7.2 Lexikalisierung pro Jahrhundert

Betrachtet man anhand der jeweils frühesten Belegstellen der *DHYDCD*-Lexeme aus der Tabelle `the_words` die Anzahl neuer Lexikalisierungen pro Jahrhundert auf einer Zeitachse, wird die chronologische Abdeckung und Gewichtung der Daten sichtbar (Abb. 5.8).[249] Sie lässt unterschiedliche Rückschlüsse und Interpretationen zu.

— 1. Die vor dem **4. Jh. v. u. Z.** verhältnismäßig dünne Datenlage zeigt sich klar. Sie ist darauf zurückzuführen, dass aus den ersten Jahrhunderten des Betrachtungszeitraums überhaupt nur wenige (längere) Texte überliefert sind. Zudem ist die Datierung prä-hanzeitlicher Texte meist ungenau, so dass teils auf grobe Schätzungen zurückgegriffen werden muss.[250] Zudem wurde ein Großteil der bis heute überlieferten frühen Texte während der Han-Zeit durch Liu Xiang 劉向 (77–6 v. u. Z.) und seine Mitarbeiter redigiert und standardisiert, so dass sie uns gewissermaßen gefiltert vorliegen.[251]

— 2. Im *DHYDCD* stehen für Quellen aus dem **20. Jh.** meist weniger Metadaten zur Verfügung als sonst.[252] Dadurch fällt die Lexikalisierung auch hier übertrieben gering aus, obwohl gerade

249 Lexeme, die wg. einer zu ungenauen Datierung der Belegstelle nicht eindeutig einem Jahrhundert zugeordnet werden können, werden gemäß ihrer Datierung anteilig zugeordnet. Vorgehensweise und Berechnung dafür werden in Kapitel 6.2.1, ab S. 184 beschrieben.
250 Siehe auch die Angaben in der Tabelle der häufig zitierten Texte in Abschnitt 5.7.4, S. 150.
251 Siehe z. B. Kern 2004, S. 46.
252 Siehe Abschnitt 5.5.2, S. 129.

im 20. Jh. durch den viel intensiveren Kontakt mit dem Westen,[253] den technischen Fortschritt und den Erfolg der geschriebenen Umgangssprache (*baihuawen* 白話文)[254] zahllose Neologismen entstanden sind.

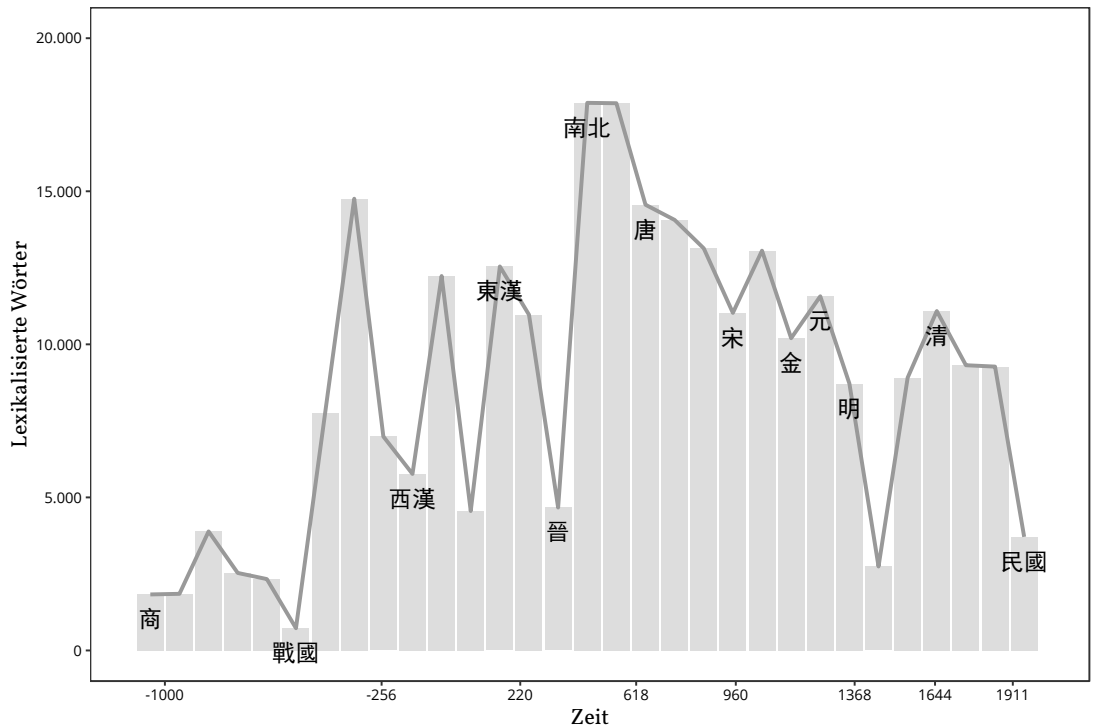


Abbildung 5.8 Lexikalisierung im *HYDCD* nach Jahrhundert. Die Balken stehen für die Anzahl der Lexeme, deren älteste Belegstelle (etwa) aus dem jeweiligen Jahrhundert stammt.

— 3. Auf den ersten Blick scheint die restliche Lexikalisierung (und damit die Entstehung) neuer Wörter einer zyklischen bzw. dynastischen **Schwankung** zu unterliegen, angedeutet durch die Zickzackkurve in Abb. 5.8. Dies ist tatsächlich plausibel, da auch für andere Sprachen größere Schwankungen im Wortschatz im Zusammenhang mit wichtigen historischen Ereignissen in Verbindung gebracht werden konnten.[255] Dazu passt besonders auch der Abwärtsknick in der Lexikalisierung während der gegenüber Einflüssen von außen als besonders verschlossen geltenden Ming 明-Dynastie.[256] Eder mahnt jedoch zurecht, dass „jeder Versuch, direkte Zusammenhänge zwischen historischen Ereignissen und stilistischen Veränderungen

253 Siehe z. B. Lackner, Amelung und Kurtz 2001, S. 2.

254 Vgl. z. B. Elisabeth Kaske 2007: *The Politics of Language in Chinese Education, 1895–1919*. Leiden: Brill, v. a. S. 30–31.

255 Siehe Mikhail V. Arapov 1983: „Word Replacement Rates for Standard Russian (A.D. 1100–1850)". In: *Historical Linguistics*. Hrsg. von Barron Brainerd. Quantitative Linguistics 18. Bochum: Dr. N. Brockmeyer, S. 50–61, S. 60; siehe z. B. auch Eder 2018, S. 364: „Our study corroborated the hypothesis that epochs of substantial stylistic drift are followed by periods of stagnation, rather than forming purely linear trends." ; vgl. auch Bochkarev, Solovyev und Wichmann 2014, S. 4.

256 Siehe z. B. Richard von Glahn 1996: *Fountain of Fortune: Money and Monetary Policy in China, 1000–1700*. Berkeley: University of California Press, S. 90. Hinter der konservativen Isolationspolitik der Ming standen allerdings vor allem wirtschaftspolitische Beweggründe.

zu finden, menschlichen Vorurteilen unterliegt."[257] Dass lexikalischer Wandel durch Krisen beschleunigt werden kann, bestätigt sich aber auch an der Anzahl der Neologismen, die vom Leibniz-Institut für Deutsche Sprache (IDS) für das Deutsche im Kontext der „Coronakrise" aufgezeichnet wurden.[258] Allerdings bleibt abzuwarten, welche und vor allem wie viele der 2.433 gesammelten Wortschöpfungen (Stand: November 2022) mittelfristig im Sprachgebrauch erhalten bleiben bzw. lexikalisiert werden.[259]

— 4. Auffällig ist zudem die besonders hohe Lexikalisierung im **5. und 6. Jh.**, die andeuten könnte, dass zu dieser Zeit besonders viele neue Wörter Eingang in die chinesische Sprache fanden. Ein plausibler historischer Grund hierfür wäre die verstärkte Verbreitung des Buddhismus in China.[260] Durch die Übersetzung von Sutren gelangen damit zahlreiche Sanskrit-Begriffe in die chinesische Sprache.[261] Ein Abgleich der gefundenen Lexeme mit dem buddhistischen chinesischen Wörterbuch von William E. Soothill und Lewis Hodous[262] zeigt jedoch, dass dies nur einen verhältnismäßig kleinen Beitrag leistet, während die Hauptursache in der hohen Gewichtung des *Hou Han shu* 後漢書 als Primärquelle gesehen werden kann.[263] Vor dem Hintergrund, dass ein Kriterium der Herausgeber des *HYDCD* war, dass die Aufnahme von Fachvokabular beschränkt sein sollte auf Begriffe, die zum allgemeingebräuchlichen Wortschatz gezählt werden können,[264] ist die im Verhältnis geringe Lexikalisierung buddhistischer Termini wenig überraschend.

— 5. Die Auswahl der Belegstellen wurde durch Neigungen bzw. Präferenzen der Herausgeber:innen und auch durch das zur Verfügung stehende Material beeinflusst, so dass gut erschlossene Texte unverhältnismäßig häufig zitiert werden. Ein Vergleich mit dem *OED*, für das das Vorhandensein von Konkordanzen offensichtlich die Auswahl der Belegstellen beeinflusst hat, legt dies ebenfalls nahe.[265] Ein dadurch entstehendes ***Bias*** bedingt, dass die Lexikalisierung der betrachteten Jahrhunderte unterschiedlich gut dokumentiert ist.

Berücksichtigt man die zusätzlichen, früheren Belegstellen aus *zhengshi* 正史 und Loewe-Korpora[266] (Abb. 5.9) lässt sich eine teilweise Verschiebung nach links beobachten, am grundsätzlichen Verlauf des Balkendiagramms ändert sich aber kaum etwas. Das 4. Jh. v. u. Z. weist nunmehr die höchste Konzentration von *Loci classici* auf, was klar auf die Gewichtung des verwendeten Loewe-Korpus zurückzuführen ist, aus dem ein hoher Anteil der ergänzten Belegstellen stammt.[267] Ließe man die zyklischen bzw. zufälligen Schwankungen in der

257 Eder 2018, S. 363, übersetzt durch den Verfasser. Eder untersucht Sprachwandel im Englischen mithilfe des *Google n-Gram Service* und entdeckt entsprechende *peaks* im Kontext des amerikanischen Bürgerkriegs in den 1870er Jahren und der *Great Depression* in den 1920er Jahren.

258 Leibniz-Institut für Deutsche Sprache (IDS), Hrsg. 2020: *Neuer Wortschatz rund um die Coronapandemie*. Online-Neologismenwörterbuch OWID. Mannheim. URL: https://www.owid.de/docs/neo/listen/corona.jsp (besucht am 09. 11. 2022).

259 Vgl. auch Jing-Schmidt und Hsieh 2019, S. 523: „[T]he majority of new words in fact fail to become established in language."

260 Siehe z. B. Vogelsang 2012, S. 219.

261 Wang Li 王力 2011 [1958], S. 590–591. Siehe auch Kapitel 2.2, ab S. 16.

262 William E. Soothill und Lewis Hodous 2003 [1937]: *A Dictionary of Chinese Buddhist Terms*. Online Version. URL: http://mahajana.net/texts/soothill-hodous.html (besucht am 28. 11. 2017).

263 Siehe Abschnitt 5.7.4, ab S. 150; ausführlicher dazu siehe auch T. Schalmey 2020, S. 79 u. S. 84–85.

264 „[…]对专科词的收录以进入一般语词范围的为限[…]" Yu Zhangrui 余章瑞 1988.

265 Siehe Kapitel 5.2, 111.

266 Siehe dazu auch Abschnitt 5.5.4, S. 134.

267 Siehe dazu Kapitel 4.2, ab S. 66.

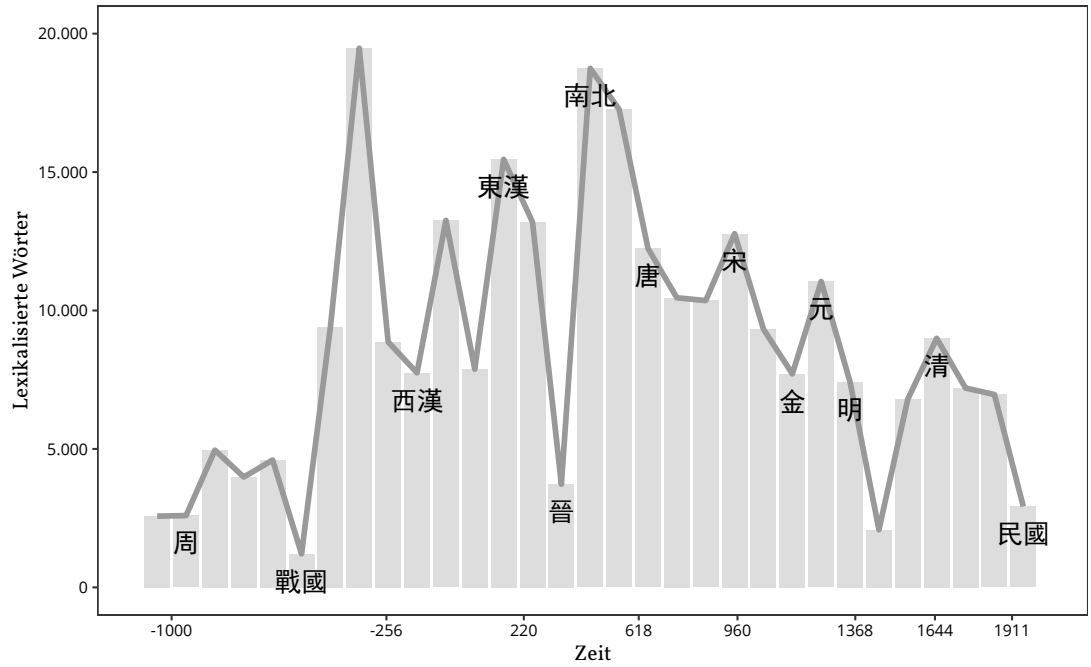


Abbildung 5.9 Lexikalisierung mit zusätzlichen Korpusbelegstellen aus Abschnitt 5.5.4

Lexikalisierung außer Acht, so kann insgesamt (abgesehen von den Randbereichen) eine gleichbleibend hohe Lexikalisierung mit durchschnittlich 8.727 neuen Einträgen pro Jahrhundert im *DHYDCD* beobachtet werden.[268]

Dass gerade zu Beginn und zum Ende des Betrachtungszeitraums eine geringere Lexikalisierung zu beobachten ist, legt einen Blick auf das kumulative Wachstum des Wortschatzes nahe. Besitzt das Piotrowski-Gesetz[269] auch einen Erklärungsgehalt für das Wortschatzwachstum im Chinesischen? Betrachten wir anhand der über die Belegstellen im *DHYDCD* datierbaren Lexikalisierung das kumulative Wortschatzwachstum pro Jahrhundert, so folgt es tatsächlich einer *s*-förmigen Kurve (Abb. 5.10).[270] Vom 5. Jh. v. u. Z. bis zum Ende des 3. Jhs., sowie vom 14. bis 20. Jh. sind zudem kürzere *s*-Kurven innerhalb des Gesamtverlaufs erkennnbar.[271] Auch wenn entsprechende Gesetzmäßigkeiten Spekulation bleiben müssen, spricht vieles dafür, von einem natürlichen, logistischen Wortschatzwachstum auszugehen, welches durch einschneidende historische Ereignisse beeinflusst werden kann.

268 Berücksichtigt werden dabei nur die insgesamt 270.525 Einträge in `the_words`, die sich chronologisch einordnen lassen. Einträge ohne Belege, bzw. mit Belegen ohne ausreichende bibliographische Daten, können nicht gezählt werden. Die *tatsächliche* Lexikalisierung würde bedeutend höher ausfallen. Auch die unvermeidbare Unvollständigkeit des Wörterbuchs sollte nicht vergessen werden.

269 Siehe Kapitel 2.1, ab S. 14.

270 Die Kurve der idealisierten *s*-förmigen Lexikalisierung in Abb. 5.10 wird in *R* mithilfe der Funktion `drm` (*Dose-Response Model*) geschätzt. Sie ist Teil des Pakets `drc`, das sich primär an Epidemiolog:innen richtet. Siehe Christian Ritz 2016: *drc Analysis of Dose-Response Curves, Version 3.0-1. R* package. URL: `rdocumentation.org/packages/drc/versions/3.0-1` (besucht am 10. 02. 2021).

271 Vgl. auch Aitchison 2001 [1991], S. 92: „A closer look at each *S*-curve, however, suggests that many *S*-curves are themselves composed of smaller *S*-curves. Each little *S*-curve covers one particular linguistic environment."

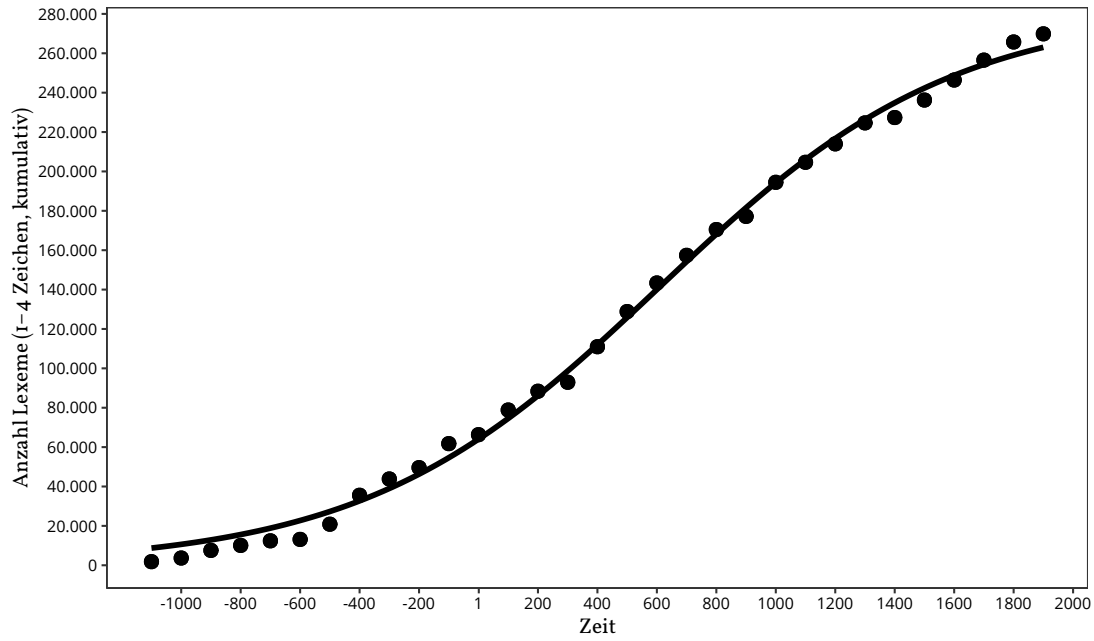


Abbildung 5.10 Lexikalisierung im *DHYDCD* nach Jahrhundert (ohne zusätzliche Belegstellen)

5.7.3 Mono- und Polysyllabizität

Ein Großteil der heute verwendeten Schriftzeichen stand bereits spätestens während der Han 漢-Zeit zur Verfügung. Große Zeichenwörterbücher wie das *Hanyu da zidian* 漢語大字典[272] beweisen zwar eindrucksvoll, dass auch in den folgenden knapp zwanzig Jahrhunderten kontinuierlich neue Zeichen entstanden sind. von denen sich allerdings nur wenige durchsetzen konnten. Die Rigidität der chinesischen Schrift bzw. der tatsächlichen Zeichennutzung zeigt sich anhand der Belege im *DHYDCD*. Die Lexikalisierung neuer Schriftzeichen (Abb. 5.11) nimmt im zeitlichen Verlauf tendenziell klar ab.[273]
Diese Beobachtung steht scheinbar im Widerspruch zu der von Best und Zhu Jinyang beobachteten „Zunahme der Schriftzeichen“[274] – es bleibt dabei allerdings anzumerken, dass im *DHYDCD* kaum historische oder lokale Zeichenvarianten aufgeführt sind.

Auf der anderen Seite ist als Trend erkennbar, dass die Lexikalisierung 3- und 4-silbiger Wörter, die in frühen Texten noch wenig belegt sind, im Laufe des Betrachtungszeitraums insgesamt kontinuierlich zunimmt (Abb. 5.12). Dabei überwiegen zunächst 4-silbige Lexeme klar – im Laufe der Jahrhunderte wird diese Verteilung aber zunehmend gleichmäßiger. Trotz der scheinbaren Prävalenz quadrisyllabischer Ausdrücke wie *chengyu* 成語 in der chinesischen Sprache, werden also ca. ab dem 7. Jh. ungefähr gleich viele trisyllabische Wörter lexikalisiert.[275]

272 *HYDZD*.
273 Ausführlicher dazu siehe auch T. Schalmey 2020, S. 80–82.
274 Best und Zhu Jinyang 2006, S. 208.
275 Siehe dazu auch T. Schalmey 2020, S. 81–82. Über die tatsächliche Häufigkeit dieser Lexeme in Texten kann hier natürlich keine Aussage getroffen werden.

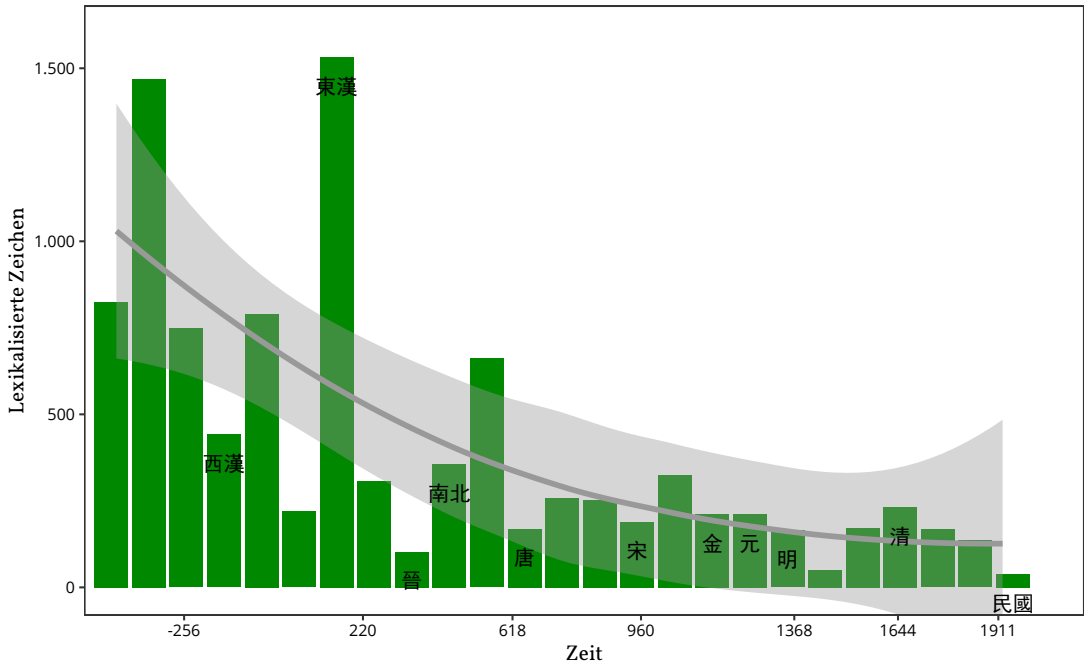


Abbildung 5.11 Lexikalisierung neuer Schriftzeichen im *DHYDCD*

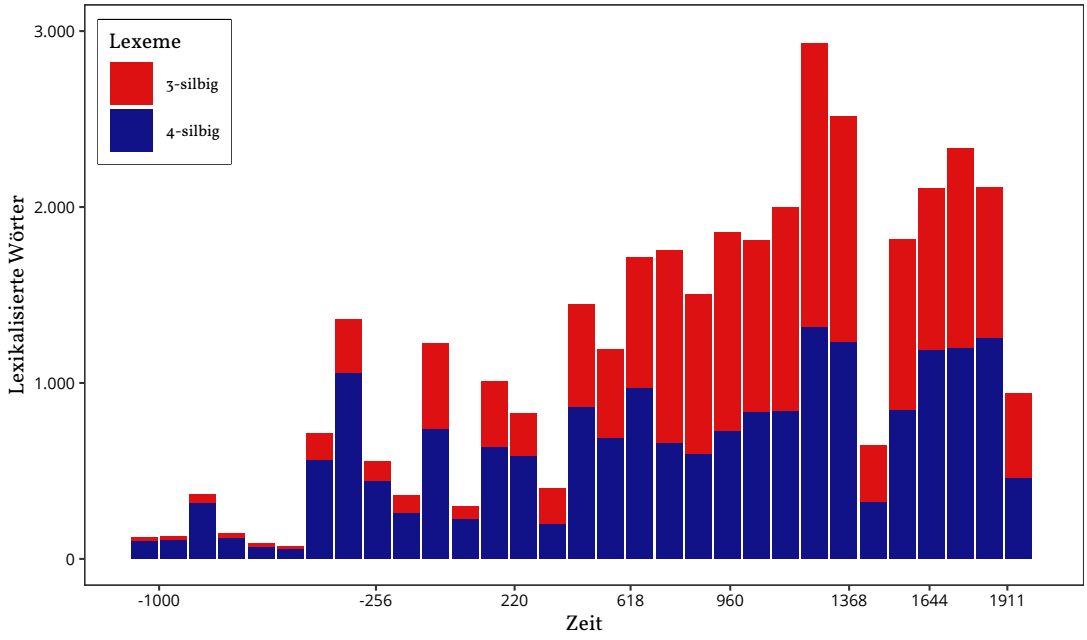


Abbildung 5.12 Lexikalisierung 3- und 4-silbiger Wörter

Die insgesamt klar zunehmende Entstehung mehrsilbiger Wörter mag den steigenden Bedarf daran widerspiegeln, komplexere oder zahlreichere Konzepte sprachlich eindeutiger auszudrücken. Bei langfristiger diachroner Betrachtung spricht sie auch dagegen, dass das allgemein für Sprachwandel typische „crunching“,[276] das Kompakter-werden sprachlicher Ausdrücke für das Chinesische uneingeschränkt zutrifft.[277] Auch im Hinblick auf Übertragungen schriftsprachlicher (*wenyan* 文言) oder gar klassischer (*guwen* 古文) Texte in die schriftliche Form der modernen Umgangssprache *baihua wen* 白話文 lässt sich diese Beobachtung für die chinesische Sprache nicht allgemein bestätigen.

Gleichzeitig stellt eine Länge von vier Zeichen ein typisches Maximum dar. Zwar sind im *DHYDCD* vereinzelt bis zu 17-silbige Ausdrücke lexikalisiert, doch bereits der Anteil an 5-silbigen ist verschwindend gering (Abb. 5.13).[278]

Der zeitliche Verlauf der Aufnahme neuer disyllabischer Lexeme entspricht etwa dem der gesamten Lexikalisierung.[279] Werden die Gesamtanteile *aller* datierbaren *DHYDCD*-Lexeme nach Anzahl der Silben betrachtet, ist das wenig überraschend: über 80 % aller Lexeme sind disyllabisch (Abb. 5.13). Der hohe Anteil erklärt sich durch eine starke Präferenz für Zusammensetzungen.[280] Typische Beispiele umfassen Wortbildungen aus zwei bedeutungsgleichen oder -ähnlichen Morphemen wie *bao+hu* 保護 („beschützen“), *xiao+shou* 銷售 („verkaufen“) oder *gou+mai* 購買 („kaufen“), Abkürzungen wie *Beida* 北大 („Uni Peking“), sowie einsilbige Ortsbezeichnungen, die eine Kategorieangabe erfordern, wie *faguo* 法國 („*Fa*-Land“, Frankreich).[281]

Für die moderne Hochsprache beobachtet Breiter auf Basis des *Xiandai Hanyu pinlü cidian* 現代漢語頻率詞典 *(Häufigkeitswörterbuch der modernen chinesischen Sprache)* eine ähnliche Verteilung, die sich lediglich durch einen deutlich höheren Anteil an monosyllabischen Lexemen unterscheidet (1-silbig 11,90 %, 2-silbig, 73 %, 3-silbig ca. 8,7 %, 4-silbig ca. 6,5 %, längere Lexeme weniger als 1 %). Untersucht man anstatt des Vorhandenseins von Lexemen die Wortlängenverteilung in *Texten*, bzw. in verschiedenen Textgattungen, ergibt sich ein anderes Bild. Sogar im modernen *putonghua* 普通話 dominieren einsilbige Wörter mit 64,3 % des von Breiter untersuchten Korpus.[282] Während in literarischen Texten einsilbige *tokens* den größten Anteil ausmachen, dominieren in juristischen, wissenschaftlichen- und Zeitungstexten zweisilbige Wörter.[283] Auch

276 Aitchison 2001 [1991], S. 116.

277 Vgl. Reinhard Köhler 1986: *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Quantitative linguistics 31. Bochum: Dr. N. Brockmeyer, S. 75–78. Köhler sieht die „Minimierung des Produktionsaufwandes“ als „Systembedürfnis“, welches Sprachwandel bedingt und sich auf die Länge und Komplexität sprachlicher Ausdrücke auswirkt. Ronald Langacker bezeichnet Sprachen sogar als „gigantic expression-compacting machines“, die sprachliche Ausdrücke typischerweise im Laufe der Jahrhunderte kompakter werden lassen. Siehe Ronald W. Langacker 1977: „Syntactic reanalysis“. In: *Mechanisms of Syntactic change*. Hrsg. von Charles N. Li. Austin: University of Texas Press, S. 57–139, S. 106; zitiert in Aitchison 2001 [1991], S. 116; Andererseits vermuten Arapov und Cherc einen Zusammenhang zwischen Silbenlänge und Alter von Wörtern dahingehend, dass neuere Wörter häufiger eine höhere Anzahl an Silben aufweisen. Siehe Mikhail V. Arapov und Maja M. Cherc 1983 [1974]: *Mathematische Methoden in der historischen Linguistik [Matematičeskiye metody v istoričeskoy lingvistike, Математические методы в исторической лингвистике]*. Übers. von Reinhard Köhler und Peter Schmidt. Quantitative Linguistics 17. Bochum [Moskau]: Dr. N. Brockmeyer [Nauka], S. 49–50.

278 Siehe auch Kapitel 4.5.2, S. 92.

279 Siehe Abb. 5.8, S. 143.

280 Ausführlicher dazu siehe Wong Kam-Fai 黃錦輝 et al. 2010, S. 11–18.

281 Siehe z. B. Lü Shuxiang 呂叔湘 1963: „Xiandai Hanyu danshuang yinjie wenti chutan 现代汉语单双音节问题初探 (Vorläufige Studie zum Problem von Mono- und Disyllabizität im modernen Chinesischen)“. In: *Zhongguo yuwen* 中国语文 1, S. 10–22; zitiert in Wong Kam-Fai 黃錦輝 et al. 2010, S. 10.

282 Siehe Breiter 1994, 224ff. zitiert in Schindelin 2005a, S. 959.

283 Siehe Zhu Jinyang und Karl-Heinz Best 1992: „Zum Wort im modernen Chinesisch“. In: *Oriens extremus* 35, S. 45–60, S. 52f. zitiert in Schindelin 2005a, S. 960.

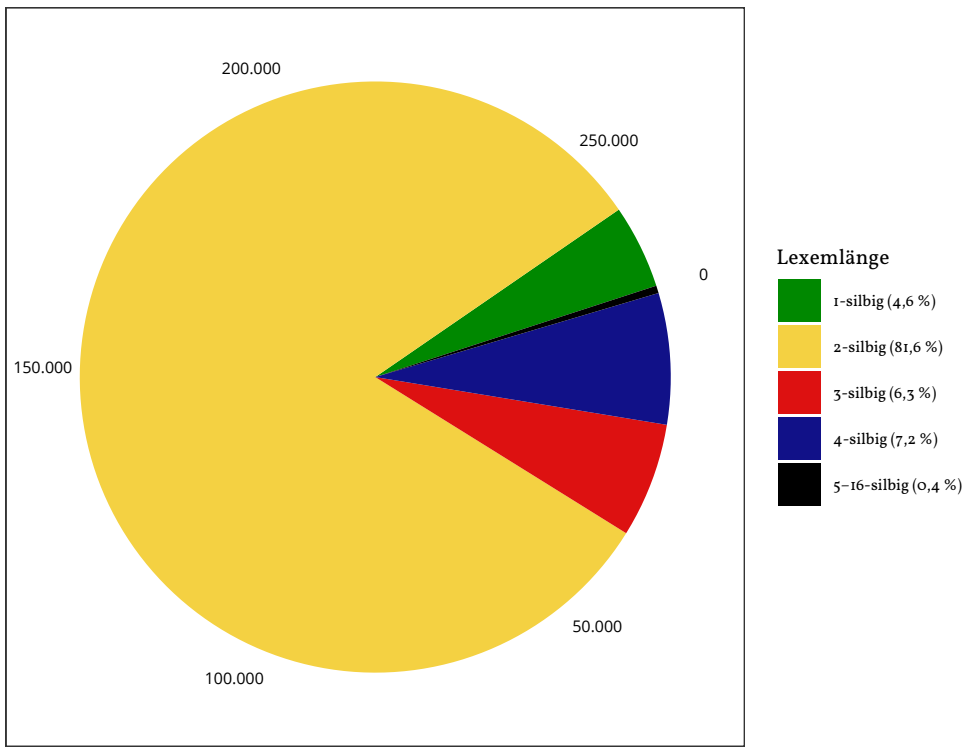


Abbildung 5.13 Chronologisierbare Lexikalisierung im *DHYDCD* nach Länge der Lexeme

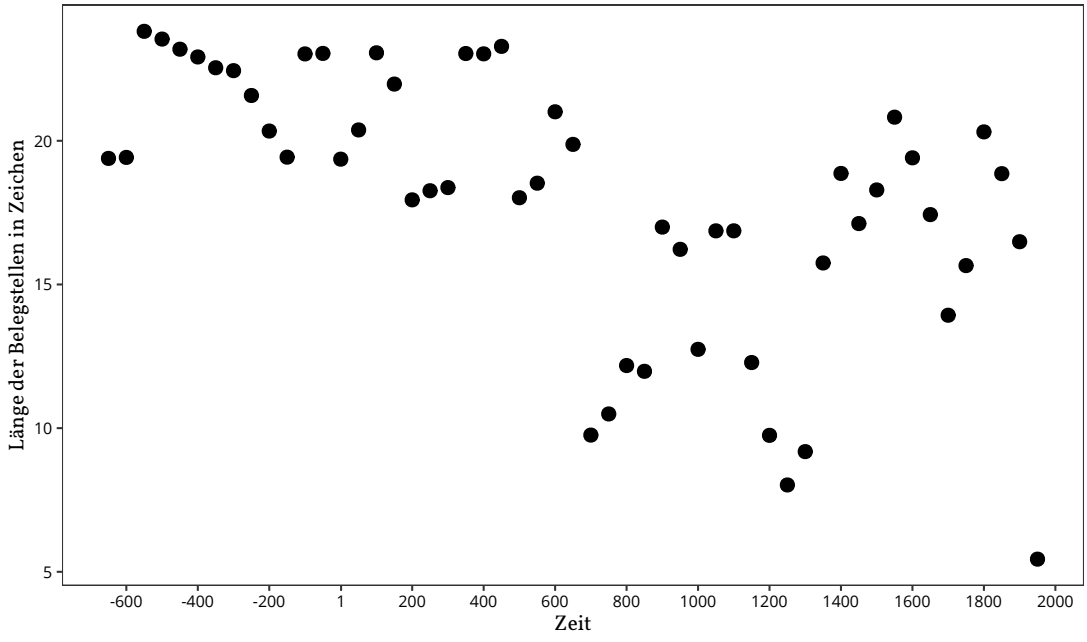


Abbildung 5.14 Länge der Belegstellen im *DHYDCD* nach Jahrhundert (in *zi* 字, inkl. Interpunktion)

über eine diachrone Analyse von Texten aus den vergangenen 2.500 Jahren lässt sich die zunehmende durchschnittliche Wortlänge statistisch nachweisen.[284]

Ungeachtet der gestiegenen Länge lexikalisierter und verwendeter Wörter, nimmt die Länge der Belegstellen im *DHYDCD* bei diachroner Betrachtung nicht zu. Abgesehen von wenigen Ausreißern schwankt sie um ein konstantes Niveau von knapp 20 Zeichen und scheint insgesamt sogar eher abzunehmen (Abb. 5.14). Hoffmann zeigt bei einer Untersuchung der *attestations* aus dem 11.–20. Jh. im englischsprachigen *OED* hingegen durchaus eine Zunahme der Länge der Textbelegstellen.[285] Bei den vorliegenden Daten scheint die Ableitung langfristiger, sprachgeschichtlicher Trends in Bezug auf die Satzlänge zu spekulativ, auch da die Belegstellen gekürzt sein können.

5.7.4 Lexikalisierung nach *Locus classicus*

Auffällig an den *Locus classicus*-Angaben im *DHYDCD* ist, dass einige Werke unverhältnismäßig oft genannt werden. In Tabelle 5.3 sind die 30 am häufigsten als früheste Belegstelle angegebenen aufgelistet. Zusammen werden sie für etwas mehr als 100.000 Lexeme (ca. 30,7 % der belegten Einträge) herangezogen.[286] Dominant sind die frühesten Textzeugnisse und wichtige kanonische und philosophische Texte vertreten, deren Datierung ist aber leider „nicht immer mit der Genauigkeit möglich, die sich Leser vielleicht wünschen würden“[287] Besonders häufig werden auch *zhengshi* 正史 wie *Hou Han shu* 後漢書 (*HHS*), *Han shu* 漢書, *Shiji* 史記, *Jin shu* 晉書 usw. angeführt.

Tabelle 5.3 Die 30 häufigsten *Locus classicus*-Angaben im *DHYDCD*

#	Text (Konkordanz)	Datierung	*Locus classicus*-Angaben	Länge in 1.000 Zeichen
1	*Hou Han shu (H 41)* 後漢書	ca. 400–500	9.091	917,5
2	*Han shu (H 36)* 漢書	92	8.115	712,2
3	*Shiji (H 40)* 史記	-91	7.468	512,6
4	*Shijing (HS 9)* 詩經	ca. -1100–-700	5.547	30,5
5	*Wen xuan [H 26]* 文選	520–530	5.278	996,2
6	*Zuo zhuan (HS 11)* 左傳	ca. -500–-400	5.058	180,5
7	*Liji (H 27)* 禮記	ca. -400–-300	4.513	98
8	*Jin shu [H 32]* 晉書	646	4.351	1.167
9	*Xin Tang shu [H 32]* 新唐書	1060	4.303	1.800,8
10	*Zhou li (H 37)* 周禮	ca. -150–23	3.777	52,8
11	*Sanguo zhi (H 33)* 三國志	ca. 280–297	3.551	390,4
12	*Shangshu (TJ)* 尚書	ca. -1100–-300	3.410	25,7
13	*Zhuangzi (HS 20)* 莊子	ca. -400–-200	3.036	65,1
14	*Song shu* 宋書	492–493	2.698	811,1
15	*Guanzi* 管子	ca. -720–-645	2.546	57,5
16	*Hong lou meng* 紅樓夢	ca. 1730–1764	2.324	731,1

284 Siehe Best und Zhu Jinyang 2006, S. 209–211. Best und Zhu beobachten ein quasi lineares Wachstum der Wortlänge, von etwa 1,15 Zeichen vor- und während der Han 漢-Zeit, auf bis zu 1,8 im 20. Jh., stellen aber ebenfalls eine breite Streuung fest.

285 Siehe Hoffmann 2004, S. 25. Er betont aber, dass die Zitatlänge „proves to be fairly constant, particularly for the time between 1450 and the end of the 19th century“ und begründet eine starke Zunahme im 20. Jh. mit der Präferenz der Herausgeber:innen der 2. Ausgabe für mehr Kontext.

286 Hinter den *Pinyin*-Titeln sind vorhandene Konkordanzen bzw. Indexbände dazu angegeben (*H = Harvard-Yenching*, *HS = Harvard-Yenching Supplement*, *TJ = Shangshu tongjian*), siehe auch Fußnote 293. Datierungen sind im Wesentlichen den entsprechenden Artikeln aus Loewe 1993 bzw. Wilkinson 2000, S. 503–505, entnommen.

287 Loewe 1993, S. xi, übersetzt durch den Verfasser.

Tabelle 5.3 (Fortsetzung)

#	Text (Konkordanz)	Datierung	*Locus classicus*-Angaben	Länge in 1.000 Zeichen
17	*Chu ci* 楚辭	ca. -329–-278	2.154	29,8
18	*Song shi [H 34]* 宋史	1345	2.069	4.037
19	*Shui hu zhuan* 水滸傳	ca. 1320–1372	2.063	437,3
20	*Yijing (HS 10)* 易	ca. -850–-800	2.033	21,6
21	*Huainanzi* 淮南子	-139	1.969	130,8
22	*Xunzi (HS 22)* 荀子	ca. -300–-238	1.943	64,9
23	*Guoyu* 國語	ca. -500–-300	1.918	70,4
24	*Nan shi* 南史	ca. 643–659	1.748	676,2
25	*Wei shu* 魏書	ca. 551–554	1.716	999
26	*Han Feizi* 韓非子	ca. -350	1.703	108,9
27	*Baopuzi* 抱樸子	ca. 265–420	1.534	152,2
28	*Ernü yingxiong zhuan* 兒女英雄傳	1878	1.502	472,1
29	*Jiu Tang shu* 舊唐書	945	1.496	2.001,9
30	*Lun heng* 論衡	80	1.442	164,1

Legt man die Daten aus Tabelle 5.3 und Abb. 5.8 (S. 143) übereinander, wird der Einfluss der meistzitierten Texte noch deutlicher (Abb. 5.15, S. 152). Die grauen Balken stellen wie in Abb. 5.8 die Gesamtlexikalisierung jeweils eines Jahrhunderts dar. Die 30 am häufigsten zitierten Texte sind gemäß ihrer in Tabelle 5.3 angebenen Datierung (x) und der Häufigkeit der *Locus classicus*-Angabe darübergelegt.[288] Die Visualisierung veranschaulicht, wie Zitate aus einzelnen Texten für einen großen Teil der Lexikalisierung des jeweiligen Jahrhunderts „verantwortlich" sein können. Als Extrembeispiel sticht das *Hou Han shu* 後漢書 (*HHS*) heraus: Mit über 9.000 darin belegten Lexemen ist es nicht nur Primärquelle für einen Großteil der Lexikalisierung aus dem 5. Jh., es ist auch der am häufigsten im *DHYDCD* als älteste Belegstelle angeführte Text.

Zwar wurde das *HHS* in seiner heute erhaltenen Form von Fan Ye 范曄 (398–445) erst im 5. Jh. kompiliert, es entstand aber in erster Linie aus überlieferten Materialien der östlichen Han 東漢-Zeit (25–220) wie dem *Dongguan Hanji* 東觀漢記.[289] Die im *HHS* enthaltenen „neuen" Lexeme dürften also größtenteils spätestens Han-zeitlich sein und die Datierung der Lexikalisierung über das *HHS* kann damit als teilweise „verspätet" angesehen werden. Ähnliches gilt sicherlich auch für die anderen prominent vertretenen *zhengshi*-Texte, sowie das im 6. Jh. kompilierte *Wenxuan* 文選 (mehr als 5.000 Angaben), eine heterogene Zusammenstellung von Texten, die teilweise mehrere Jahrhunderte früher datieren.[290]

Hieran zeigt sich einmal mehr, dass sich aus dem *DHYDCD* zwar Belege extrahieren lassen, wann ein Wort *spätestens* sicher belegt ist, ein Teil der so datierten Lexeme aber durchaus bereits mehrere Jahrhunderte früher verwendet worden sein kann. Die Gewichtung einzelner Texte liefert zudem einen wichtigen Erklärungsansatz für die Schwankungen in der Neulexikalisierung pro Jahrhundert[291] – sie können in der Arbeitsweise der Herausgeber:innen des *HYDCD* bei der Auswahl der *attestations* begründet liegen. Vermutlich wurden wichtige, leicht zugängliche Texte

288 Die Datierung früher Texte wie *Shijing* 詩經 oder *Yijing* 易經 kann bestenfalls eine grobe Schätzung sein, die sich über mehrere Jahrhunderte erstreckt. Da die Texte als Mittelpunkte dieser Perioden dargestellt werden, kommt es hier vereinzelt zu einer höheren y-Platzierung als die der Gesamtlexikalisierung des jeweiligen. Jh. Die Zuordnung der dadurch ungenau datierbaren Lexeme erfolgt anteilig auf die Jahrhunderte der Schätzperiode (siehe dazu Kapitel 6.2.1, S. 184)

289 Zur Entstehungsgeschichte des *Hou Han shu* siehe Bielenstein 1954, S. 9–17.

290 Siehe auch T. Schalmey 2020, S. 79.

291 Vgl. Abschnitt 5.7.2, ab S. 142.

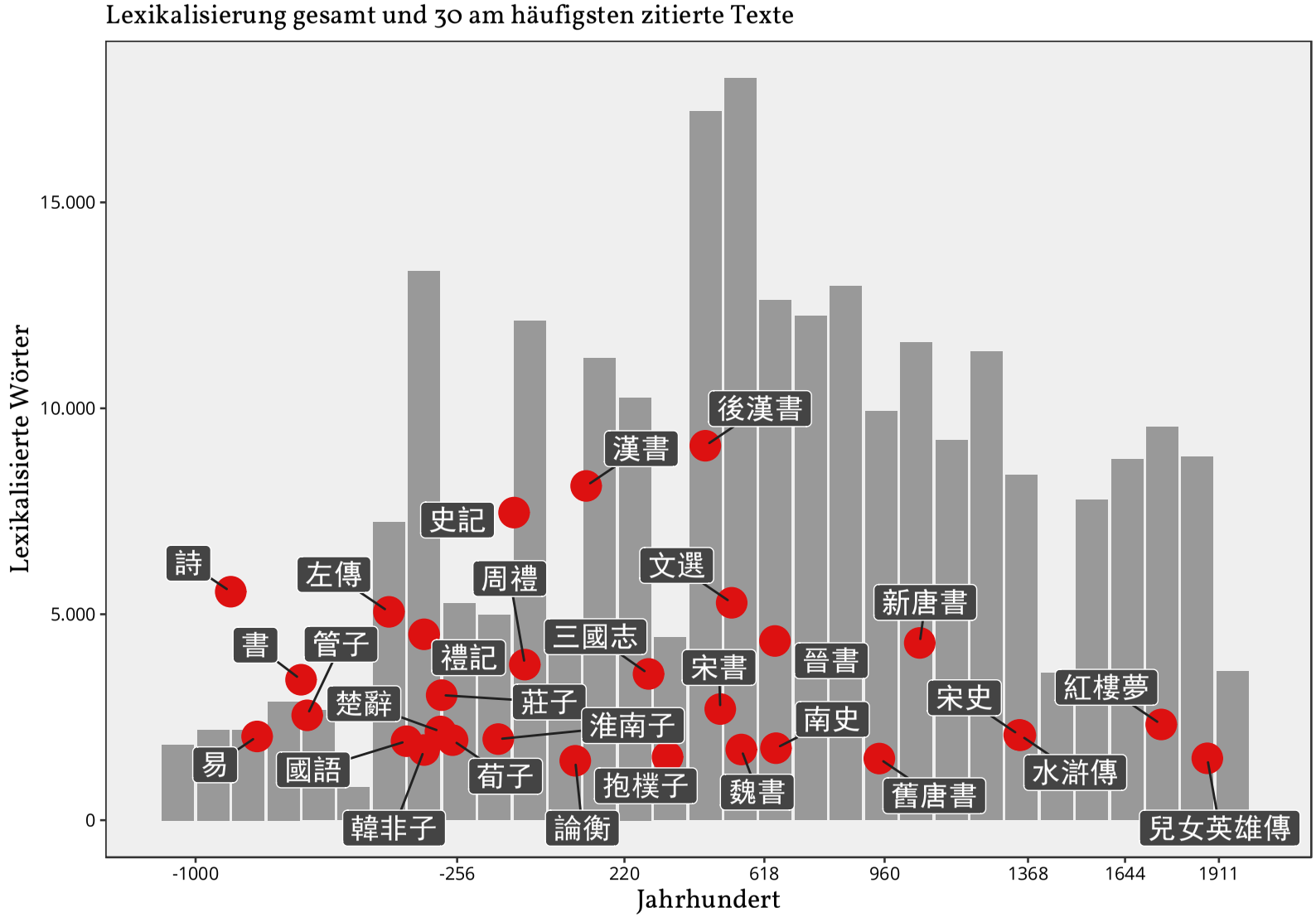


Abbildung 5.15 Lexikalisierung im *DHYDCD* nach Jahrhundert – häufigste *Locus classicus*-Texte

genau gelesen und daraus Belege exzerpiert. Auch wenn man sicher bemüht war, für jedes Lexem das ältestmögliche Textbeispiel zu zitieren, mag die Auswahl von der Versuchung beeinflusst gewesen sein, auf Konkordanzen bzw. Indizes zurückzugreifen. Dass – ähnlich wie beim *OED* – mit solchen Hilfsmitteln gearbeitet wurde, liegt nahe,[292] da gerade für die sehr häufig zitierten Texte wie *HHS*, *Han shu* 漢書, *Shiji* 史記, *Shijing* 詩經, *Sanguo zhi* 三國志 und weitere entsprechende Bände aus der Harvard Yenching-Reihe sinologischer Indizes vorliegen, die ab 1931 von Hong Ye 洪業 (William Hung) et al. herausgegeben wurde.[293] Als Beispiel sei der Eintrag zu *zhong gui ren* 中貴人 genannt. Darin wird die Biographie des Generals Li (*Li jiangjun liezhuan* 李將軍列傳) aus dem *Shiji* zitiert[294] – dieselbe Textstelle, die auch im entsprechenden Index unter *zhong gui ren* als erstes angegeben ist.[295] Die Prominenz früher Texte wie *Yijing* 易經, *Shangshu* 尚書 und *Shijing* 詩經 ist selbstverständlich, da aus der abgedeckten Zeit sonst wenig umfangreiches Textmaterial erhalten ist. Davon abgesehen prägt eine Vorliebe für die offiziellen Dynastiegeschichten (*zhengshi* 正史) die Liste häufig zitierter Texte.

Betrachtet man die 30 unabhängig von der Angabe als früheste Belegstelle meistzitierten Texte (Tabelle 5.4), ergibt sich eine ähnliche Liste. Einige wichtige Romane wie *Hong lou meng* 紅樓夢, *Shui hu zhuan* 水滸傳 und *Ru lin wai shi* 儒林外史 rücken auf, weitere Texte mit umgangssprachlichen Elementen wie die Geschichtensammlung *Liao zhai zhi yi* 聊齋志異, sind ebenfalls häufiger vertreten. Den 30 meistzitierten Texten sind 21,7 % aller Belege entnommen.[296]

Tabelle 5.4 30 meistzitierte Werke im *DHYDCD*

#	Text	Datierung	Belegstellen	Länge in 1.000 Zeichen
1	*Hou Han Shu* 後漢書	ca. 400–500	9.091	917,5
2	*Han shu* 漢書	111	8.115	712,2
3	*Shiji* 史記	-94	7.468	512,6
4	*Xin Tang shu* 新唐書	1060	5.547	1.800,8
5	*Wenxuan* 文選	ca. 520–530	5.278	996,2
6	*Jin shu* 晉書	648	5.058	1.167
7	*Zuo zhuan* 左傳	ca. -500–-400	4.513	180,5
8	*Shijing* 詩經	ca. -1100–-700	4.351	30,5

292 Siehe Kapitel 5.2, ab S. 111.

293 Siehe Tabelle 5.3, S. 150; vgl. u. a. Hong Ye 洪業 (William Hung), Hrsg. 1966 [1949]: *Combined indices to Hou Han shu and the notes of Liu Chao and Li Hsien (*後漢書及注釋綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 41. Taipei 台北 [Beijing 北京]: Harvard-Yenching [Yanjing xue she 燕京學社]; Hong Ye 洪業 (William Hung) et al., Hrsg. 1966 [1940]: *Combined indices to Han Shu and the notes of Yen Shih-ku and Wang Hsien-ch'ien (Hanshu ji buzhu zonghe yinde* 漢書及補註綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京學社引得) 36. Taipei 台北 [Beijing 北京]: Harvard Yenching Institute [Yanjing da xue tu shu guan 燕京大學圖書館]; Hong Ye 洪業 (William Hung), Hrsg. 1955 [1947]: *Combined indices to Shih chi and the notes of P'ei Yin, Ssu-ma Cheng, Chang Shou-chieh, and Takigawa Kametaro (Shi ji ji zhu shi zong he yin de* 史記及注釋綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 40. Cambridge, MA [Beijing 北京]: Harvard University Press [Yenching University Press]; Hong Ye 洪業 (William Hung) et al., Hrsg. 1934: *A concordance to Shih ching (Mao shi yin de* 毛詩引得*)*. Harvard-Yenching Institute Sinological Index Series Supplement (Hafo Yanjing xue she yin de te 哈佛燕京大學圖書館引得特刊) 9. Beijing 北京: Harvard-Yenching (Hafo Yanjing xueshe 哈佛燕京學社); Hong Ye 洪業 (William Hung) et al., Hrsg. 1938: *Combined Indices to San Kuo Chih and the Notes of P'ei Sung-chih (*三國志及裴注綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 33. Beijing 北京: Harvard-Yenching (Hafo Yanjing xueshe 哈佛燕京學社).

294 *DHYDCD*, 中貴人.

295 Siehe Hong Ye 洪業 (William Hung) 1955 [1947], S. 62.

296 Insgesamt werden nach den in Abschnitt 5.5.2 (ab S. 128) beschriebenen Kritierien mehr als 155.000 Quellenangaben unterschieden. Eine Liste der am häufigsten zitierten 20.804 würde 80 %, 63.727 dann 90 % der insgesamt 919.280 identifizierten Belegstellen abdecken.

Tabelle 5.4 (Fortsetzung)

#	Text (Konkordanz)	Datierung	*Locus classicus*-Angaben	Länge in 1.000 Zeichen
9	*Hong lou meng* 紅樓夢	ca. 1730–1764	4.303	731,1
10	*Liji* 禮記	ca. -400–-300	3.777	98
11	*Sanguo zhi* 三國志	ca. 280–297	3.551	390,4
12	*Shui hu zhuan* 水滸傳	ca. 1320–1372	3.410	437,3
13	*Song shi* 宋史	1345	3.036	4.037
14	*Zhou li* 周禮	ca. -150–23	2.698	52,8
15	*Song shu* 宋書	ca. 492–493	2.546	811,1
16	*Liaozhai zhiyi* 聊齋志異	1740	2.324	381,4
17	*Shangshu* 尚書	ca. -1100–-300	2.154	25,7
18	*Jiu Tang shu* 舊唐書	945	2.069	2.001,9
19	*Ernü yingxiong zhuan* 兒女英雄傳	1878	2.063	472,1
20	*Ming shi* 明史	1643	2.033	2.081,9
21	*Nan shi* 南史	ca. 643–659	1.969	676,2
22	*Zhuangzi* 莊子	ca. -400–-200	1.943	65,1
23	*Huainanzi* 淮南子	ca. -139	1.918	130,8
24	*Guanzi* 管子	ca. -720–-645	1.748	57,5
25	*Guoyu* 國語	ca. -500–-300	1.716	70,4
26	*Baopuzi* 抱樸子	ca. 265–420	1.703	152,2
27	*Xunzi* 荀子	ca. -300–-238	1.534	64,9
28	*Zi zhi tong jian* 資治通鑒	1084	1.502	1.933,1
29	*Chu ci* 楚辭	-329–-278	1.496	29,8
30	*Ru lin wai shi* 儒林外史	1749	1.442	231,9

Die Auswahl der am häufigsten im *DHYDCD* zitierten Texte zeigt eine insgesamt stark in der klassischen Tradition verwurzelte Textrezeption durch die Herausgeber:innen, die durchaus an das *KXZD* erinnert. Im Gegensatz zu diesem wird aber zumindest einer Auswahl an moderneren, umgangssprachlichen Quellen ebenfalls Gewicht verliehen.

Der Schwerpunkt dieses Kapitels lag in der Erschließung des *DHYDCD* als diachrone Lexemdatenbank und der Analyse dieser Ressource. Die Datenbank dient als Basis für die Entwicklung der Datierungsmethoden für schriftsprachliche chinesische Texte, die in Kapitel 6.2 (ab S. 179) und 6.3 (ab S. 210) beschrieben werden. Die Erzeugung eines diachronen Behelfskorpus aus dem *DHYDCD* erlaubt es zudem, die in Kapitel 6.1 (ab S. 156) für schriftsprachliche chinesische Texte adaptierten Methoden für einen langen Betrachtungszeitraum zu evaluieren.

6 Textdatierung für schriftsprachliches Chinesisch

> *„While the majority of diachronic studies focus on change in language, we should also not forget the flipside of language change, stability over time, which is equally interesting."*[1]
>
> Vaclav Brezina

Dieses Kapitel widmet sich computerlinguistischen Methoden zur chronologischen Einordnung chinesischsprachiger Texte. Wie in Kapitel 3.3 (ab S. 45) skizziert, lassen sich im Wesentlichen zwei Arten von Datierungsaufgaben unterscheiden: **1.** Die Klärung der Frage, *wann* ein Text etwa verfasst wurde und **2.** eine temporale Einordnung des *Inhalts*, z. B. in welcher Zeit sich die Handlung eines literarischen Texts abspielt, bzw. welcher Zeitraum in einem historiographischen Text beschrieben wird. Hier soll es primär um erstere Herausforderung gehen.

Dass bislang keine Arbeiten zur Datierung chinesischsprachiger Texte vorliegen, in denen die in Kapitel 3.3 (ab S. 45) beschriebenen Methoden eingesetzt werden, ist vermutlich auf den Mangel an geeigneten (Trainings-)korpora zurückzuführen.[2] Für den Zeitraum von 1475 bis ca. 1925 steht mit dem von CrossAsia veröffentlichten *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌) ein großer Datensatz mit den Häufigkeiten der 1–3-Gramme von 11.081 Lokalchroniken (*difangzhi* 地方誌, *DFZ*) zur Verfügung,[3] deren Entstehungszeit bekannt ist und aus dem sich dadurch entsprechende *Statistical Language Models* (*SLM*) berechnen lassen.

In Kapitel 6.1 (ab S. 156) implementiere ich die wesentlichen Ansätze aus den in Kapitel 3.3 vorgestellten Methoden, die auf solchen *chronon*-Sprachmodellen basieren, um ihre Eignung für die (domänenspezifische) Datierung chinesischsprachiger Texte zu testen.

Um eine genreübergreifende Datierung von Texten über einen größeren Zeitraum hinweg zu ermöglichen, arbeite ich zudem zur Erzeugung der temporalen Sprachmodelle mit dem Behelfskorpus aus den Beispielsätzen des *DHYDCD*.[4]

Da die teilweise Rigidität einiger Genres der chinesischen Schriftsprache und das Fehlen umfangreicherer Trainingskorpora die Zuverlässigkeit der Datierung bei der Verwendung dieser Methoden auf Basis statistischer Sprachmodelle stark einschränken können, wird zudem ein innovativer, datenbankgestützter Ansatz zur chronologischen Einordnung von Texten vorgestellt, dem die historischen Lexikalisierungsdaten aus dem *DHYDCD* zugrunde liegen.[5] Werden

1 Vaclav Brezina 2018: *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge & New York: Cambridge University Press, S. 220.

2 Siehe auch Kapitel 4.2, S. 62.

3 *DFZ*, siehe auch Kapitel 4.2, S. 62. Zwar sind in diesem Datensatz vereinzelt deutlich frühere Texte aus der Zeit der Tang 唐-Dynastie (618–907) enthalten, jedoch ist erst für den Zeitraum ab dem Ende des 15. Jahrhunderts eine ausreichende Menge an Texten vorhanden, um sowohl Trainings-, als auch Testdatensätze zur Verfügung zu haben. Die Veröffentlichung erfolgt als *n*-Gramme, da eine Bereitstellung der Volltexte lizenzrechtlich problematisch wäre.

4 Siehe Kapitel 5.6, ab S. 137.

5 Siehe Kapitel 5, S. 107–154.

Veröffentlicht in: Tilman Schalmey: *Computerlinguistische Datierung schriftsprachlicher chinesischer Texte*. Heidelberg: HASP, 2023. DOI: https://doi.org/10.11588/hasp.1153

diese für alle in einem Text enthaltenen Lexeme herangezogen, lässt sich ein *Neologismusprofil* erzeugen, welches das (stilistische) Alter des Textes visualisiert (Kapitel 6.2, ab S. 179) und von Philolog:innen zu Analysezwecken herangezogen werden kann, um die zeitliche Einordnung von Texten zu erleichtern. Mittels Regression auf die Daten solcher Neologismusprofile lassen sich überdies für die automatisierte Datierung schriftsprachlicher chinesischer Texte teilweise bessere Ergebnisse erzielen, als dies mit den für die Datierung westlichsprachiger Texte gängigen *SLMs* möglich ist.

Auch mit dem experimentellen Ansatz von Berechnungen auf Basis des durchschnittlichen Jahres der Lexikalisierung der in einem Text enthaltenen Lexeme (*Average Year of Lexicalization*, *AYL*) lassen sich überraschende Ergebnisse erzielen. Mittels Regression kann so ebenfalls das Alter eines Textes geschätzt werden, allerdings – wegen der stilistischen Unterschiede zwischen unterschiedlichen Textgattungen – nur in einem eng gesteckten Rahmen (Kapitel 6.3, ab S. 210).

Zuletzt werden die wesentlichen Ergebnisse und Unterschiede zwischen den untersuchten Datierungsmethoden, sowie sich daraus ergebende Herausforderungen und Limitationen zusammengefasst und eine Benutzer:innenoberfläche für die erarbeiteten Möglichkeiten präsentiert (Kapitel 6.4, ab S. 229).

6.1 Datierung als Kategorisierungsproblem

„What am I gonna do?
I'm gonna train myself a classifier*!"*
Paul Vierthaler

Für die Textdatierung auf Basis statistischer Sprachmodelle werden zunächst aggregierte Wort- bzw. *n*-Gramm Häufigkeiten für bestimmte Zeitabschnitte (*chronons*) berechnet. Ein zu datierender Text *d* wird dann dem *chronon* mit der größten Übereinstimmung zugeordnet.[6] Anhand des *DFZ*-Datensatzes,[7] sollen zunächst folgende Fragen ergründet werden:

1. Sind statistische Sprachmodelle aus Trainingsdaten dieses Korpus geeignet, um andere Texte innerhalb desselben Korpus zu datieren?
2. Mit welchen Ähnlichkeitsmaßen lässt sich die beste Performance für die Datierung chinesischsprachiger Texte erzielen?
3. Wie wirkt sich eine Reduktion auf „Lexemdimensionen" und die Erweiterung dieser um Namen, Ortsnamen und temporale Ausdrücke auf die Performance aus?
4. Wie wirken sich weitere Parameter, z. B. Mindesthäufigkeiten als Möglichkeit der Reduktion der zu betrachtenden Dimensionen (*features*) aus?
5. Wie verhält es sich mit einer Erweiterung oder Reduktion des verwendeten *n*-Gramm-Raums auf 1-Gramme, 1–2-, 1–3-Gramme?
6. Kann die Performance der Modelle durch Glättung (*smoothing*) verbessert werden?

Die Performance der Modelle und Ähnlichkeitsmaße wird dabei am Anteil der „richtig" datierten Texte (*Accuracy*), sowie am mittleren Abstand der jeweiligen Datierung von den tatsächlichen

6 Ausführlicher in Kapitel 3.3, ab S. 45.
7 Siehe auch Kapitel 4.2, S. 66.

Jahresangaben aus den Metadaten der Texte (*mean error*) gemessen. Als *richtig* werden diejenigen Datierungen angesehen, bei denen das Jahr der Veröffentlichung bzw. das mittlere Jahr des Erscheinungszeitraums (`cleanyear`) in das datierte *chronon* fällt. Durch die Überlappung der *chronons* können also in der Regel bis zu zwei *chronons* als richtig gewertet werden.

Der *mean error* D_{mean} der Datierung eines Textes ist dabei definiert als arithmetisches Mittel des Intervalls der Jahre von Beginn und Ende des datierten *chronons* zum `cleanyear` des zu datierenden Textes:

$$D_{mean} = \frac{(|D_{start}| + |D_{end}|)}{2}$$

Ist also z. B. die Veröffentlichung eines Texts mit 1525–1527 angegeben, der korrekt auf das *chronon* 1500–1550 datiert wird, so beträgt D_{mean} 25 Jahre. Dieser minimale Wert ergibt sich bei jeder *richtigen* Datierung eines Textes:

$$D_{mean} = \frac{(|\frac{1527+1525}{2} - 1500| + |\frac{1527+1525}{2} - 1550|)}{2} = \frac{(|1526 - 1500| + |1526 - 1550|)}{2} = \frac{(26 + 24)}{2} = 25$$

Für eine Gesamtanzahl *N* an Testtexten ist der *mean average error* (*MAE*) also angegeben als

$$MAE = \frac{1}{N}\sum_{n=1}^{N} D_{mean}$$

und spiegelt so die ungefähre Genauigkeit einer Datierung wider.

Als Ähnlichkeitsmaße werden *Cosine similarity* (*CS*) ohne und mit Gewichtung mittels *inverse document frequency* (*idf*),[8] Jaccard *similarity*, sowie *Normalized-Log-Likelihood-Ratio* (*NLLR*) und Kullback-Leibler-Divergenz (*KLD*) getestet, jeweils ohne und mit Gewichtung mittels temporaler Entropie (TE).[9]

Ebenfalls anhand des *DFZ*-Datensatzes soll zudem erörtert werden, ob eine Co-Datierung von Dokumenten gegenüber der Verwendung von *chronon*-Modellen vorzuziehen ist.[10] Um den Vergleich auf eine solidere Grundlage zu stellen, wird dieses Experiment mit einem weiteren Datensatz, dem *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*)[11] wiederholt.[12]

Um in der Lage zu sein, Texte unabhängig von ihrer Zugehörigkeit zu einem stilistisch und vor allem temporal stark eingeschränkten Korpus zeitlich einordnen zu können, werden zudem Experimente mit Sprachmodellen durchgeführt, die aus dem Behelfskorpus der *DHYDCD*-Belegstellen erzeugt werden.[13] Anhand derselben Daten lassen sich zudem grobe quantitative Beobachtungen über Veränderungen der Wortnutzung und des Wortschatzes in dem von diesem Korpus abgedeckten Zeitraum von 700 v. u. Z. bis zum 20. Jh. machen.

8 Siehe dazu S. 168; siehe auch Kapitel 3.3, S. 53.
9 Ausführlich dazu siehe Kapitel 3.3, ab S. 45.
10 Vgl. Kapitel 3.3, S. 50; siehe auch de Jong, Rode und Hiemstra 2005, S. 7; Bamman et al. 2017, S. 4.
11 CrossAsia, Staatsbibliothek zu Berlin 2019b: *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書. Datenset. Version 0.0.1-20190307. DOI: 10.5281/zenodo.2586940.
12 Siehe Abschnitt 6.1.2, ab S. 169.
13 Siehe 6.1.3, ab S. 171.

6.1.1 Datierung mit *difangzhi* 地方誌 Sprachmodellen

Aus dem *DFZ*-Datensatz wird zunächst ein temporales Sprachmodell mit 17 *chronons* für einen Zeitraum von 450 Jahren erzeugt. Hierfür wird eine *chronon*-Dauer von 50 Jahren mit jeweils 25-jähriger Überlappung gewählt.[14] Das erste *chronon* deckt den Zeitraum von 1475–1525 ab, das zweite 1500–1550 und das letzte die Jahre 1875–1925.[15]

Für jedes *chronon* werden jeweils 50 zufällig ausgewählte Lokalchroniken als Trainingsdaten verwendet.[16] Dabei werden nur Texte berücksichtigt, für die beide chronologischen Einordnungen (Angaben zum Inhalt und zeitliche Angaben zu Herausgeber- bzw. Autorschaft) vorliegen. Außerdem werden Texte, bei denen beide Angaben 50 oder mehr Jahre auseinander liegen, ausgeschlossen. So soll sichergestellt werden, dass es sich nicht um eine (überarbeitete) Neuausgabe eines eigentlich deutlich älteren Texts handelt und die Verfasserschaft tatsächlich ungefähr in den angegebenen Veröffentlichungszeitraum fällt. Ist anstatt des Erscheinungsjahres ein Zeitraum angegeben, wird das mittlere Jahr dieses Zeitraums für die chronologische Einordnung verwendet, also z. B. 1909 für 1908–1910. Durch die Überlappung der *chronons* kann sich die Textauswahl für die Trainings-Subkorpora teilweise überschneiden.[17] Für alle *chronons* werden die absoluten (*tokens*) und relativen Häufigkeiten (*tf*), sowie temporale Entropie (TE) und *term frequency inverse document frequency* (*tf-idf*) vorberechnet.[18] Außerdem werden die Gesamthäufigkeiten des Trainingskorpus vorgehalten. Da der *DFZ*-Datensatz zahlreiche Variantenzeichen (*yitizi* 異體字) enthält, wird die in Kapitel 4.3 beschriebene Normalisierung lediglich zur Erkennung mehrsilbiger Ausdrücke als Lexeme verwendet, um einen potenziellen Datenverlust zu minimieren.[19] Als *Baseline* für die unterschiedlichen Ähnlichkeitsmaße dient ein Zufallsgenerator, der eines der 17 *chronons* auswählt und theoretisch mit einer Wahrscheinlichkeit von etwa 11 % die Testdokumente korrekt datieren wird.

Werden solche Sprachmodelle für alle 1–2-Gramme berechnet, entsteht eine Dokumentensammlung mit insgesamt mehr als 7 Mio. *n*-Gramm *types*. Etwa 40 % davon sind *Hapax legomena* mit nur einem einzigen Vorkommen im Korpus, die nur mit sehr geringer Wahrscheinlichkeit in einem *Query*-Dokument *d* vorkommen. Berechnungen mit einer so großen Zahl an Vektoren sind zudem wenig performant.

Obwohl die *chronon*-Sprachmodelle jeweils aus der gleichen Anzahl von Texten aggregiert werden, kann die Anzahl der enthaltenen *types* stark schwanken. So enthält z. B. das *chronon* Sprachmodell von 1550–1600 ca. 1,5 Mio. *n*-Gramm *types*, das Modell von 1850–1900 mit über 2,1 Mio. deutlich mehr.[20] Diese Diskrepanz kann bei der Berechnung einiger Ähnlichkeitsmaße problematisch sein, da große *chronons* so lediglich deswegen „ähnlicher" zu einem *Query*-Dokument sein können, weil mehr Dimensionen verglichen werden können. Dieser Effekt kann durch die Verwendung von ausgeglichenen Modellen reduziert werden, indem alle *chronons* so lange um seltene *types* reduziert werden, bis die Anzahl der betrachteten *n*-Gramm-Vektoren

14 Die Erzeugung versetzt überlappender *chronons* ermöglicht einen besseren Umgang mit *chronon*-„Rändern". Siehe dazu de Jong, Rode und Hiemstra 2005, S. 4.

15 Der Datensatz enthält zwar vereinzelt auch deutlich ältere Texte ab dem 8. Jh. Als Trainingsdaten reichen diese aber noch nicht aus. Erst ab der Ming-Zeit ist eine substanzielle Anzahl an *DFZ* erhalten geblieben. Vgl. auch Dennis 2015, S. 2.

16 Aus der Datenqualität des Korpus ergeben sich dabei Einschränkungen. Die Kritierien für die Auswahl der Texte werden in Kapitel 4.2 dargelegt (siehe S. 67).

17 Bei der verwendeten Auswahl werden etwa 10 % der Texte in zwei *chronons* verwendet.

18 Siehe dazu Kapitel 3.3. S. 53, S. 53.

19 Siehe S. 70; siehe auch Kapitel 4.2, S. 66.

20 Insgesamt ist im Betrachtungszeitraum ein eindeutiger Trend zu längeren Texten mit mehr *types* nachweisbar.

derjenigen des „kürzesten“ *chronons* entspricht. Durch eine derartige Reduktion der Dimensionen können andererseits allerdings auch Merkmale eliminiert werden, die eine temporale Diskriminierung der Dokumente überhaupt erst ermöglichen würden.[21]

Als Testdaten werden unter Ausschluss der Trainingsdaten 216 zufällig ausgewählte *Difangzhi* verwendet, deren Veröffentlichung gleichmäßig über den Betrachtungszeitraum von 1475–1925 verteilt ist. Die Häufigkeit $P(\hat{\theta} \mid c)$ von Wörtern aus einem zu datierenden Text d, die in einem Vergleichs-*chronon* c nicht auftreten, aber für die Berechnung von *NLLR* und *KLD* benötigt werden (sogenannte *unseen events*),[22] wird zunächst angenommen als Hälfte der niedrigsten Häufigkeit eines Wortes im längsten *chronon* c_{long}, also $P(\hat{\theta} \mid c) = \lambda \times P_{min}(w \mid c_{long})$ mit $\lambda = 0,5$. Dadurch soll vereinfacht modelliert werden, dass die Wahrscheinlichkeit dieses *unseen events* niedriger ist als die geringste tatsächlich in einem *chronon* auftretende Worthäufigkeit. Diese naive Glättungsmethode bezeichne ich im Folgenden als *Largest chronon minimum* (*LCM*) *smoothing*.[23]

Experimente mit *difangzhi* 地方誌-Sprachmodellen

Im Folgenden werden Experimente zur Beantwortung der eingangs formulierten Fragen durchgeführt. Tabelle 6.1 (S. 165) gibt einen Gesamtüberblick über die Ergebnisse dieser Untersuchungen.

1–2-Gramme.
Als Basismodell wird ein vollständiges (d. h. bei Verwendung aller verfügbaren *types* unabhängig von ihrer Häufigkeit und Sinnhaftigkeit) 1–2-Gramm Modell verwendet.[24] Während die Texte mit Jaccard *similarity* kaum besser datiert werden als mit einem Zufallsgenerator, stellen sich alle anderen verwendeten Metriken als aussagekräftig heraus. *Tf-idf* zur Gewichtung der *CS* verbessert die Performance im Vergleich zur einfachen *CS* erheblich, während die aufwändigere Gewichtung von *KLD* und *NLLR* mit temporaler Entropie zunächst eher einen negativen Effekt zu haben scheint (Abb. 6.1). Ein Grund dafür ist sicherlich die große Menge an *Hapaxen*, bzw. dass *types*, die ein einziges Mal in einem *chronon* auftreten tendenziell übergewichtet werden, da ihr seltenes Auftreten auch zufällig sein kann, ohne ein tatsächliches Indiz für die Entstehungszeit des Textes zu sein. *KLD* und *NLLR* sind tendenziell aussagekräftiger als CS_{tfidf}, *KLD* und *NLLR* quasi gleichwertig.[25]

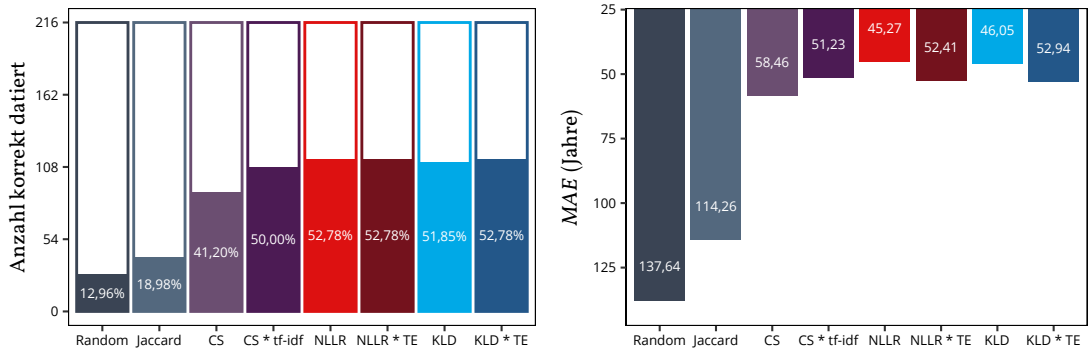
21 Alternativ könnte ein Ausgleich über eine unterschiedliche Anzahl von Texten oder eine unterschiedliche Länge der *chronons* erreicht werden.

22 Siehe Kapitel 3.3, ab S. 45.

23 Diese Herangehensweise ist prinzipiell ähnlich einem *add one* bzw. Laplace-*smoothing*, aber ohne Veränderung der Grundmasse bzw. ohne die tatsächliche Vorkommenshäufigkeit zu verändern. Die Verwendung anspruchsvollerer *smoothing*-Methoden wird in Abschnitt 6.1.1 (S. 164) beschrieben und untersucht.

24 Dies ist nicht direkt mit der Verwendung eines Bigramm-Modells für westliche Sprachen vergleichbar, da Bigramme hier sowohl zwei Wörter mit je einem Zeichen, ein Wort aus zwei Zeichen, oder die Kombination aus dem letzten Zeichen eines Wortes mit mehreren Schriftzeichen und dem ersten Zeichen des nächsten Wortes usw. repräsentieren können. Siehe auch Kapitel 3.3, S. 47.

25 Vgl. auch Kraaij 2004, S. 208.



(a) Anteil richtig datierter *difangzhi* **(b)** Durchschnittliche Abweichung in Jahren (*MAE*)

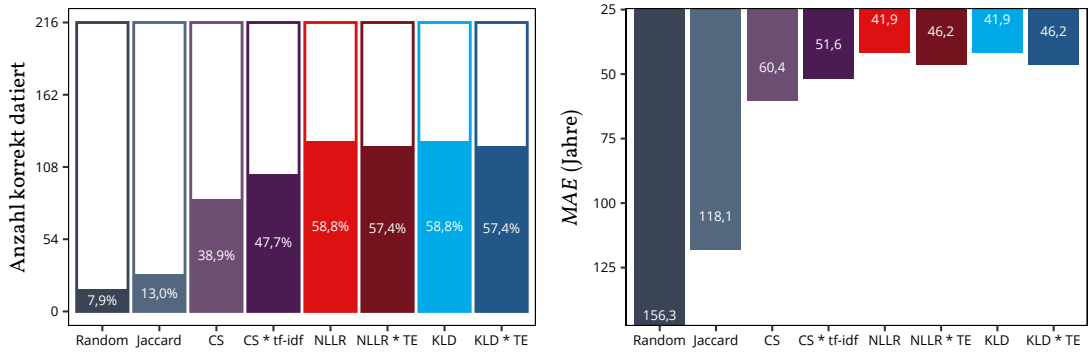
Abbildung 6.1 Performance mit 1–2-Gramm *difangzhi*-Sprachmodell

Reduktion auf Lexemdimensionen.
Durch eine Reduktion der verwendeten Dimensionen um etwa 96 % auf diejenigen 238.707 Uni- und Bigramme des Modells, die als Lexeme im *DHYDCD* aufgeführt sind, lässt sich die Dauer der Berechnung um etwa 95 % reduzieren. Die so erzeugten *SLM*s bezeichne ich im Folgenden als Lexem-Modelle. Im Unterschied zu einem klassischen *BoW*-Modell handelt es sich hierbei streng genommen um stark dimensionsreduzierte Zeichen-*n*-Gramm-Modelle.[26] Dabei werden zunächst auch Vorkommen von Einzelzeichen aus dem Modell eliminiert, wenn diese im *DHYDCD* nicht lexikalisiert sind und dadurch auch ca. 8.000 Unigramm-*types* bzw. chinesische Schriftzeichen weniger berücksichtigt, als das beim reinen *n*-Gramm Modell der Fall war.[27] Das Lexem-Modell enthält mit nur 21.750 (knapp 10 %) zudem deutlich weniger *Hapaxe* als das *n*-Gramm-Modell.[28] Die relativen Häufigkeiten und die Gewichte *tf-idf* und *TE* des so entstandenen Lexem-Sprachmodells werden sowohl für die *chronons* als auch das gesamte Modell auf Grundlage der Lexem-*tokens* neu berechnet. Trotz der Verwendung nur eines Bruchteils der *features* erhöht sich der Anteil der korrekt datierten Texte mit *NLLR* und *KLD* und der *MAE* sinkt um etwa 5 Jahre. Lediglich die Performance von *CS* bzw. CS_{tfidf} geht durch die Reduktion der Dimensionen leicht zurück (Abb. 6.2; Tabelle 6.1).

26 Siehe dazu auch Kapitel 4.5.3, S. 94.

27 Siehe auch Experiment Nr. 10 in Tabelle 6.1: Die Verwendung dieser zusätzlichen Zeichen für die Modellberechnung führt lediglich zu marginalen Unterschieden. Es handelt sich bei diesen Zeichen um im *DHYDCD* fehlende Einträge wie *bing* 丙 und *mei* 美 (siehe dazu Kapitel 5.4.1, S. 120), nicht automatisch normalisierte Zeichenvarianten, z. B. 吳 für *wu* 吳/吴, 出 für *chu* 出, 蒙 für *meng* 蒙 (siehe dazu Kapitel 4.3, ab S. 70), seltene Zeichen, die nicht im *HYDCD* lexikalisiert sind, wie *chun* 蓴 („Kräutermedizin"), *niang* 蘘 („Medizinkräuter"), einige Zeichen, die sehr wahrscheinlich durch Codierungsfehler entstanden sind, wie der javanische Buchstabe *ba* ꦧ, sowie Symbole aus Schriftsystemen ethnischer Minderheiten wie der *Yi* 彝.

28 In einer Studie zur Anwendbarkeit von Zipf's Gesetz auf das Chinesische stellt Xiao fest, dass in einem „großen", modernen Korpus über 40 % *Hapax legomena* auftreten. Siehe Xiao Hang 2008: „On the Applicability of Zipf's Law in Chinese Word Frequency Distribution". In: *Journal of Chinese Language and Computing* 18.1, S. 33–46, S. 44. Dies entspricht etwa dem hier (s. o.) für *n*-Gramme beobachteten Anteil, obwohl Xiao mit einem getaggten, segmentierten Korpus arbeitet. Es muss also vermutet werden, dass ein relevanter Teil der in den Texten enthaltenen Wörter nicht im *HYDCD* lexikalisiert ist.



(a) Anteil richtig datierter *difangzhi* **(b)** Durchschnittliche Abweichung in Jahren (*MAE*)

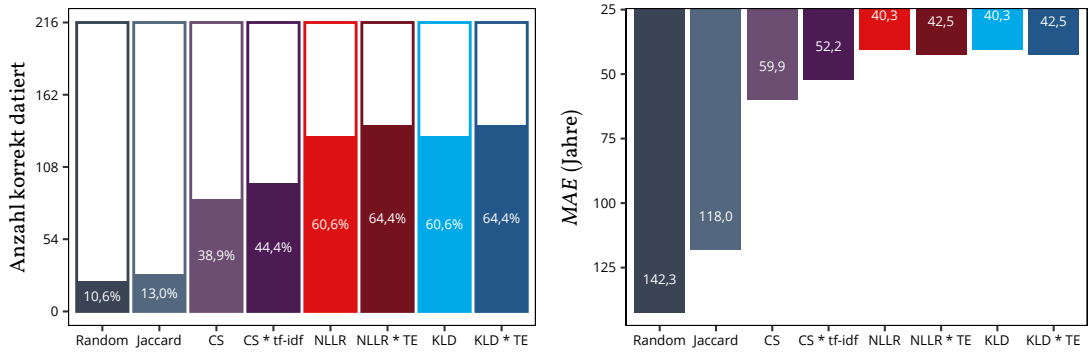
Abbildung 6.2 Performance von 1–2-Zeichen *difangzhi*-Lexem-Sprachmodellen

Verwendung gleichförmiger *chronons*.
Hierfür wird die Anzahl der *types* jedes *chronon*s durch Entfernen niedrigfrequenter *types* so lange reduziert, bis sie der Anzahl der *types* des „kürzesten" *chronons* (hier 112.281) entspricht, also eine weitere Dimensionsreduktion um je 0–27 %. Hierdurch ist nur eine entscheidende Veränderung in der Performance feststellbar: die Anzahl der mit Jaccard *similarity* richtig datierten Texte erhöht sich um fast 30 Prozentpunkte auf 41.7 % (ohne Abb.). Das Ergebnis reicht nicht an die komplexeren bzw. gewichteten Metriken heran, weist aber darauf hin, dass es für die Datierung schriftsprachlicher Texte – zumindest in diesem Textkorpus – vielleicht wichtiger ist, *welche types* in einem Text auftreten, als wie häufig sie auftreten und dass – wie auch die relative „Erfolglosigkeit" der Gewichtung mittels temporaler Entropie zeigt – der Sprachwandel im Hinblick auf die Veränderung der Häufigkeiten von Wörtern im Betrachtungszeitraum in diesem Korpus verhältnismäßig gering ist.

Festlegung einer Mindesthäufigkeit von *types* für die Berücksichtigung in Metriken. Durch die Festlegung einer Mindesthäufigkeit von *types* könnte der Effekt übergewichteter, seltener *types* minimiert werden. Erhöht man die Mindesthäufigkeit von *types* für die Erzeugung des Modells und die Berechnung der Ähnlichkeitsmaße auf 2, lässt sich allerdings keine nennenswerte Veränderung der Performance feststellen, da der positive Effekt der *feature reduction* vermutlich durch den Verlust seltener, diskriminativer *types* ausgeglichen wird.

Verwendung von Namen und Zeitausdrücken.
Während die Erweiterung der verwendbaren „sinnvollen" *types* um Personen- und Ortsnamen aus der *CBDB* keinen positiven Einfluss auf die verwendeten Metriken hat,[29] kann durch eine Verwendung von Zeitausdrücken[30] die Performance des bisher besten Modells (1–2 Gramm-Lexeme, keine Mindesthäufigkeit, ohne gleichförmige *chronons*), noch leicht gesteigert werden:

29 Die Ergebnisse der Experimente 7–14 sind in Tabelle 6.1, S. 165 zusammengetragen.
30 Zu diesem Zweck wird die *DHYDCD*-Lexemliste um Äranamen und *tiangan dizhi* 天干地支 Ausdrücke erweitert. Siehe dazu auch Kapitel 4.8, ab S. 103.



(a) Anteil richtig datierter *difangzhi* **(b)** Durchschnittliche Abweichung in Jahren (*MAE*)

Abbildung 6.3 Performance von 1–2-Zeichen *difangzhi*-Lexem-Sprachmodellen mit zusätzlichen temporalen Ausdrücken

Mit einer *Accuracy* von 0,64 zeigt sich erstmals auch ein spürbar positiver Effekt der Gewichtung mit TE. Diese funktioniert deutlich besser, wenn mehr häufigere Begriffe auch wirklich in verschiedenen *chronons* unterschiedlich stark auftreten, was bei „gewöhnlichen" Lexemen offensichtlich weniger der Fall ist. Ohne *TE* kann ein etwas geringerer *MAE* erzielt werden.

Abb. 6.4 zeigt die Detailergebnisse der Datierung von 216 *difangzhi* mit diesem *SLM* bei der Verwendung von *NLLR* und *TE*. Unterhalb der gestrichelten Linie ist ein etwas größerer Anteil als „zu alt" datierter Texte erkennbar. Insgesamt sind die Datierungen und Abweichungen über den gesamten Zeitraum etwa gleichmäßig verteilt.
Das Fehlen einer zeitlich diskriminativen Kraft von Ortsnamen ist wenig überraschend, da die Angaben zur ersten Nennung in der *CBDB* offensichtlich viele frühere Quellen unberücksichtigt lassen.[31] Der fehlende Effekt der Verwendung von Personennamen verdient jedoch weitere Aufmerksamkeit, insbesondere da in vielen *DFZ* mehrere hundert Übereinstimmungen mit Personennamen vorhanden sind[32] und erwartbar ist, dass in Chroniken zu unterschiedlichen Zeiten über unterschiedliche Akteure berichtet wird. Ursachen für eine ausbleibende Verbesserung der Datierungsergebnisse können in der bereits in Kapitel 4.7 erörterten Problematik liegen.[33] Zeichenfolgen, die in ihrer lexikalisierten Bedeutung vorkommen, können fälschlich als Name identifiziert werden. Hinzu kommt die Problematik mehrfacher Namensträger.[34] Möglich ist auch, dass unterschiedliche Erwähnungen von Personen in Lokalchroniken stärker eine geographische als eine temporale Zuordnung stützen.

In Tabelle 6.1 sind die Ergebnisse der oben beschriebenen Experimente zusammenfassend dargestellt. Die Spalte *DHYDCD* gibt an, ob die *types* auf die Lexeme des *DHYDCD* reduziert sind.[35] „Namen" bzw. „Orte" gibt an, ob der Lexem-Raum um die Personen- bzw. Ortsnamen aus der *CBDB* erweitert ist,[36] „Zeit" gibt an, ob die Zeitausdrücke aus der *DDBC*-Datenbank

31 Siehe auch Kapitel 4.7, ab S. 97.
32 Vgl. dazu Abb. 6.25 in Kapitel 6.2.
33 Siehe S. 97.
34 Siehe Kapitel 4.7, ab S. 97.
35 Die Angabe „+1-*grams*" (Experiment Nr. 10) bedeutet, dass hier nicht nur die lexikalisierten, sondern alle vorkommenden Einzelzeichen berücksichtigt wurden, also 7.912 zusätzliche *types*.
36 Siehe auch Kapitel 4.7, ab S. 97.

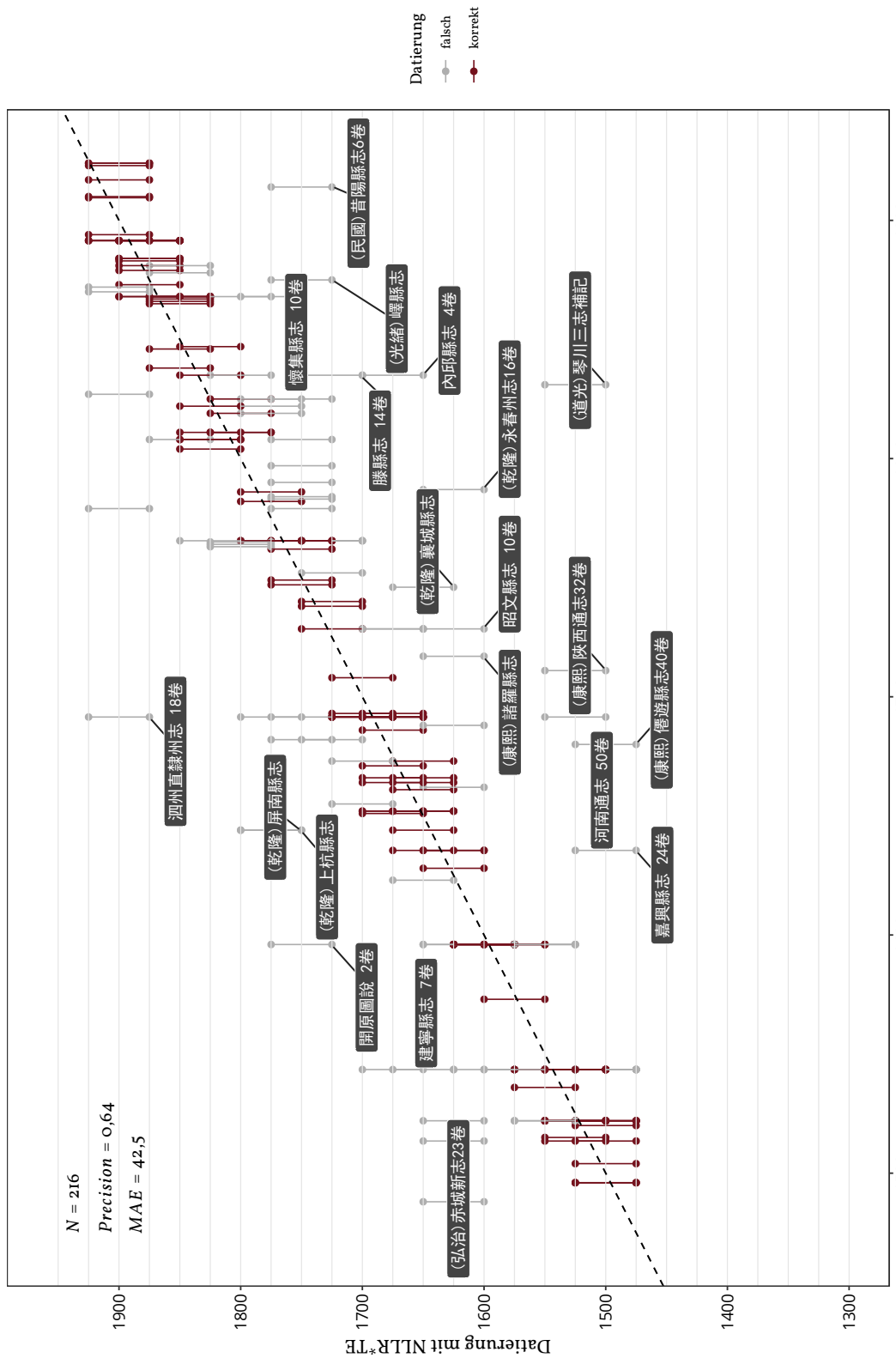


Abbildung 6.4 Performance von 1-2-Zeichen *difangzhi*-Lexem-Sprachmodellen mit zusätzlichen temporalen Ausdrücken

hinzugezogen wurden.[37] Experimente mit detaillierterer Darstellung sind mit Verweisen auf die entsprechenden Abbildungen versehen.

Smoothing von *DFZ*-Sprachmodellen

Wie beschrieben wurde bisher für alle *unseen events* in Vergleichs-*chronons* eine Wahrscheinlichkeit $P(\hat{\theta} \mid c) = \lambda P_{min}(w \mid c_{long})$ mit $\lambda = 0{,}5$ angenommen, also die Hälfte der niedrigsten Wahrscheinlichkeit des längsten *chronons* (*LCM smoothing*). Diese einfache Annahme, mit der *unseen events* zuverlässig eine niedrige Häufigkeit zugewiesen wird, scheint sich bisher gut zu bewähren. Im Gegensatz zu gewöhnlich im Kontext statistischer Sprachmodelle eingesetzter *Smoothing*-Techniken bleibt allerdings die Häufigkeit des jeweiligen Wortes bzw. *n*-Gramms im Korpus unberücksichtigt, d. h. allen *unseen events* wird dieselbe Häufigkeit zugeschrieben. Auch de Jong, Rode und Hiemstra verwenden eine niedrige Häufigkeit für *types*, die im Query-Dokument, aber nicht in einem *chronon*-Modell auftreten und stellen fest, dass die Auswirkung von ihnen getesteter *Smoothing*-Methoden (Dirichlet-*Smoothing*, *linear interpolation*) bei der Verwendung von *chronon*-Modellen gering ist.[38] Für die Verwendung ungewichteter *NLLR* trifft das auch für die *DFZ*-Modelle zu: Die Auswirkungen sowohl eines Weglassens der entsprechenden Dimensionen, als auch die Anpassung der angenommenen Häufigkeit in einem niedrigen Bereich zwischen 0,000.000.01 und 0,000.000.1 sind gering. Das mag damit zusammenhängen, dass der Anteil der *unseen events* aus Query-Dokumenten in *chronons* hier zwischen etwa 1 und 10 % liegt, also zur Berechnung von *KLD* oder *NLLR* nur 1–10 % der *types* überhaupt von diesem vereinfachten *smoothing* betroffen sind. Bei der Verwendung von Gewichten wie *TE* kann die Auswirkung der *Smoothing*-Methode und Parameter allerdings immens werden, da gerade diejenigen *types*, die nur in wenigen oder in einem *chronon*-Modell auftreten, eine hohe Entropie haben. Zhai Chengxiang, der sich intensiv mit *Smoothing*-Methoden im Kontext von *SLM*s beschäftigt, stellt fest, dass „nonoptimal smoothing can degrade retrieval performance significantly."[39]

Im Folgenden werden daher einige der im Rahmen von Textdatierungen mittels *SLMs* üblichen *Smoothing*-Methoden[40] auf ihre Eignung für die Datierung von *DFZ*-Texten überprüft.

1. Laplace / *add one smoothing*. Dabei wird angenommen, dass die Häufigkeit aller *events* um einen gewissen Wert λ höher ist. Dadurch erhalten *unseen events* in jedem *chronon* eine Häufigkeit von λ:

$$P(w, \theta) = \frac{c(w, D) + \lambda}{|D|}$$

   Mit $\lambda = 1$ wird von *add one smoothing* gesprochen. Relative Häufigkeiten und Gewichte werden entsprechend neu berechnet.

37 Siehe dazu auch Kapitel 4.8, ab S. 103.

38 Siehe de Jong, Rode und Hiemstra 2005, S. 3. Es wird eine „very small (non-zero) probability" verwendet, ohne auf die Anwendung bzw. Parameter-Bestimmung für das angeblich verwendete *linear interpolation smoothing* und *Dirichlet smoothing* einzugehen.

39 Siehe Zhai Chengxiang 2008: „Statistical Language Models for Information Retrieval: A Critical Review". In: *Foundations and Trends in Information Retrieval* 2.3, S. 137–213. doi: 10.1561/1500000008, S. 154.

40 Siehe dazu Kapitel 3.3, S. 54.

Tabelle 6.1 Ergebnisse der beschriebenen Experimente mit *difangzhi-SLM*

	Modell	*#types*	gleichförmige *chronons*	Beschränkung von *types* auf *DHYDCD*	Namen	Ortsnamen	Zeitausdrücke	◆ *NLLR* *Accuracy* (%)	◆ *NLLR* *MAE* (Jahre)	◆ *NLLR * TE* *Accuracy* (%)	◆ *NLLR * TE* *MAE* (Jahre)
1	1-*grams*	23.437	-	-	-	-	-	54,2	42,69	49,1	50,69
2	1–2 (Abb. 6.1)	7.400.012	-	-	-	-	-	52,8	45,27	52,8	52,41
3	1–2 (Abb. 6.2)	238.707	-	x	-	-	-	58,8	41,86	57,4	46,19
4	1–2	265.721	-	x	x	-	-	59,7	41,54	56,9	47,02
5	1–2	232.704	x	x	x	-	-	58,8	42,09	57,4	47,37
6	1–3	315.794	-	x	x	-	-	60,2	41,29	56,5	49,1
7	1–3	268.623	x	x	x	-	-	60,2	41,5	57,4	48,15
8	1–2	239.701	-	x	-	x	-	59,7	41,74	56,5	46,54
9	1–2 (Abb. 6.3)	238.978	-	x	-	-	x	60,6	**40,33**	**64,4**	42,54
10	1–2	246.890	-	+ 1-*grams*	-	-	x	59,7	**39,84**	60,6	42,65
11	1–2	266.954	-	x	x	x	x	59,7	41,36	61,1	45,4
12	1–2 (min. 2 *tokens*)	147.402	-	x	-	-	x	57,9	41,1	62	43,32
13	1–3	250.537	-	x	-	-	x	60,6	40,33	64,4	42,29
14	2-*grams*	223.452	-	x	-	-	x	62	42,18	61,1	43,94

2. Dirichlet *smoothing*:

$$P(w,\theta) = \frac{|Q|}{(|Q| + \mu)} \times P(w,Q) + \frac{\mu}{(\mu + |Q|)} \times P(w,C)$$

3. Jelinek-Mercer, auch *linear interpolation smoothing*[41] genannt:

$$P(w,\theta) = (1 - \beta) \times P(w,Q) + \beta \times P(w,C)$$

4. *Nearest neighbour smoothing*. Kanhabua und Nørvåg schlagen vor, eine Interpolation der Trainingsdaten dahingehend vorzunehmen, dass *unseen events* innerhalb eines *chronon* durch Daten aus benachbarten *chronons* ergänzt werden, je nach Datenlage durch Verwenden der niedrigsten Häufigkeit aus benachbarten *chronons* oder durch Verwendung des Durchschnittswertes aus Häufigkeiten der benachbarten *chronons*.[42] Diese Art der Interpolation bezeichne ich als *nearest neighbour smoothing*. Die Verwendung des Mittelwertes der benachbarten *chronons* scheint die Verwendung der Gewichtung durch *TE* absolut unbrauchbar zu machen (Experiment Nr. 5 in Tabelle 6.2). Ein Zusammenhang besteht sicherlich mit der stark unterschiedlichen Anzahl *types* in den *chronons*, wodurch die Häufigkeiten aus Nachbar-*chronons* über- oder untergewichtet werden können. Eine bessere Annahme scheint daher zu sein, dass die Häufigkeit eines *unseen events* die jeweils niedrigste aus den vorangehenden und nachfolgenden *chronons* ist (Experiment Nr. 4 in Tabelle 6.2). Allerdings kann auch so die *Accuracy* mit *NLLR * TE* und *LCM smoothing* nicht übertroffen werden.

Tabelle 6.2 gibt einen Überblick über die Ergebnisse.

Tabelle 6.2 Test von *Smoothing*-Methoden mit unterschiedlichen Parametern

			◆ *NLLR*		◆ *NLLR * TE*	
	Smoothing-Methode	Parameter	*Accuracy* (%)	*MAE* (Jahre)	*Accuracy* (%)	*MAE* (Jahre)
1	*Baseline*: Largest *chronon* minimum (Abb. 6.3)	$\lambda = 0{,}5$	60,6	**40,33**	**64,4**	42,54
2	Laplace (»add one«)	$\lambda = 1$	59,7	40,93	62	42,19
3	Laplace	$\lambda = 0{,}5$	60,6	41	12	200
4	Neighbour minimum & LCM	$\beta = 0{,}5$	**61,1**	40,63	55,6	48,36
5	Neighbour mean & LCM	$\beta = 0{,}5$	61,1	40,72	12,5	199,79
6	Jelinek-Mercer	$\beta = 0{,}01$	57,9	41,57	56,9	45,93
7	Jelinek-Mercer	$\beta = 0{,}1$	60,2	40,65	60,2	44,52
8	Jelinek-Mercer	$\beta = 0{,}4$	57,9	41,57	56,9	45,93
9	Jelinek-Mercer	$\beta = 0{,}9$	27,3	55,99	36,1	67,47
10	Dirichlet	$\mu = 0{,}1$	52,8	44,6	50,5	48,8
11	Dirichlet	$\mu = 0{,}4$	54,6	43,17	51,4	48,21
12	Dirichlet	$\mu = 0{,}9$	54,6	43,02	51,9	47,6

41 Siehe Kraaij 2004, S. 209.

42 Siehe Kanhabua und Nørvåg 2008, S. 362; Die Konsultation benachbarter *chronons* wird auch von A. Kumar 2013, S. 52, vorgeschlagen. Weder Kumar noch Kanhabua u. Nørvåg führen allerdings weitergehende Experimente mit dieser Art der Interpolation durch. Erst Bamman et al. 2017, verwenden erfolgreich ein *moving average* mit Daten der benachbarten *chronons*, wobei allerdings mit einem deutlich umfassenderen Korpus gearbeitet werden kann und alle Häufigkeiten durch diese Glättung angepasst werden. (Siehe S. 4.)

Fazit. Gegenüber der einfachen Annahme, die für das *LCM Smoothing* getroffen wird, lassen sich innerhalb dieses Korpus durch die Anwendung aufwändigerer *Smoothing*-Methoden keine Performanceverbesserungen bei der Verwendung statistischer Sprachmodelle zur Datierung erzielen. Die Häufigkeit von im Vergleichs-*chronon* nicht vorkommenden Lexemen sollte leicht unter der minimalen Häufigkeit liegen, die in einem *chronon* möglich ist. Dies bestätigt die ursprüngliche Annahme von de Jong, Rode und Hiemstra, dass die *Smoothing*-Effekte gering sind,[43] zeigt aber in einigen der durchgeführten Experimente auch, dass ungeeignete Glättungsmaßnahmen die Aussagekraft der Modelle massiv verschlechtern können.[44] In ihrem viel beachteten Aufsatz „An Empirical Study of Smoothing Techniques for Language Modeling" schreiben Chen und Goodman:

> Whenever data sparsity is an issue, smoothing can help performance, and data sparsity is almost always an issue in statistical modelling. In the extreme case where there is so much training data that all parameters can be accurately trained without smoothing, one can almost always expand the model, such as by moving to a higher-order *n*-gram model, to achieve improved performance. With more parameters, data sparsity becomes an issue again, but with proper smoothing the models are usually more accurate than the original models. […][45]

Durch die Verwendung der über einen Zeitraum von je 50 Jahren aggregierten *chronon*-Modelle aus jeweils 50 Texten sind die Trainingsdaten hier tatsächlich relativ umfangreich. Da *DFZ* nur als 1–3-Gramm Zählungen vorliegen, ist eine Erweiterung des „echten" *n*-Gramm Raumes nur sehr bedingt möglich, da keine Daten über die Reihenfolge des Auftretens der 2–3-Gramme vorhanden sind. Dass hier mit keiner der getesteten *Smoothing*-Techniken eine Verbesserung der Ergebnisse zu erzielen war, deutet an, dass durch das *Smoothing* regelmäßig in höherem Maß temporale Diskriminatoren „weggeglättet" werden, als durch die Glättung eine Verbesserung des Modells stattfindet. Es bleibt allerdings zu erforschen, wie eine Glättung sich bei der Verwendung deutlich kürzerer bzw. kleinerer *chronons*, oder sogar bei der Verwendung des „ähnlichsten" Dokuments zur Datierung (*document co-dating*) auswirken würden.

Notizen zu Berechnungen bei der Verwendung statistischer Sprachmodelle

— 1. ***Bag of words*** (*BoW*). Wie bereits angedeutet handelt es sich bei den hier als „Lexem-Modelle" bezeichneten Textrepräsentationen letztlich nicht um eine klassische *BoW*,[46] da bei Betrachtung von Wörtern mit einer Länge von 1–*n* Zeichen fast alle Vorkommen von Wörtern mit 2 oder mehr Zeichen Länge zusätzlich als Vorkommen der enthaltenen Einzelzeichen gezählt werden. Versuche, dies z. B. durch entsprechende Abzüge von Vorkommen längerer Wort-*n*-Gramme von den Unigramm-Zählungen auszugleichen, um eine stärkere Annäherung an eine tatsächliche *bag of **words*** zu erreichen, wirken sich aber tendenziell negativ auf die Leistungsfähigkeit der oben beschriebenen Modelle aus. Durch die Betrachtung von 2–*n*-Grammen wird der angesprochene Effekt allerdings ohnehin minimiert.

— 2. **Logarithmen**. Bei der Berechnung sowohl von *idf* und *TE* als auch von Ähnlichkeitsmaßen wie der *NLLR* werden zur Normalisierung Logarithmen verwendet. In der wissenschaftlichen Fachliteratur wird das Symbol *log* austauschbar sowohl für den *logarithmus naturalis* (*ln*) mit Basis

43 Siehe de Jong, Rode und Hiemstra 2005, S. 3.
44 Vgl. auch Chen und Goodman 1998, S. 59.
45 Ebd., S. 58.
46 Siehe auch Kapitel 4.5.3, S. 94.

e, als auch den Logarithmus mit Basis 2 (log_2) oder sogar 10 (log_{10}) verwendet. Obwohl „the precise base of the logarithm is not material to ranking“,[47] konnten oben durch Verwendung von log_2 für *NLLR* und *KLD* die eindeutig besten Ergebnisse erzielt werden.

— 3. ***Term frequency.*** Die *natürliche* bzw. rohe Worthäufigkeit $f_{w,d}$ ist definiert als Anzahl der Vorkommen von w in einem Dokument d.[48] Es ist offensichtlich, dass bei der Arbeit mit einem Korpus aus Dokumenten unterschiedlicher Länge eine Normalisierung erfolgen muss. In der Fachliteratur sind unterschiedliche Varianten einer solchen normalisierten *term frequency* verbreitet.

— 3.1 Normalisierung auf das häufigste Wort des Dokuments. $tf_{cmax}(w,d) = a + (1-a)\frac{f_{w,d}}{f_{max,d}}$, wobei a zwischen 0 und 1 gewählt werden kann und gewöhnlich auf 0,4 oder 0,5 gesetzt wird.[49] Mit $a = 1$ gilt dann $tf_{cmax}(w,d) = \frac{f_{w,d}}{c_{max,d}}$.

— 3.2 Normalisierung auf die Länge des Dokuments. Hierbei ist $|d|$ definiert als die Anzahl aller *tokens* in d.[50]

$$tf(w,d) = \frac{f_{w,d}}{|\,d\,|}$$

— 3.3 Zusätzlich kann die Größe des verfügbaren Wortschatzes $|\,V\,|$ betrachtet werden, also die Anzahl der unterschiedlichen *types* in d oder eines Korpus C,[51] z. B.:

$$tf_{vocsize}(w,d) = \frac{f_{w,d}}{|\,d\,| + |\,V\,|}$$

In einer experimentellen Überprüfung der Wirkung unterschiedlicher Interpretationen der *term frequency* auf die *Accuracy* des besten Modells (1–2 Zeichen Lexeme + temporale Ausdrücke) stellt sich der Unterschied zwischen den letzten beiden Varianten $tf_{vocsize}$ und tf erwartungsgemäß als marginal heraus, Die tf_{cmax} hingegen eignet sich nicht für Berechnungen mit *NLLR* oder *KLD*.[52]

— 4. ***Inverse document frequency (idf).*** Wie bei der tf gibt es auch für die zur Gewichtung verwendete inverse Dokumentenhäufigkeit idf zahlreiche Möglichkeiten der Berechnung. Buck und Koehn geben ohne Anspruch auf Vollständigkeit sechs unterschiedliche Definitionen an und bemerken, dass überdies „in der freien Wildbahn noch geringfügige Variationen dieser Definitionen zu finden sind.“[53] Aus Gründen der Einfachheit wird hier lediglich die in Kapitel 3.3 implizierte Variante verwendet:

$$idf_{w,c} = \log_2\left(\frac{N}{df_w}\right)$$

47 Christopher D Manning, Prabhakar Raghavan und Hinrich Schütze 2008: *Introduction to Information Retrieval*. Cambridge & New York: Cambridge University Press, S. 109.
48 Siehe z. B. ebd., S. 107.
49 Siehe z. B. ebd., S. 117.
50 Vgl. z. B. Zhai Chengxiang 2008, S. 145.
51 Vgl. z. B. Chen und Goodman 1998, S. 8.
52 Die *Accuracy* der *NLLR-/KLD*-Datierung sinkt von über 60 auf 21.8, während sie bei Verwendung von *CS* gleich mittelmäßig bei 38,9 bzw. 44,4 mit tf_{cmax} – idf bleibt.
53 Christian Buck und Philipp Koehn 2016: „Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance“. In: *Proceedings of the First Conference on Machine Translation, Berlin, Germany, August 11-12, 2016*. Bd. 2. Berlin: Association for Computational Linguistics, S. 672–678, S. 674, übersetzt durch den Verfasser.

— 5. ***Cosine similarity (CS)* und *tf-idf*.**[54] Zur Gewichtung der *tf* für die Berechnung der *CS* liefert bei der Verwendung von *DFZ-chronon*-Sprachmodellen der Vergleich der Vektoren mit *idf*-gewichteten *tf* des *chronon* mit denen der ungewichteten *tf* des zu datierenden Dokuments eine höhere *Accuracy* und einen niedrigeren *MAE*.[55] Für Modelle auf Dokumentebene ist dies nicht der Fall. Hier ist es erforderlich, die Worthäufigkeiten des zu datierenden Dokuments gleichermaßen zu gewichten.[56]

— 6. **Temporale Entropie (*TE*).** Für die Berechnung der *TE* werden hier die relativen Häufigkeiten *tf*, also $P(w, C) = \frac{f(w,c)}{|C|}$ usw. und nicht die absoluten Vorkommen verwendet. Die so berechneten Gewichte liegen zwar nicht zwischen 0 und 1, die Ergebnisse sind aber deutlich besser. Dies hängt ebenfalls mit der stärkeren Überbewertung von *Hapaxen* zusammen, die mit absoluten Häufigkeiten immer das maximale Gewicht von 1 erhalten.

— 7. ***NLLR* vs. *KLD*.** Ein nennenswerter Unterschied in der Performance ist nicht feststellbar. Kraaij bemerkt: „for ad hoc search, KL(Q|D) is essentially equivalent to the length normalized query likelihood [...] since the query entropy [...] is a constant which does not influence document ranking."[57]

6.1.2 Co-Datierung von Dokumenten

Wie unter anderem von de Jong, Rode und Hiemstra (2005) und Bamman et al. (2017) vorgeschlagen,[58] kann alternativ zur Verwendung aggregierter *chronons* dem zu datierenden Dokument der Zeitstempel des ähnlichsten Dokuments aus den Trainingsdaten zugewiesen werden.

Hierfür wird ein neues Korpus-*SLM* aus den einzelnen Trainingsdokumenten des *DFZ*-Datensatzes erzeugt. Aus Gründen der Vergleichbarkeit werden hierfür dieselben Texte aus dem Zeitraum 1475–1925 verwendet, wie für die Erzeugung der *chronon* Sprachmodelle in Abschnitt 6.1.1.[59] Das Modell enthält die relativen Häufigkeiten der 1–2 Zeichen-Lexeme und temporalen Ausdrücke von 772 Einzeltexten aus dem Zeitraum 1475–1925 (ursprünglich 17 *chronons*).

Anstatt wie oben den Zeitstempel des ähnlichsten *chronons* zu vergeben, wobei mit *NLLR* eine *Accuracy* von 60,6 % bei einem *MAE* von 40,3 Jahren erreicht wurde, wird hier der Zeitstempel des ähnlichsten Dokuments vergeben und zur besseren Vergleichbarkeit der Zeitraum des entsprechenden *chronons* genutzt. Die Datierung der 216 Texte des Testdatensatzes mit Jaccard *similarity*, *CS* mit und ohne *idf*-Gewichtung und *NLLR* mit einem *LCM-smoothing* mit $\lambda = 0,5$ wird mit den entsprechenden Ergebnissen der bereits durchgeführten Datierung mit

54 Siehe dazu auch Kapitel 3.3, v. a. S. 51 u. S. 53.

55 Bei Betrachtung von 1–2 Gramm Lexemen wird so eine *Accuracy* von 47,7 % und ein *MAE* von 51,9 Jahren erzielt (S. 161). Werden die *term frequencies* des zu datierenden Dokuments ebenfalls *idf*-gewichtet, reduziert sich die *Accuracy* auf 45,4 %, bei einem *MAE* von 58,4 Jahren. Beide Varianten liefern bessere Ergebnisse als die ungewichtete *CS*.

56 Siehe Abschnitt 6.1.2, S. 170. Bei Betrachtung von 1–2-Gramm Lexemen wird so eine *Accuracy* von 42,1 bei einem *MAE* von 60,9 Jahren erzielt. Die Verwendung von *CS* als Vergleich der ungewichteten *tf* des zu datierenden Dokuments mit den gewichteten *tf* der Vergleichsdokumente liefert eine *Accuracy* von 3,2 %, bei einem *MAE* von 224 Jahren.

57 Kraaij 2004, S. 208.

58 Siehe Kapitel 3.3, S. 50; siehe auch de Jong, Rode und Hiemstra 2005, S. 7; Bamman et al. 2017, S. 4.

59 Siehe S. 158.

chronon-Sprachmodellen verglichen (Tabelle 6.3).[60] Die Laufzeit der Berechnung beträgt etwa das 45-fache derjenigen bei Verwendung aggregierter *chronons*.

Tabelle 6.3 Ergebnisse mit *DFZ* Co-Datierung vs. *chronon-SLM*

				beschränkt auf		◆ *NLLR*		◆ *CS*		◆ *CS*tf-idf*	
	Modell	*#types*	*Smoothing*	*DHYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 / *chronons*	238.978	$\lambda = 0{,}5$	x	x	**60,6**	**40,33**	38,9	59,94	44,4	52,21
2	1–2 / *documents*	238.978	$\lambda = 0{,}5$	x	x	46,3	59,29	40,7	62,2	42,1	60,93

Tabelle 6.3 (Fortsetzung)

				beschränkt auf		◆ *Jaccard*		◆ *Random*	
	Modell	*#types*	*Smoothing*	*DHYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*
1	1–2 / *chronons*	238.978	$\lambda = 0{,}5$	x	x	13	117,97	10,6	142,32
2	1–2 / *documents*	238.978	$\lambda = 0{,}5$	x	x	39,4	66,78	15,7	139,43

Auch bei Verwendung eines dokumentenbasierten Modells lässt sich mit *NLLR* die höchste *Accuracy* (46,3 %) und der niedrigste *MAE* (59,3 Jahre) erzielen. Die Jaccard *similarity* gewinnt an Aussagekraft, übertrifft die Performance der komplexeren Metriken, die auch Worthäufigkeiten berücksichtigen, jedoch nicht. Insgesamt bleiben die Ergebnisse jedoch deutlich hinter denen der *chronon*-Datierung zurück.

Für die Entscheidung zwischen aggregierten *chronon*-Modellen und einem Direktvergleich von Dokumenten können neben der gewünschten Granularität der zu vergebenden Zeitstempel auch praktische Erwägungen eine Rolle spielen: die Verfügbarkeit geeigneter Trainingsdaten, sowie die Art der zu datierenden Texte.

Eine Wiederholung des Experiments mit dem *XXSKQS*-Datensatz mit 1–3-Grammen schriftsprachlicher chinesischer Texte,[61] deutet an, dass mit Co-Datierung unter Umständen ähnlich gute Ergebnisse erzielt werden können wie mit *chronons*. Im Fall des *XXSKQS* hängt dies sicherlich mit der heterogenen Natur des Korpus zusammen.[62]

Aus dem Datensatz wird je ein *chronon*-Sprachmodell und ein Co-Datierungs-Sprachmodell mit 1–2 Zeichen-Lexemen, Namen und Zeitausdrücken erzeugt. Für beide Modelle werden dieselben 717 Texte als Trainingsdaten verwendet und ein Testdatensatz mit 176 Texten zufällig ausgewählt.[63] Wie bereits in Abschnitt 6.1.1 werden Texte aus dem Zeitraum von 1475–1925 berücksichtigt. Um Defizite in den zur Verfügung stehenden Metadaten auszugleichen, werden bei der Auswahl der Trainings- und Testdaten nur Texte ausgewählt, bei denen die Namen der Verfasser:innen (mit *zhuan* 撰, „zusammenstellen, verfassen, komponieren“) in den Metadaten

60 Die Berechnung von *TE* zur Gewichtung der *features* für die Berechnung von *NLLR* oder *KLD* erscheint bei der Betrachtung einzelner Dokumente wenig intuitiv. Denkbar wäre aber ein hybrider Ansatz, der auf die *TE* der entsprechenden *chronons* zurückgreift.

61 *XXSKQS*, siehe auch Kapitel 4.2, S. 68, siehe auch Abschnitt 6.1.3, v. a. S. 174.

62 Siehe Kapitel 4.2, S. 68.

63 Es wurde eine ausgewogene Aufteilung der Testdaten auf die *chronons* angestrebt. Für die Zeiträume 1475–1525, sowie 1675–1725 stehen aber wg. der Auswahlkriterien und Trainingsdaten nur noch einzelne Texte zur Verfügung. Daher muss der Testdatensatz kleiner ausfallen als bei den *DFZ*.

angegeben sind und die Dynastie der Veröffentlichung zu deren biographischen Daten passt.[64] Tabelle 6.4 zeigt die Ergebnisse bei der Datierung der 176 Texte aus den Testdaten mit *chronon*- und Co-Datierungs-Modellen im direkten Vergleich. Es wurde jeweils ein *LCM Smoothing* mit $\lambda = 0,5$ angewendet und *DHYDCD*-Lexeme, Namen und Zeitausdrücke berücksichtigt.

Tabelle 6.4 Ergebnisse mit *XXSKQS* Co-Datierung vs. *chronon-SLM*

			◆ *NLLR*		◆ *CS*		◆ *CS*tf-idf*		◆ *Jaccard*		◆ *Random*	
	Modell	*#types*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 / *chronons*	267.349	23.3	98.11	19,9	108,52	**33,5**	**80,76**	11,9	103,81	11,9	141,54
2	1–2 / *documents*	267.349	21.6	101.32	25	94	27,3	94,26	**29,5**	**86,17**	11,4	142,84

Insgesamt bleiben die Ergebnisse erwartungsgemäß weit hinter denen der *DFZ*-Experimente zurück, da die *XXSQKS* als diachrones Korpus deutlich problematischer sind.[65] Erneut lassen sich mit einem aggregierten *chronon*-Modell bessere Ergebnisse erzielen. Die Diskrepanz zwischen *chronon*- und dokumentenbasiertem Modell ist aber deutlich kleiner.

Die höchste *Accuracy* (33,5 %) und der niedrigste *MAE* (80,8 Jahre) werden mit ersterem bei Verwendung von *CS* mit *tf-idf*-Gewichtung erreicht. Ein vergleichsweise niedriger *MAE* (86,2) kann jedoch auch mit Co-Datierung und Jaccard *similarity* erreicht werden. Die Zuordnung erfolgt dabei unabhängig von Worthäufigkeiten allein aufgrund einer Überschneidung der vorkommenden Lexeme. Die Erfolgschancen einer dokumentenbasierten Datierung dürften stärker als bei der *chronon*-Datierung davon abhängen, dass dem zu datierenden Dokument inhaltlich und damit im Wortschatz ähnliche Texte in den Trainingsdaten vorhanden sind. Insgesamt sollte es damit wenig Anwendungsfälle geben, in denen eine dokumentenbasierte Vorgehensweise vorzuziehen ist, die überdies deutlich höhere Laufzeiten mit sich bringt.

6.1.3 Datierung mit *DHYDCD*-Sprachmodellen

Mit aus dem *DFZ*-Korpus generierten Sprachmodellen lassen sich mit *NLLR* und *TE*, einer vereinfachten *Smoothing*-Methode und der Einschränkung auf Lexeme und temporale Ausdrücke bereits eine *Accuracy* von über 60 % bei einem *MAE* von etwa 40 Jahren erreichen. Die Datierung ist dabei jedoch durch die Trainingsdaten auf den Zeitraum vom 15. bis zum Anfang des 20. Jahrhunderts beschränkt. Auch für die Datierung von Texten außerhalb des vorgegebenen Genres sind die Erfolgsaussichten als gering einzustufen. Um die ungefähre zeitliche Einordnung von Texten für die gesamte chinesische Schrifttradition und über Genregrenzen hinweg zu ermöglichen, verwende ich als Trainingskorpus die datierbaren Textzitate aus dem *DHYDCD*.[66] Aus den 53 Segmenten dieses Behelfskorpus mit repräsentativem Textmaterial unterschiedlicher Genres aus dem Zeitraum von ca. 700 v. u. Z. bis zum 20. Jh. können temporale Sprachmodelle mit einer Auflösung von 100 Jahren und einer Überlappung von 50 Jahren berechnet werden.[67] 700–

64 Das *XXSKQS* enthält zahlreiche spätere Ausgaben, bei denen der Name des Herausgebers bzw. eines Kommentators angegeben ist. Solche Werke aufgrund ihrer *n*-Gramm-Häufigkeiten zu datieren ist aussichtslos, da Textmaterial aus unterschiedlichen Zeiträumen in unterschiedlichen Anteilen untrennbar vermischt ist. Texte mit fragwürdigen Angaben (siehe dazu auch Kapitel 4.2, ab S. 67.) werden ebenfalls ausgeschlossen.

65 Datierungsergebnisse mit dem *XXSQKS*-Korpus werden in Abschnitt 6.1.3, ab S. 175, ausführlicher diskutiert.

66 Siehe Kapitel 5.6, ab S. 137.

67 Für eine kürzere *chronon*-Dauer reicht die Genauigkeit der Datierung der zugrunde liegenden Daten leider nicht aus. Vgl. auch Kapitel 5.7, S. 139.

600 v. u. Z. ist also das erste, 650–550 v. u. Z. das zweite usw. und 1900–2000 das letzte *chronon* solcher Modelle.

Die aus den in Abschnitt 6.1.1 (ab S. 158) beschriebenen Experimenten als am effektivsten hervorgegangenen Herangehensweisen werden mit diesen *DHYDCD*-Sprachmodellen erprobt. Hierfür werden — 1. Die 25 offiziellen Dynastiegeschichten[68] gegen die Textbeispiele aus dem *DHYDCD* datiert. — 2. dieselben zufälligen 216 Texte aus dem *difangzhi* 地方誌-Korpus datiert. — 3. Da keine Trainingsdaten aus dem *Difangzhi*-Datensatz benötigt werden, kann dieser über einen größeren Zeitraum von 1300–1925 genutzt werden. — 4. Um eine Datierung von Texten mit dem *DHYDCD*-Sprachmodell über Genregrenzen hinweg zu testen, werden überdies zufällige Texte aus dem *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書 datiert.

Als *Baseline* dient erneut ein Zufallsgenerator, der – bei 53 überlappenden *chronons* – Texte mit einer Wahrscheinlichkeit von etwa 4 % korrekt zuordnet. Eine Co-Datierung von Texten wie in 6.1.2 ist so weder sinnvoll noch möglich, da die aggregierten Trainingsdaten keine konkreten Texte mehr repäsentieren. Ebensowenig erfolgt eine Unterteilung des Korpus in Test- und Trainingsdaten.[69]

Experiment 1: *zhengshi* 正史

Im ersten Test der aus dem Zitatmaterial des *DHYDCD* erzeugten *SLM*s werden die 25 offiziellen Dynastiegeschichten zugeordnet. Die Sprachmodelle können so mit einem Referenzkorpus getestet werden, dessen Texte aus einem sehr großen Zeitraum von 2.019 Jahren, von ca. 91 v. u. Z. bis 1928, stammen. Dabei darf nicht vergessen werden, dass ein großer Teil der *zhengshi* als Belegstellen im *DHYDCD* häufig bis sehr häufig vertreten ist[70] und dadurch bedingt eine hohe Übereinstimmung an *types* zwischen den zu datierenden Texten und den jeweils korrekten *chronons* besteht. Die Testreihe ist also teilweise inzestuös. Sie kann aber Aufschluss über die Eignung der unterschiedlichen Metriken und die geeignete Größe des *Smoothing*-Parameters λ für das *LCM Smoothing* geben. Da die *zhengshi* als Volltext vorliegen, besteht keine Beschränkung des *n*-Gramm-Raums auf eine Länge von 3 Zeichen mehr, so dass auch die Verwendung von Lexemen von 1–4 Zeichen Länge geprüft werden kann.[71]

Tabelle 6.5 Ergebnisse mit *zhengshi* und mit *DHYDCD-SLM*s

				beschränkt auf		◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Jaccard*	
	Modell	*#types*	λ	*HYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 *grams*	1.992.349	0,90	-	-	84	**88,68**	88	65,68	68	67,28	64	95,64
2	1–2 *grams*	1.992.349	0,50	-	-	80	77,04	80	**59,16**	68	67,28	64	95,64
3	1–2 *grams*	1.992.349	0,10	-	-	68	72,64	76	59,52	68	67,28	64	95,64
4	1–2 *grams*	1.992.349	0,01	-	-	60	87,28	76	59,52	68	67,28	64	95,64
5	1–2	228.191	0,90	x	-	84	88,68	80	72,4	60	77,96	64	90,48
6	1–2	228.191	0,50	x	-	80	87,04	76	62,76	60	77,96	64	90,48
7	1–2	228.191	0,10	x	-	80	78,24	76	60,76	60	77,96	64	90,48
8	1–2	228.191	0,01	x	-	68	73,84	72	62,32	60	77,96	64	90,48
9	1–2	228.438	0,90	x	x	84	86,68	76	73,96	60	77,96	64	90,48

68 Siehe dazu Kapitel 2.3, ab S. 20.

69 Theoretisch möglich wäre die Erzeugung von Pseudotext aus den Belegstellen jedes einzelnen zitierten Texts – was allerdings in den wenigsten Fällen eine ausreichende Menge an Sprachmaterial ergäbe.

70 Siehe dazu auch Kapitel 5.7.4, ab S. 150

71 Aufgrund ihrer geringen Zahl ist die Berücksichtigung von Wörtern mit einer Länge von 5+ Zeichen wenig aussichtsreich. Siehe dazu Kapitel 5.7.3, ab S. 146 und Kapitel 4.5.2, S. 92.

Tabelle 6.5 (Fortsetzung)

				beschränkt auf		◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Jaccard*	
	Modell	*#types*	λ	*HYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
10	1–2	228.438	0,50	x	x	80	87,04	76	62,76	60	77,96	64	90,48
11	1–2	228.438	0,01	x	x	68	73,84	72	61,12	60	77,96	60	97,8
12	1–3	240.601	0,90	x	x	84	86,68	76	73,96	60	77,96	68	82,68
13	1–4	256.800	0,90	x	x	84	86,68	76	73,96	60	77,96	68	82,68

1. Die beste *Accuracy* von 88 wird bei Verwendung aller 1–2-Gramme mit *NLLR*TE* und einem *Smoothing*-Parameter von $\lambda = 0,9$ erreicht, der geringste *MAE* von 59,5 Jahren mit einem niedrigeren *Smoothing*-Faktor – auf Kosten einer gesunkenen *Accuracy*.
2. Auch wenn aus den Ergebnissen für die nur 25 hier datierten Texte keine voreiligen Schlüsse gezogen werden sollten, scheint ein hoher Wert von λ eine tendenziell bessere *Accuracy* und ein niedriger Wert einen geringeren *MAE* zu ermöglichen. Mit $\lambda = 0,9$ werden also mehr Texte korrekt datiert, mit $\lambda = 0,01$ weniger starke Abweichungen erzielt.
3. Durch Eingrenzung der betrachteten *n*-Gramme auf Lexeme ergibt sich erwartungsgemäß eine drastische Steigerung der Verarbeitungsgeschwindigkeit, da nur etwas mehr als 10 % der *types* genutzt werden. Die Ergebnisse der Datierung verschlechtern sich dabei marginal.
4. Eine Erweiterung der Lexeme um Zeitausdrücke hat hier kaum Auswirkungen, da die Trainingsdaten nur sehr wenige Vorkommen aufweisen.
5. Eine Vergrößerung des *n*-Gramm-Raums auf Lexeme mit 1–3 bzw. 1–4 Zeichen Länge bringt ebenfalls keine deutliche Verbesserung mit sich, da Wörter mit einer Länge von mehr als 2 Zeichen einen zu geringen Anteil haben. Aufgrund der großen Anzahl an *zhengshi*-Zitaten im Korpus wird bei der Verwendung von Jaccard *similarity* ein geringer Effekt sichtbar.
6. Wieder ist *NLLR* den einfacheren Ähnlichkeitsmaßen überlegen, durch *TE*-Gewichtung wird eine etwas geringere *Accuracy* bei besserem *MAE* erzielt.

Experiment 2: 216 *Difangzhi* 地方誌
Im zweiten Versuch werden die für Abschnitt 6.1.1 zufällig für die Datierung mit *DFZ-SLM*s ausgewählten Texte erneut unter Verwendung der *DHYDCD SLM*s datiert. Es bestätigt sich, dass mit einem weniger spezialisierten Trainingskorpus bzw. einem längeren Betrachtungszeitraum zunächst deutlich schlechtere Ergebnisse erzielt werden können, als dies in Abschnitt 6.1.1 bzw. in Experiment 1 der Fall war. Die beste erreichte *Accuracy* von 16,2 % liegt deutlich unter der Performance bei Verwendung der *DFZ* selbst zur Erzeugung des temporalen Sprachmodells, ebenso wie der *MAE* von 140,6 Jahren.[72] Es sollte jedoch bedacht werden, dass nun längere *chronons* und ein sechsmal längerer Betrachtungszeitraum von 2.700 Jahren verwendet werden. In dieser Relation ist eine Datierung mit einer durchschnittlichen Genauigkeit von 140 Jahren als durchaus aussagekräftig anzusehen. Die genauen Ergebnisse von Experiment 2 sind in Tabelle 6.6 wiedergegeben.[73]

72 Dieselben Texte konnten in Abschnitt 6.1.1 zu 64 % korrekt bei einem *MAE* von etwa 40 Jahren datiert werden. Siehe S. 162.

73 Um einen möglichst geringen *MAE* zu erzielen, wurde für *unseen events* hier ein *LCM smoothing* mit $\lambda = 0,01$ angewandt.

Tabelle 6.6 Ergebnisse mit *Difangzhi* und *HYDCD-SLM*s.

				◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Random*	
	Modell	*#types*	Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 Gramme	1.992.349	-	16,2	152,92	15,7	141,92	**19,9**	**136,04**	5,1	1.072,06
2	1–2 字 Lexeme	228.191	-	15,3	165,97	16,2	141,45	6	322,31	4,6	1.134,55
3	1–3 字 Lexeme	240.601	x	16,2	158,53	16,2	140,79	17,1	202,69	4,6	1.058,95
4	1–2 字 Lexeme	228.438	x	15,7	157,45	**16,2**	**140,56**	17,1	210,7	4,2	1.165,05

Zudem deutet sich an, dass bei kargerer Datenlage *CS* mit *tf-idf* bessere Ergebnisse liefern kann als die komplexeren Metriken, allerdings nur bei Verwendung aller 1–2-Gramme – bei Reduktion der Dimensionen auf Lexeme und Zeitausdrücke sind die *Accuracy*-Unterschiede zur *NLLR* marginal. Zudem wird bei Verwendung von *NLLR* ein geringerer *MAE* erzielt, der durch Gewichtung mittels *TE* auf 140,56 Jahre reduziert werden kann.

Experiment 3: *Difangzhi*, 1300–1925

Der *DFZ*-Datensatz enthält Texte aus dem Zeitraum vom 8. bis zum Anfang des 20. Jhs. Erst ab dem 15. Jh. sind jedoch ausreichend Texte im Korpus, um temporale *SLM*s zu erzeugen. Als reine Testdaten können jedoch auch ältere Texte ab etwa 1300 eingesetzt werden. In Experiment 2 soll festgestellt werden, ob sich eine neue, zufällige Textauswahl aus dem längeren Zeitraum 1300–1925 ebenso gut datieren lassen wie die Texte aus Experiment 2 bzw. Abschnitt 6.1.1. Hierfür werden 1–2 Zeichen-Lexeme und temporale Ausdrücke verwendet und 119 Texte zufällig zur Datierung ausgewählt. Das Experiment wird viermal mit identischen Parametern wiederholt, so dass jeweils teilweise unterschiedliche Texte datiert werden.

Tabelle 6.7 Ergebnisse mit *Difangzhi* 1300–1925 und *HYDCD-SLM*s.

	◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Random*	
	Accuracy (%)	*MAE* (Jahre)	*A* (%)	*MAE* (Jahre)	*A* (%)	*MAE* (Jahre)	*A* (%)	*MAE* (Jahre)
1	14,3	162,98	20,2	133,9	21	214,95	4,2	1.071,46
2	14,3	167,12	17,6	135,76	22,7	218,22	5,9	1.044,2
3	10,1	174,71	16,8	137,5	18,5	227,52	4,2	1.002,95
4	9,2	163,97	15,1	136,32	15,1	198,79	4,2	990,39

Die Ergebnisse zeigen ein sehr ähnliches Bild wie in Experiment 2, die Wiederholungen geben zudem Aufschluss über die Schwankungsbreite bzw. Stabilität der Datierungsergebnisse. *NLLR* liefert einen konstant niedrigen *MAE*, der sich durch *TE*-Gewichtung auf ca. 135 Jahre reduzieren lässt. Mit *CS* und *tf-idf* lassen sich erneut tendenziell etwas mehr Texte einem korrekten *chronon* zuordnen, der *MAE* bleibt jedoch deutlich schlechter, als das mit *NLLR* der Fall ist. Die Erweiterung des Datierungszeitraums wirkt sich nicht nachteilhaft aus.

Experiment 4: *Xu xiu si ku quan shu* 續修四庫全書, 14. Jh.–1927

In Experiment 4 soll festgestellt werden, ob eine ungefähre Datierung mit den *DHYDCD-SLM*s auch über Genregrenzen hinweg möglich ist. Aus dem *N-gram dataset of Xu xiu si ku quan shu* 續修四

庫全書 wird ein Testdatensatz von 105 Texten aus dem Zeitraum vom 14. Jh. bis 1927 gleichmäßig verteilt zufällig ausgewählt.[74]

Die Texte werden mit *NLLR*, *NLLR*TE*, *CS*tf-idf* und einem Zufallsgenerator als Baseline datiert, für *NLLR* wird ein *LCM Smoothing* mit $\lambda = 0,01$ verwendet. Die Ergebnisse sind in Tabelle 6.8 aufgelistet.

Tabelle 6.8 Ergebnisse mit *Xu xiu si ku quan shu* und *HYDCD-SLM*s.

			◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Random*	
	Modell	Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 Gramme	-	21,9	323,79	21,9	298,92	21	346,67	3,8	1.135,09
2	1–2 字 Lexeme	-	21	358,49	23,8	284,37	13,3	390,55	5,7	1.140,45
3	1–2 字 Lexeme	x	21	354,81	23,8	284,37	13,3	386,74	3,8	1.009,33
4	1–3 字 Lexeme	x	21	354,81	22,9	286,45	13,3	384,36	2,9	1.104,76

Mit einer *Accuracy* von 23,8 % bei Verwendung von *NLLR* mit *TE*-Gewichtung ist hier sogar ein etwas größerer Anteil an Texten korrekt datiert als in den Experimenten 2 und 3 mit Texten aus dem *DFZ*-Datensatz. Der *MAE* ist allerdings mit fast 300 Jahren deutlich größer. Weder die Erweiterung des *n*-Gramm-Raums auf Lexeme mit bis zu drei Zeichen, noch die Berücksichtigung von Zeitausdrücken verursachen einen spürbaren Effekt. Werfen wir einen Blick auf die Detaildarstellung der 105 mit *NLLR*TE* datierten Texte, lassen sich schnell zwei wesentliche Probleme erkennen (Abb. 6.5).

— 1. Texte werden übermäßig häufig auf das *chronon* 1550–1650 datiert. Diese Präferenz ist kein Zufall, da für dieses *chronon* Daten zu mehr *types* zur Verfügung stehen, als dies bei den benachbarten *chronons* der Fall ist. Durch Verwendung „ausgewogener" Modelle kann diese Problematik zwar aufgelöst werden, der Datenverlust führt aber insgesamt zu einer deutlichen Verschlechterung der Ergebnisse.[75] — 2. Die Titel einiger stark zu früh datierter Texte (Abb. 6.5), z. B. *Sanguo zhi zhu bu* 三國志注補, (etwa: „Ergänzung zur kommentierten Ausgabe der Chroniken der Drei Reiche", 1644), *Shiji kaozheng* 史記考證 (etwa: „Untersuchungen über die Authentizität des *Shiji*", 1788), *Chuci yi* 楚詞譯 (etwa: „Interpretation der Elegien von Chu", 1901) usw., deuten darauf hin, dass es sich dabei um neue, kommentierte Ausgaben der jeweils im Titel genannten Texte (*Sanguo zhi*, Ende des 3. Jahrhunderts, *Shiji*, 1. Jh. v. u. Z., *Chuci*, ca. 3. Jh. v. u. Z. usw.) handelt, oder zumindest um Werke, die diese Texte in hohem Maße zitieren und damit einen sehr hohen Anteil an früherem Sprachmaterial enthalten. Texte wie das lt. Metadaten 1409 entstandene *Sheng xue xin fa* 聖學心法, datiert auf das *chronon* 400–300 v. u. Z, zeigen aber, dass auch stilistisch „alte" Texte für die linguistische Datierung ein großes Problem darstellen, vor allem, wenn sie kaum oder kein zeitgenössisches Vokabular enthalten.

In Experiment 4 werden zwar nur rund ein Viertel der Texte korrekt datiert und der *MAE* ist mit 284 Jahren sehr hoch, dennoch datiert ein Großteil der Texte ungefähr in den korrekten Zeitraum. Extremwerte falscher Datierungen sind teilweise auf die Natur des *XXSKQS* mit einem hohen Anteil an Texten, die frühere Texte ganz oder teilweise enthalten – z. B. kommentierte Ausgaben – zurückzuführen. Diese Art der Intertextualität, die bei traditionellen Herangehensweisen für die Datierung eines Textes hilfreich sein kann, denn „if text A cited text B, then text

74 Da hier ein größerer Zeitraum betrachtet werden kann als in Abschnitt 6.1.2 (ab S. 169), wird ein neuer Testdatensatz generiert.

75 Ohne Abb. Mit *NLLR * TE* wird noch eine *Accuracy* von $A = 11,4$ bei einem *MAE* von 433 Jahren erreicht. Detaillierte Versuche zur Verwendung gleichförmiger *chronons* in Abschnitt 6.1.1, S. 161 zeigen sehr ähnliche Verschlechterungen.

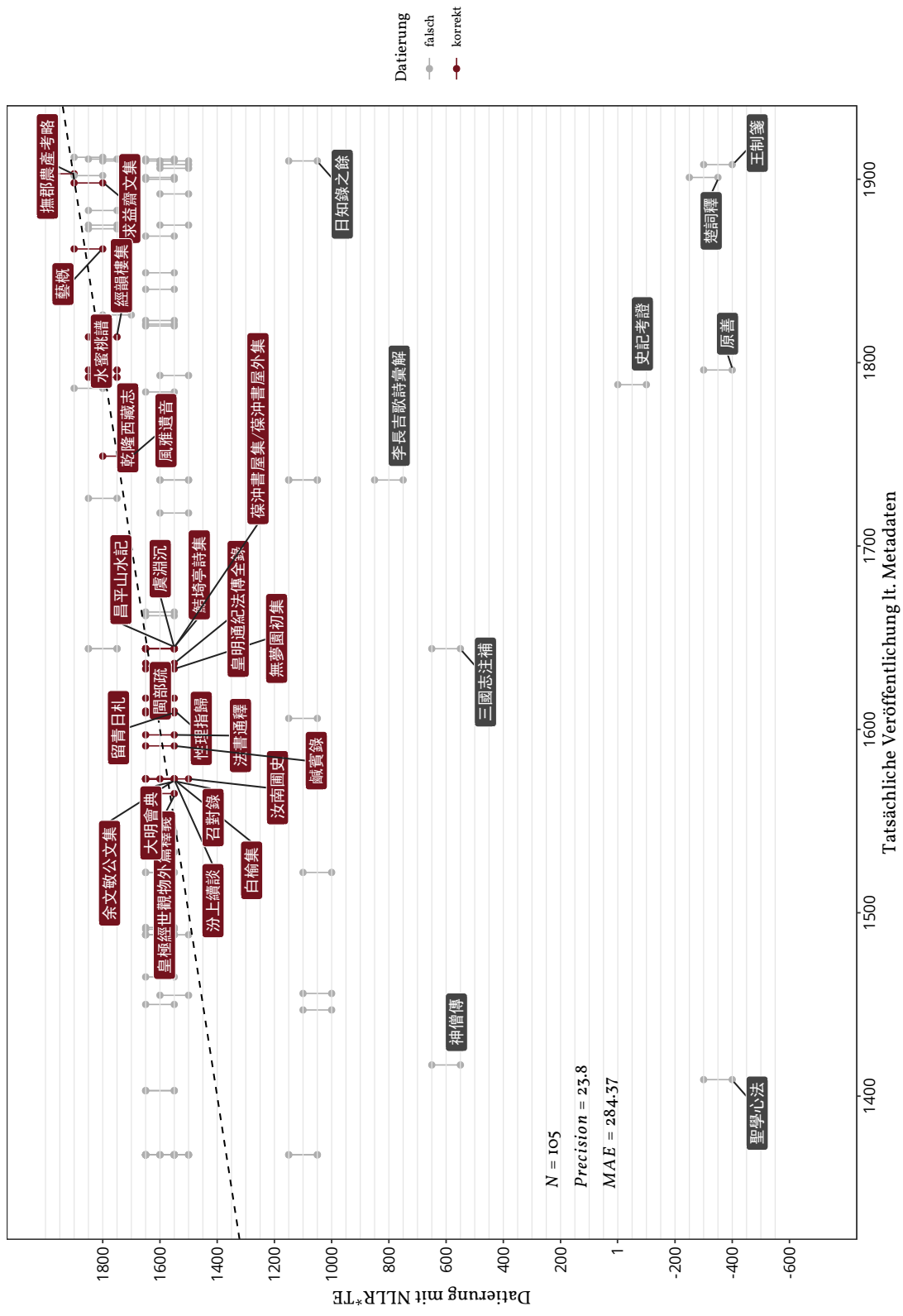


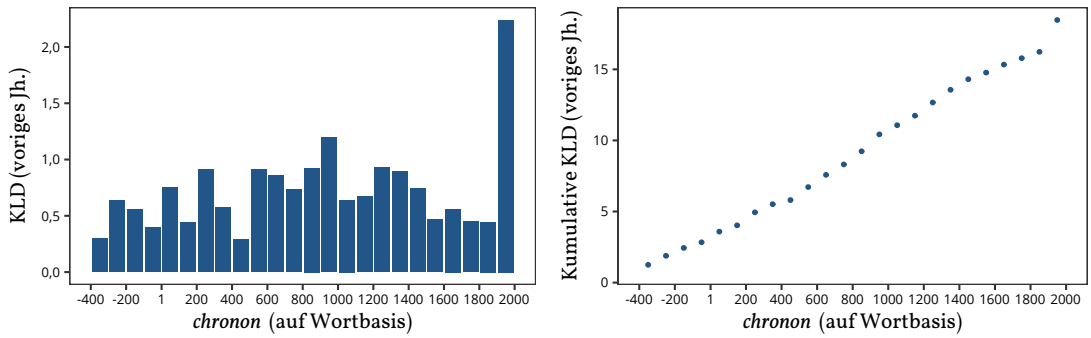
Abbildung 6.5 Experiment 4.4: 105 zufällige Texte aus dem *XXSKQS*

B must be older than text A"[76] ist für eine quantitative bzw. statistische Betrachtung, bei der Kommentar und Haupttext nicht unterschieden werden können, problematisch. Das Beispiel des Qing-zeitlichen *Yuan shan* 原善 (1796), hier dem *chronon* 300–200 v. u. Z. zugeordnet, zeigt, dass einige Texte resistent gegenüber einer statistischen Analyse sein dürften.[77]

Dass es bei Verwendung ausgewogener Modelle zu einer spürbaren Verschlechterung der *MAE*-Performance kommt, deutet auf einen für diese Zwecke wohl zu geringen Umfang der *DHYDCD*-Trainingsdaten hin. Die Datierungen sind dabei nicht abwegig, sondern schwanken um einen Zeitraum von mehreren hundert Jahren um die tatsächliche Entstehung der jeweiligen Texte. Diese Ungenauigkeit spiegelt die Problematik eines nur langsamen Sprachwandels des schriftsprachlichen Stils für Datierungsversuche auf Basis statistischer Sprachmodelle wider.

6.1.4 Sprachwandel im Sprachmodell

Wie die Experimente in 6.1.1 bis 6.1.3 gezeigt haben, können mit *SLM*s schriftsprachliche chinesische Texte zwar nicht exakt datiert werdem, mit *KLD* oder *NLLR* aber – abhängig von untersuchtem Material und Trainingsdaten – ungefähr zeitlich eingestuft werden. Durch Betrachtung des Unterschieds der einzelnen *chronon* Sprachmodelle voneinander, kann mit derselben Methodik auch die Veränderung des Wortgebrauchs als Aspekt des Sprachwandels quantifiziert werden.[78] Die Beobachtungen sollten dabei allerdings nur als grobe Richtung bzw. linguistische Trends betrachtet werden, da die Belegstellen aus dem *DHYDCD* ein stark begrenztes Textmaterial darstellen, das zudem ein *Bias* mit Präferenzen für bestimmte Texte und Textgattungen aufweist.[79]



(a) KLD, -100 Jahre

(b) KLD (kumulativ), -100 Jahre

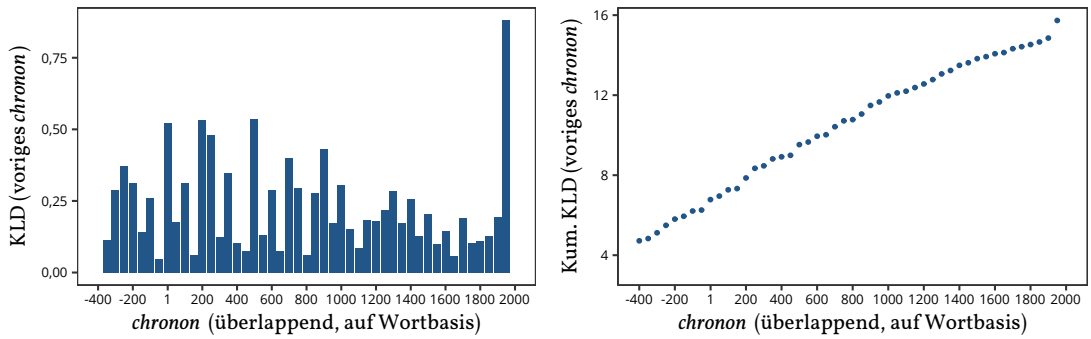
Abbildung 6.6 Kullback-Leibler-Divergenz zum *chronon* des jeweils vorangegangenen Jahrhunderts

Abbildung 6.6a stellt die Veränderung eines jeden *chronon* zum Modell des jeweils vorangegangenen Jahrhunderts dar, d. h. es wird z. B. die *KLD* des *chronon* 400–500 zum *chronon* 300–400 gemessen.[80] Eine außergewöhnlich starke Veränderung ist dabei vom 19. hin zum 20. Jh. sicht-

76 Toner und Han Xiwu 2019, S. 17–18. Siehe auch Kapitel 3, S. 39.
77 Der Text wird in Abschnitt 6.2.5 ausführlicher diskutiert, siehe S. 209.
78 Vgl. auch de Jong, Rode und Hiemstra 2005, S. 1.
79 Siehe dazu Kapitel 5.7, ab S. 138.
80 Die Kullback-Leibler-Divergenz bietet sich hier als sinnvolleres Maß an, da sie – im Gegensatz zur *NLLR* – den Unterschied zwischen zwei Sprachmodellen misst.

bar, in der sich wahrscheinlich unter anderem die Auswirkungen der Bewegung des 4. Mai (*wusi yundong* 五四運動, ab 1919) mit einer stärkeren Verschriftlichung der Umgangssprache (*baihua* 白話) widerspiegeln. Die kumulative Darstellung (Abb. 6.6b) lässt zudem eine langfristig weitgehend lineare Veränderung erahnen, die etwa ab dem 13. Jh. leicht abzuflachen scheint.

Werfen wir noch einen Blick auf die Veränderung ohne Auslassung der überlappenden *chronons* (Abb. 6.7). Dies suggeriert in der kumulativen Darstellung ebenfalls eine abflachende Kurve, innerhalb derer allerdings eine gewisse Zyklik erkennbar wird.[81] Zudem sind kleinere *s*-Formen zu sehen, die an das Piotrowski-Gesetz erinnern[82] – für eine belastbare Interpretation in diese Richtung wäre allerdings umfassenderes Datenmaterial erforderlich.



(a) KLD, -50 Jahre

(b) KLD (kumulativ), -50 Jahre

Abbildung 6.7 Kullback-Leibler-Divergenz zum vorigen *chronon*

Mit einer umgekehrten Jaccard *similarity* kann zudem die reine Übereinstimmung der in den Textbeispielen verwendeten *types* zum jeweils vorangegangenen Jahrhundert unabhängig von ihrer Häufigkeit gemessen werden.[83]

Auch in Abb. 6.8 bleibt eine stärkere Veränderung des verwendeten Wortschatzes im 20. Jh. sichtbar – es werden also nicht nur die vorhandenen Lexeme stark unterschiedlich häufig verwendet, sondern auch ein größerer Anteil *anderer* Lexeme als zuvor. In der kumulativen Darstellung deutet sich ebenfalls wieder eine schwache *s*-Form an.

Mit Ausnahme eines großen Sprungs im 20. Jh. lassen die Beobachtungen aus dem Zitatmaterial des *DHYDCD* auf eine langfristig betrachtet verhältnismäßig konstante Veränderung des Wortschatzes schließen; die Veränderungsrate der Wortnutzung scheint dabei einer gewissen Schwankung zu unterliegen. Da die verwendeten Trainingsdaten sehr begrenzt sind und die Auswahl der zugrunde liegenden Texte starke Unausgewogenheiten aufweist,[84] muss aber von

81 Bereits der Sinologe Hans Georg Conon von der Gabelentz hatte ein Modell des zyklischen Sprachwandels bzw. „Spirallaufs der Sprachgeschichte" eingeführt, das natürlich nicht belegt ist. Ihm geht es dabei primär um einen syntaktischen Sprachwandel. Hans Georg Conon von der Gabelentz 1901 [1891]: *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Hrsg. von Albrecht Graf von der Schulenburg. 2. vermehrte und verbesserte Auflage. Leipzig: Tauchnitz, S. 255.

82 Siehe dazu Kapitel 2.1, S. 14 und v. a. Kapitel 5.7, S. 146

83 Die Jaccard *similarity J* ist hier als Schnittmengenanteil der *types* zweier Wortlisten definiert (siehe S. 52), gibt also den Grad ihrer Übereinstimmung an. Um die Abweichung vom *chronon* des vorangegangenen Jahrhunderts zu messen, wird 1 – *J* verwendet. Dieses Maß bezeichne ich im Folgenden als |Jaccard-*Divergenz*.

84 Siehe dazu auch Kapitel 5.7.4, ab S. 150.

allgemeingültigen Aussagen über eine Zyklik oder Geschwindigkeit des Sprachwandels im Chinesischen Abstand genommen werden.[85]

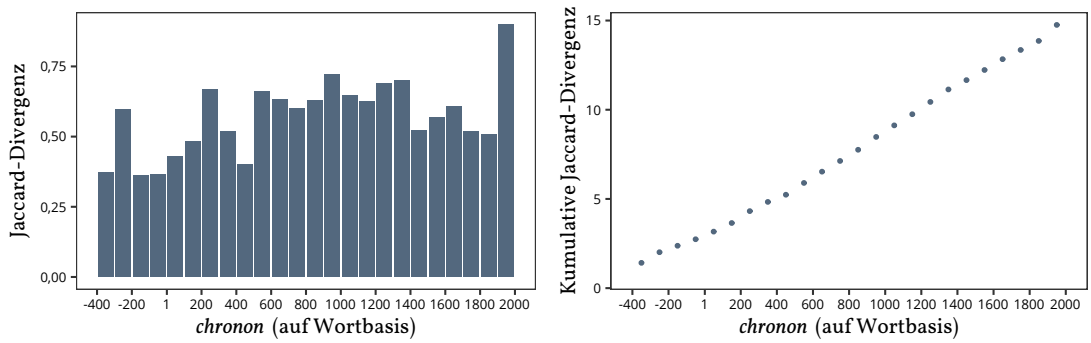


Abbildung 6.8 Jaccard-„Divergenz" zum *chronon* des vorigen Jahrhunderts

6.2 Datierung mit Neologismusprofilen

> *„[…] we foresee a role for parsed entries from historical dictionaries in this context as well."*[86]
>
> Franciska de Jong, Henning Rode and Djoerd Hiemstra

Wie in Kapitel 6.1 dargestellt, ist eine ungefähre Datierung schriftsprachlicher chinesischer Texte mit statistischen Sprachmodellen grundsätzlich möglich, wenn einige Voraussetzungen erfüllt sind: Das Genre muss bekannt sein und der Text sollte nicht in einem antiken Stil verfasst sein, der die Verwendung zeitgenössischer Lexeme und Satzkonstruktionen vermeidet. Um gute Ergebnisse zu erzielen, wird überdies ein passendes Trainingskorpus benötigt.

Die im Folgenden vorgestellte Datierungsmethodik basiert auf der aus dem *DHYDCD* erzeugten diachronen Lexemdatenbank[87] und ermöglicht eine ungefähre Altersschätzung von Texten, die genreunabhängiger ist und ohne spezifisches Trainingskorpus auskommt. Sie basiert auf der einfachen Idee, dass ein Text *mindestens* so neu ist, wie das neueste darin enthaltene Lexem (*Newest Word in Text*). Diese Herangehensweise stellt quasi eine Digitalisierung des *Lexical Dating* dar, „a method of establishing the chronology of a text through an examination of its vocabulary."[88] Würden wir den *Locus classicus jedes* Lexems bzw. jeder Kombination von Schriftzeichen und deren früheste Verwendung kennen, könnte durch einen Abgleich der *types* eines Texts mit einer entsprechenden Datenbank sein „neuestes Wort" ermittelt werden. Bei Vollständigkeit der verwendeten Daten ließe sich so implizit das *maximale* Alter des Textes ermitteln.

85 Vgl. aber Arapov und Cherc 1983 [1974], S. 88: „Die linguistische Erfahrung lehrt, daß einige Epochen der Sprachentwicklung durch stürmische Veränderungen gekennzeichnet sind, andere dagegen durch relative Stagnation. Diese Erfahrung beruht aber eher auf der Erforschung der historischen Phonetik und Morphologie und ist dann auf die schwer überschaubare Lexik übertragen worden."

86 de Jong, Rode und Hiemstra 2005, S. 1.

87 Siehe Kapitel 5.5, ab S. 120.

88 Toner und Han Xiwu 2019, S. 33–34.

In der Praxis ist der Wortschöpfungsprozess selten so genau dokumentiert. Mit den historischen Lexikalisierungsdaten aus dem *DHYDCD* stehen uns dennoch umfangreiche Daten zur Verfügung, die ernsthafte Experimente mit diesem Gedankenspiel zulassen. Einschränkungen, die sich aus der Unvollständigkeit dieser Daten ergeben, lassen sich durch Ergänzung früherer Belegstellen aus datierten Texten reduzieren.[89]

Die in einem zu datierenden Text erkannten Lexeme können auf dieser Datenbasis chronologisch zugeordnet und diese Zuordnung entsprechend visualisiert werden. Die für diese Methode verwendete Darstellung der Lexikalisierung bezeichne ich im Folgenden als *Neologismusprofil* bzw. als *temporales Profil* eines Textes. Durch die chronologische Einordnung der enthaltenen *types* können die Profile zum einen als philologisches Werkzeug genutzt werden, um Rückschlüsse über die stilistische, inhaltliche und temporale Einordnung des Textes zu ziehen. Mithilfe statistischer Überlegungen können sie zudem durch Software interpretiert werden, was auch einen groben Performancevergleich mit den Datierungsmethoden aus Kapitel 6.1 ermöglicht. Betrachten wir zur Veranschaulichung ein Neologismusprofil des *Meng xi bi tan* 夢溪筆談 (*MXBT*) von Shen Kuo 沈括 (1031–1095):[90]

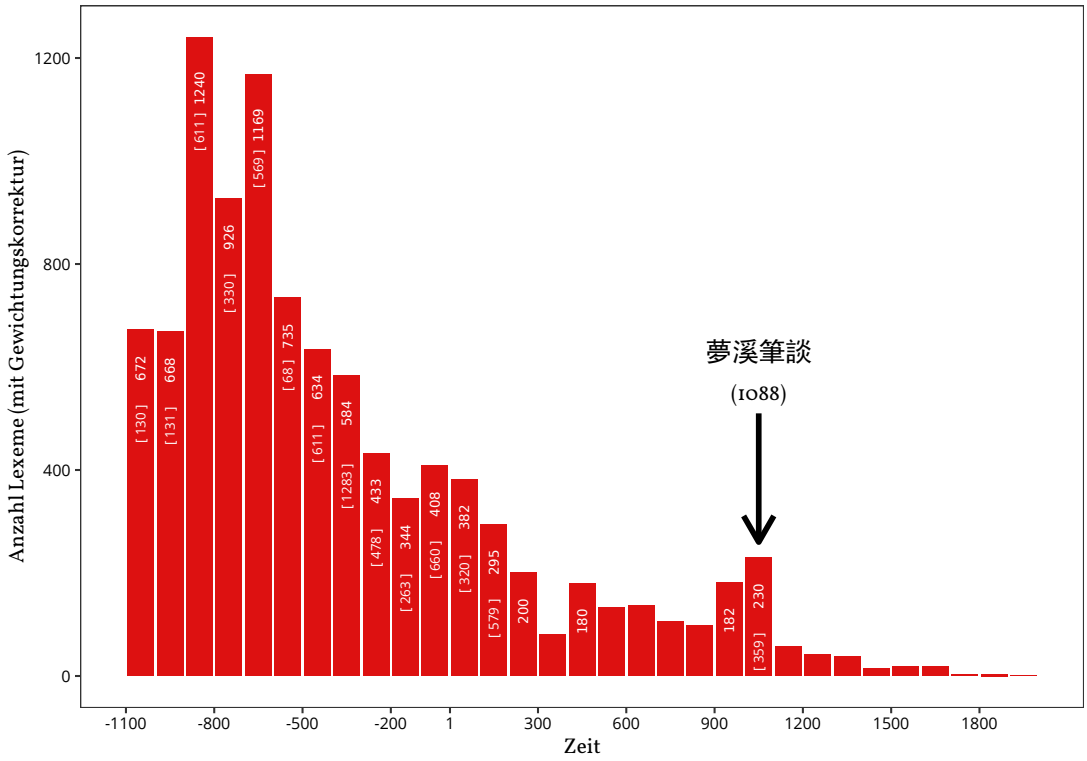


Abbildung 6.9 Neologismusprofil mit Gewichtungskorrektur für das *MXBT*, 2–4 Zeichen Lexeme

89 Siehe Kapitel 5.5.4, S. 134. Eine *vollständige* Lexikalisierungsgeschichte, die für eine naive *Newest Word in Text*-Datierung benötigt würde, könnte nur durch Betrachtung *sämtlicher* chinesischsprachiger Texte nachgezeichnet werden und bleibt damit eine theoretische Überlegung.

90 Shen Kuo 沈括 2008 [1088]: *Meng xi bi tan* 夢溪筆談 *(Pinselunterhaltungen am Traumbach)*. Project Gutenberg eBook. URL: http://www.gutenberg.net (besucht am 10. 09. 2018).

Das Balkendiagramm (Abb. 6.9) stellt die Anzahl der im *MXBT* enthaltenen Lexem-*types* pro Jahrhundert dar,[91] indem sie dem Jahrhundert ihrer frühesten bekannten Belegstelle zugeordnet werden.[92] Der Text enthält einen großen Anteil an Lexemen, die bereits sehr früh belegt sind. Zur Gegenwart hin nimmt die Anzahl der pro Jahrhundert nachgewiesenen Lexeme ab. Der Verlauf dieser Abnahme erinnert an die „arith-logarithmic equation" mit der George Zipf den Zusammenhang zwischen Häufigkeit und Alter von Wörtern herstellt.[93] Mikhail Arapov und Maja Cherc führen Zipfs Entdeckung aus:

> Es existiert ein Zusammenhang zwischen der Häufigkeit eines Wortes und der Zeit seiner Entstehung in der Sprache. Es zeigt sich, daß die Mehrheit der Wörter mit großer Auftretenshäufigkeit von den sehr alten Wörtern gebildet wird; umgekehrt ist die Chance dafür, daß es sich bei einem Wort um einen Neologismus handelt, umso größer, je geringer seine Häufigkeit ist.[94]

Obwohl die Lexeme des *MXBT* unabhängig von ihrer Häufigkeit im untersuchten Text dargestellt sind, scheint ein vergleichbarer Zusammenhang zu bestehen.[95]

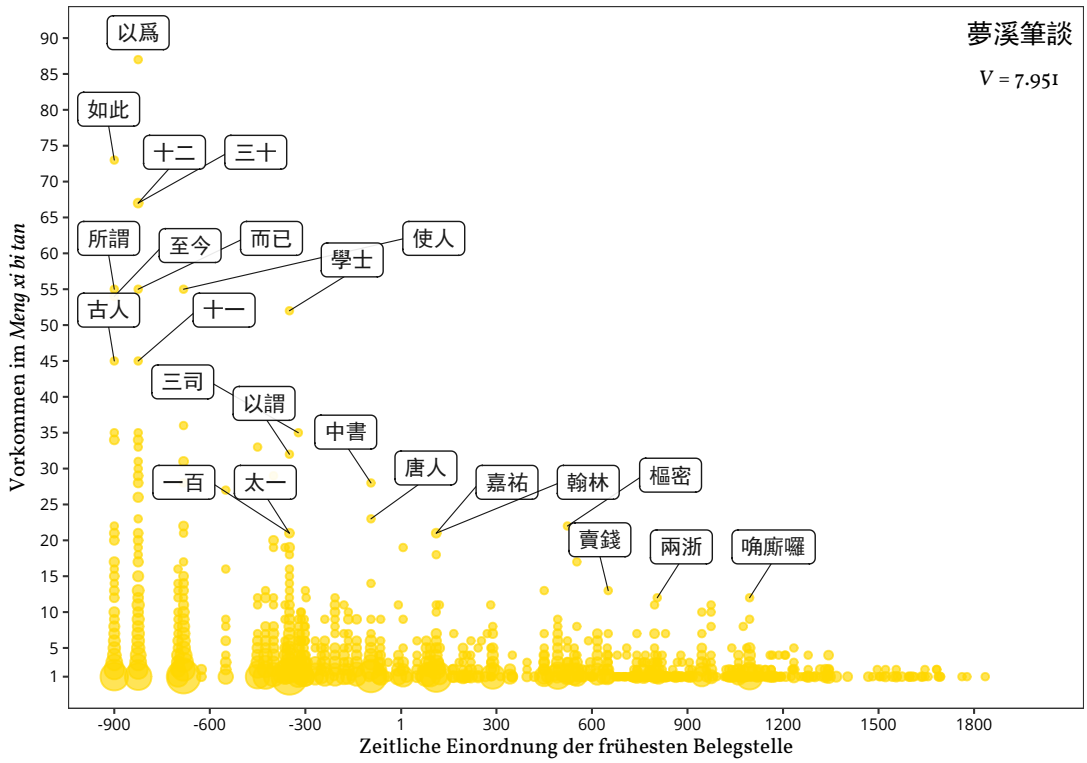


Abbildung 6.10 2–4 Zeichen Lexeme im *MXBT* chronologisch vs. Häufigkeit im Text

91 Vgl. auch Kapitel 5.7 (ab S. 138).

92 Die Erzeugung dieser Darstellung wird in Abschnitt 6.2.1, S. 182 erläutert, die gewählte Gewichtungskorrektur ab S. 185. In eckigen Klammern ist die ungewichtete für das jeweilige Jahrhundert festgestellte Anzahl *types* angegeben. Zusätzlich zu den Belegstellen aus dem *DHYDCD* werden weitere Daten herangezogen, wie in Kapitel 5.5.4 (ab S. 134) beschrieben.

93 Siehe Zipf 1947, S. 527; zitiert in Arapov und Cherc 1983 [1974], S. 29.

94 Arapov und Cherc 1983 [1974], S. 1; vgl. auch Zipf 1947, S. 523.

95 Vergleichbare logarithmische Verläufe der diachronen Verteilung des Vokabulars zeigen sich auch für andere Texte. Siehe z. B. Abb. 6.19b, S. 191, Abb. 6.20b, S. 191 und Abb. 6.25, S. 201.

Auch bei diachroner Betrachtung der Häufigkeiten von Lexemen im *MXBT* zeigt sich, dass die häufigsten *types* bereits sehr früh nachgewiesen sind (Abb. 6.10). In Richtung Gegenwart nimmt die Wahrscheinlichkeit für sehr häufige Lexeme immer weiter ab.[96]

Der überwiegende Anteil der Lexeme datiert vor die Entstehung des Textes im Jahr 1088. Zudem weist ihre chronologische Zuordnung im 11. Jh. eine Spitze auf (Abb. 6.9).[97] In den Zeitraum nach der tatsächlichen Entstehung des Textes (1100–2000) werden insgesamt noch 188 Lexeme datiert. Ihre Anzahl nimmt zwar kontinuierlich ab, zeigt aber auch, wie viele *types* allein in diesem einen Text enthalten sind, die zwar im *DHYDCD* lexikalisiert sind, deren älteste Belegstelle aber bis zu neun Jahrhunderte später datiert ist – obwohl das *MXBT* selbst im *HYDCD* zitiert wird und den Kompilator:innen offensichtlich vorlag. Einige Beispiele für Zeichenkombinationen, die deutlich früher belegbar sind als der *Locus classicus* im *HYDCD*, werden in Kapitel 5.5.4 besprochen.[98]

Während ein Großteil der als „zu neu" eingestuften Lexeme auf die „Nachlässigkeit" der *HYDCD*-Herausgeber zurückzuführen sein mag,[99] sind auch *false positives* vorhanden, die durch die naive *n*-Gramm-Segmentierung des Textes entstehen. So wird z. B. die Zeichenkombination *nancao* 南漕, im *MXBT* enthalten in *Huainan caoqu* 淮南漕渠 als Binomen dem 20. Jh. zugeordnet. Unabhängig von ihrer Ursache sind solche falschen Zuordnungen bei der Verwendung diachroner Lexikalisierungsdaten mit Text-*n*-Grammen gleichermaßen problematisch wie unvermeidbar.

6.2.1 Erzeugung von Neologismusprofilen

Die zur Erzeugung von Abb. 6.9 durchgeführten Schritte werden am Beispiel des *MXBT* erläutert:

1. Nicht verwendbare Zeichen (hier die englischsprachigen Angaben des *Project Gutenberg* zum *MXBT*) werden entfernt. Inklusive Interpunktion verbleiben 99.188 chinesische Zeichen.
2. Zur Normalisierung werden vorkommende Kurzzeichen durch Langzeichen ersetzt.[100] So werden u. a. alle Instanzen von *lu* 栌 durch 櫨 und *sha* 铩 durch 鎩 ersetzt.
3. Sonstige Varianten werden mit dem im *DHYDCD* vorgegebenen Standard normalisiert.[101] Dabei werden z. B. alle Instanzen von *wei* 為 durch 爲 (1.111 Vorkommen) und *xu* 敘 durch 敍 (52 Vorkommen) ersetzt. Insgesamt werden im *MXBT* so 49 Zeichen-*types* mit 1.594 Vorkommen ersetzt.[102]
4. Alle im Text enthaltenen 2–4-Gramme[103] werden ermittelt und gezählt.[104] Er enthält 226.029 2–4-Gramm *types*.

96 Vgl. auch Arapov und Cherc 1983 [1974], S. 51–87.
97 239 der 359 dem 11. Jh. zugeordneten Lexeme sind mit dem *MXBT* selbst belegt – die Spitze wäre also weniger markant, würde das *MXBT* nicht als Primärquelle für das *DHYDCD* dienen.
98 Siehe S. 135.
99 Siehe dazu auch Kapitel 5.3, ab S. 113.
100 Für diesen Schritt wird, wenn nötig, die Funktion `tradify` aus dem Paket `mafan` (Schaaf 2017, siehe auch Kapitel 4.3, S. 70) eingesetzt.
101 Siehe Kapitel 4.3, ab S. 69.
102 Sowohl durch `tradify` als auch `hydcd_standardize` können auch *types* erzeugt werden, die eigentlich nicht im Text enthalten sind, z. B. wenn *...li jian...* 里間 zu *lijian* 裏間 wird. Der Einsatz von Normalisierungswerkzeugen ist daher eine Abwägungsfrage, die je nach Art und Qualität der untersuchten Textdaten entschieden werden kann.
103 Einzelzeichen sind – zumindest in ihrer normalisierten Form – temporal kaum diskriminativ, wie auch in Kapitel 5.7 und 6.3 (ab S. 210) diskutiert.
104 Siehe dazu Kapitel 4.5.2, S. 91.

5. Aus der entstandenen Liste mit *n*-Gramm-*types* wird die Schnittmenge mit Einträgen im *DHYDCD* gebildet. 8.679 von den 226.029 im Text unterschiedenen 2–4-Grammen sind darin lexikalisiert.
6. Zu 8.210 (94,6 %) dieser Lexeme stehen chronologische Daten zur Verfügung, die aus der Datenbank geladen werden.
7. Jedes Lexem wird dem Jahrhundert zugeordnet, in das die älteste Quelle mit einer entsprechenden Belegstelle datiert ist. Bei ungenau datierten Quellen wird die Zuordnung – wie in Abschnitt 6.2.1 (S. 184) erläutert – ggf. aufgeteilt (*Slicing*).
8. Für die Darstellung in Abb. 6.9 wurde zudem eine Gewichtungskorrektur vorgenommen, welche die im *HYDCD* vorhandene Gewichtung ausgleichen soll.[105]

Verwendung zusätzlicher Belegstellen

Um die Problematik zu später Belegstellen im *DHYDCD* abzumildern, wurden für die zeitliche Zuordnung der Lexeme bzw. Zeichenkombinationen zusätzliche Korpus-Belegstellen herangezogen.[106] Abb. 6.11 zeigt die Neologismusprofile des *MXBT* mit Gewichtungskorrektur ohne diese zusätzlichen Daten (links), ergänzt um die Belege aus den *zhengshi*- und Loewe-Korpora (Mitte), sowie mit den *difangzhi* 地方誌-Belegen (rechts) im direkten Vergleich. Letztere stellen gerade im Bereich vom 17. bis Anfang des 20. Jhs. eine wichtige Ergänzung dar.

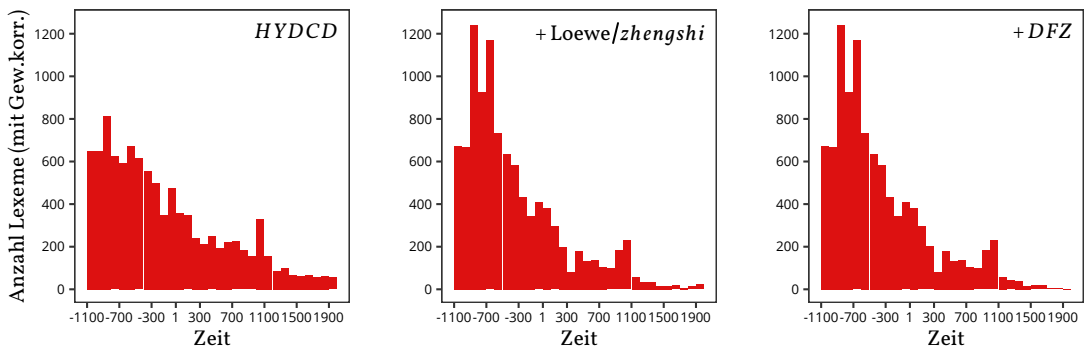


Abbildung 6.11 Profile des *MXBT* (nur *DHYDCD*-Belege; + Loewe-/*zhengshi* Belege; + *DFZ*-Belege)

Würden allein die Lexikalisierungsdaten aus dem *DHYDCD* vewendet, enthielten die meisten Texte also zahlreiche „zu neue" Wörter. Betrachtet man auf dieser Basis den Anteil verspätet belegter Zeichenkombinationen am Beispiel des *MXBT*, kann eine Fehlerquote von etwas weniger als 10 % konstatiert werden: Von den 8.136 Lexemen, zu denen chronologische Daten vorliegen, sind 7.393 dem Zeitraum zwischen 1100 v. u. Z. und 1100 zugeordnet, die restlichen sind rezenter. Mit den zusätzlichen Belegstellen sinkt diese Fehlerquote auf 2,3 %.

105 Die Gewichtungskorrektur wird ab S. 185 erläutert. Zur zeitlichen Gewichtung der *HYDCD*-Einträge siehe Kapitel 5.7.2, ab S. 142.
106 Siehe Kapitel 5.5.4, ab S. 134.

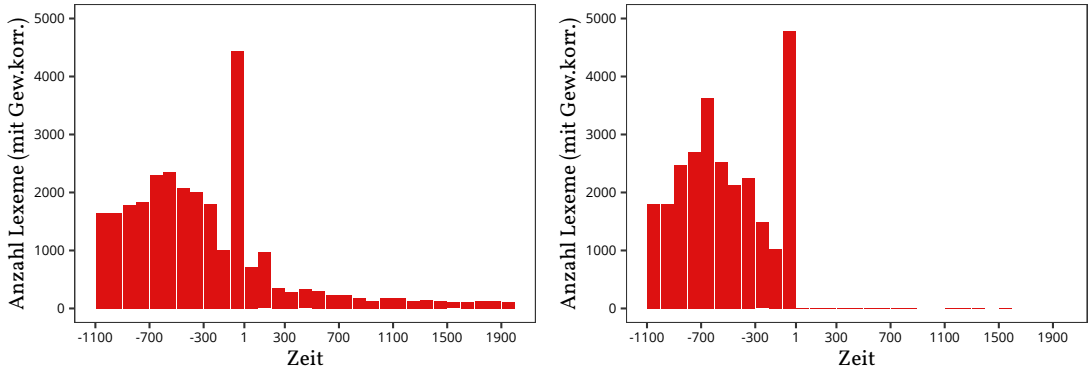


Abbildung 6.12 Neologismusprofile des *Shiji* ohne und mit zusätzlichen Korpus-Belegen

Bemerkenswert ist, dass auch im *DHYDCD* häufig zitierte Texte wie das *Shiji* 史記[107] eine hohe Anzahl an im *DHYDCD später* datierten Zeichenkombinationen enthalten können: 28,4 % der festgestellten Lexeme sind ohne die Erweiterungen der Datenbank später datiert als der Text selbst (Abb. 6.12).

Slicing – Zu welchem *chronon* gehört ein Lexem?

Um alle Lexeme eines Textes einem bestimmten Zeitraum – hier ein Jahrhundert – zuzuordnen, sollte der *Locus classicus* eindeutig auf ein bestimmtes Jahr datiert sein. Vor allem ältere Quellen sind aber schlechtestenfalls nur auf den Zeitraum einer ganzen Dynastie, oder zumindest der Lebensspanne des Autors genau datierbar.[108] Die Zuordnung muss gegebenenfalls also entsprechend aufgeteilt werden. Der Satz „鄜延境内有石油，舊説高奴縣出脂水，即此也。“[109] enthält z. B. 57 unterschiedliche 2–4-Gramm-*types*, von denen drei als Lexeme chronologisch zugeordnet werden können:

1. Die früheste Belegstelle zu *zhishui* 脂水 im *DHYDCD* stammt aus einem *fu* 賦 von Du Mu 杜牧 (803–852),[110] dessen Lebensdaten aus der *CBDB* mit 803–853 übernommen wurden. Der Datierungszeitraum fällt damit vollständig ins 9. Jh., was bei *chronons* von 100 Jahren unproblematisch ist.
2. *Shiyou* 石油 wird dem 11. Jh. zugerechnet, da das *MXBT* selbst als *Locus classicus* angegeben ist und in der *CBDB* auf das Jahr 1095 datiert ist.
3. *Jici* 即此 belegt das *DHYDCD* mit dem tangzeitlichen Gedicht *Qiu huai shi* 秋懷詩 von Han Yu 韓愈 (768–824[111]). Da der Zeitraum ungleich auf das 8. und 9. Jh. aufgeteilt ist, erfolgt eine anteilige Zurechnung zu beiden *chronons*, die ich als *Slicing* bezeichne. Von 57 Jahren entfallen 32 auf das 8., 25 auf das 9. Jh. Demnach werden dem 8. Jh. $\frac{32}{57}$ = 0,561 *types*, dem 9. Jh. $\frac{25}{57}$ = 0,439 *types* zugerechnet.

107 Siehe dazu Kapitel 5.7, ab S. 138.
108 Siehe v. a. Kapitel 5.5.2, ab S. 127 bzw. 5.5.3, ab S. 132.
109 *DHYDCD*, 石油. Die Belegstelle für *shiyou* stammt aus dem *MXBT*.
110 Siehe Emmerich 2004, S. 160.
111 Siehe ebd., S. 155, 768–825 lt. *CBDB*.

Der Beispielsatz aus dem *MXBT* hätte damit folgendes Profil, das mit nur drei Beobachtungen natürlich wenig Aussagekraft hat – es dient lediglich zur Veranschaulichung.

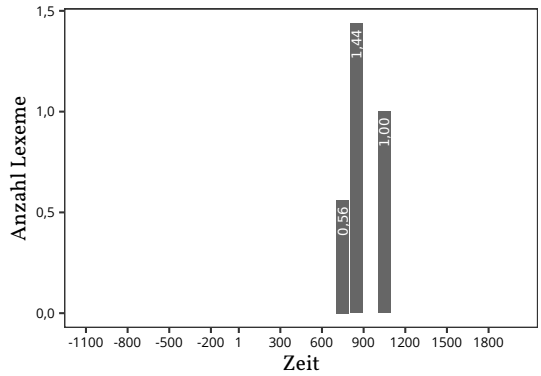


Abbildung 6.13 Neologismusprofil für „鄜延境内有石油，舊説高奴縣出脂水，即此也。“

Gewichtungskorrektur – Entfernung des *HYDCD-Bias*

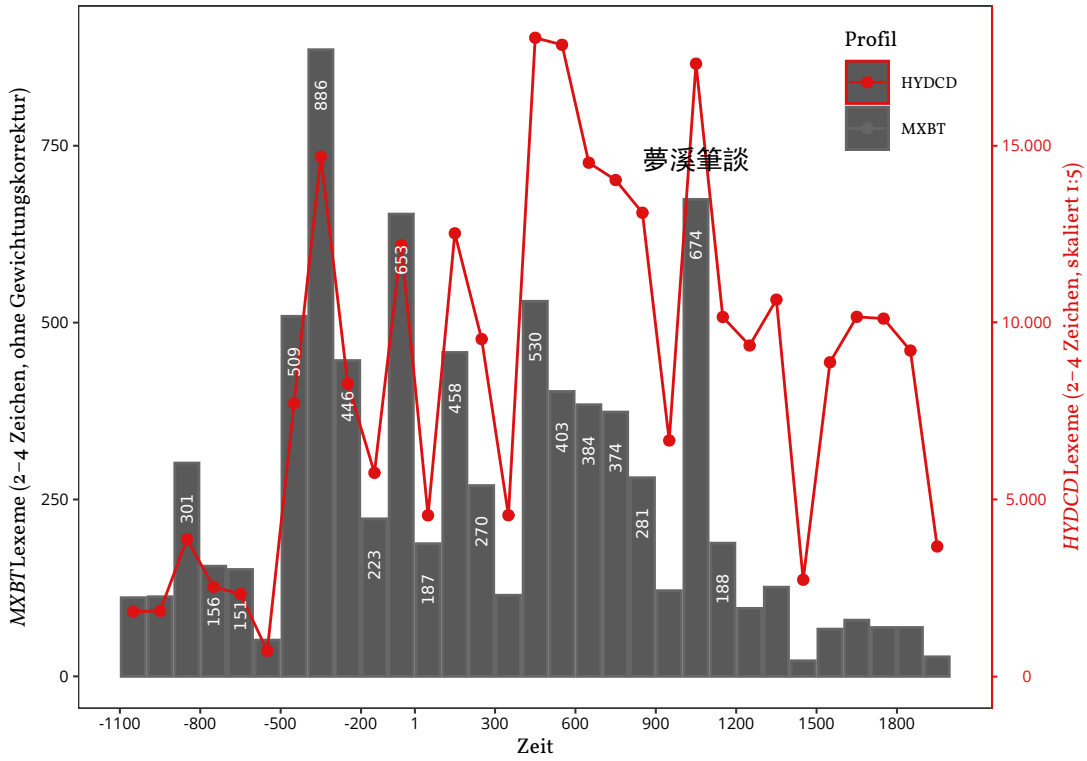


Abbildung 6.14 Neologismusprofil für das *Meng xi bi tan* vs. Lexikalisierung im *HYDCD*

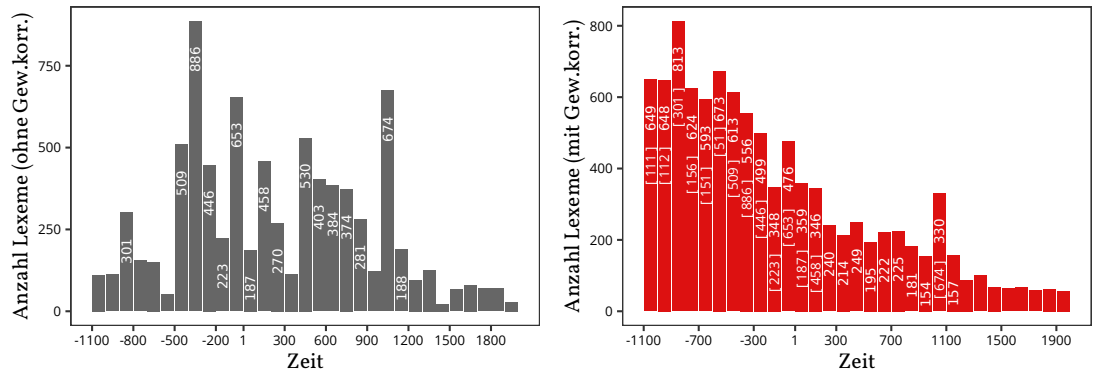
Wird – wie im gerade gezeigten Beispiel – die „rohe“ Anzahl *types* pro Jahrhundert zur Erzeugung von Neologismusprofilen verwendet, sind diese oft schwierig zu interpretieren, da die

dem Text inhärente Lexikalisierungsgeschichte vom *Bias* des *HYDCD* überdeckt wird, bzw. eine ungewollte Gewichtung stattfindet.[112] Abb. 6.14 zeigt das Profil des *MXBT* mit Rohdaten. Zum direkten Vergleich sind alle in das jeweilige Jahrhundert datierten 2–4-Zeichen-Lexeme aus dem *DHYDCD* im Maßstab 1:5 darüber gelegt (rot). Vor allem bis zur Entstehung des Textes im 11. Jh. folgen die für das *MXBT* beobachteten Schwankungen sehr deutlich der Lexikalisierung des *DHYDCD*.

Dieser Effekt lässt sich durch eine Gewichtungskorrektur gemäß der Lexikalisierungsdaten nahezu vollständig eliminieren. Zu diesem Zweck können unterschiedliche Überlegungen getroffen werden. Trifft man die stark vereinfachende Annahme einer „tendency of vocabulary to change at a uniform rate",[113] dass also die Aufnahme neuer Wörter in den Wortschatz ein insgesamt kontinuierlicher Prozess ist und in jedem Jahrhundert damit etwa gleich viele Wörter hinzukommen würden,[114] kann für jedes Jahrhundert im Beobachtungszeitraum ein Gewicht w_c berechnet werden, mit dem multipliziert sich für alle Jahrhunderte dieselbe Anzahl an *types* ergibt. Hierzu wird die Gesamtmenge aller über Belegstellen aus dem *DHYDCD* datierten Lexeme $|\ V\ |$ durch die Anzahl der Jahrhunderte im Betrachtungszeitraum $|\ C\ |$ und die Menge der auf das jeweilige Jahrhundert datierten Lexeme $|\ V_c\ |$ dividiert.

$$w_c = \frac{|\ V\ |}{\frac{|C|}{|V_c|}}$$

Wendet man w_c als Korrekturfaktor auf die y-Werte in Abbildung 6.15a an, ergibt sich ein homogeneres, einfacher zu interpretierenendes *Neologismusprofil* des *MXBT* (Abb. 6.15b, siehe auch Abb. 6.9):



(a) Ohne Gewichtungskorrektur **(b)** Mit Gewichtungskorrektur

Abbildung 6.15 Neologismusprofile des *MXBT* (ohne Korpus-Belegstellen)

112 Siehe auch Kapitel 5.7.2, Abb. 5.8, S. 143

113 Swadesh 1955, S. 122; vgl. auch Arapov und Cherc 1983 [1974], S. 89: „[I]n einigen Fällen [scheint] der lexikalische Wandel wirklich mit nahezu konstanter Geschwindigkeit zu erfolgen."

114 Diese veraltete Annahme aus der Lexikostatistik ist bestenfalls als stark vereinfachend anzusehen. Ebenfalls mögliche sprunghafte Erweiterungen des Wortschatzes durch wichtige gesellschaftliche Ereignisse, Umbrüche, Sprachkontakt, Internationalisierung, Technologiewandel etc. werden von solchen Modellen nicht berücksichtigt. Zudem konnte verschiedentlich gezeigt werden, dass der Wortschatzzuwachs, wie auch andere Manifestationen von Sprachwandel, eher einer *s*-Kurve folgt.

Gegen die Annahme einer konstanten Veränderung des Vokabulars spricht, dass auch die Lexikalisierungsdaten des *DHYDCD* suggerieren, dass seine Zunahme einer *s*-Kurve folgt. Für eine entsprechende *s*-Gewichtungskorrektur der Neologismusprofile (Abb. 6.16) kann – wie bereits in Kapitel 5.7.2 gezeigt – die Funktion der idealisierten Lexikalisierung mit *R* geschätzt und die resultierenden Werte der theoretischen Neulexikalisierung für jedes Jahrhundert berechnet werden.[115] Der *s*-Korrekturfaktor s_c ergibt sich aus dieser idealisierten und der tatsächlich gemessenen Lexikalisierung des jeweiligen Jahrhunderts.

$$s_c = \frac{| V_{c_{ideal}} |}{| V_c |}$$

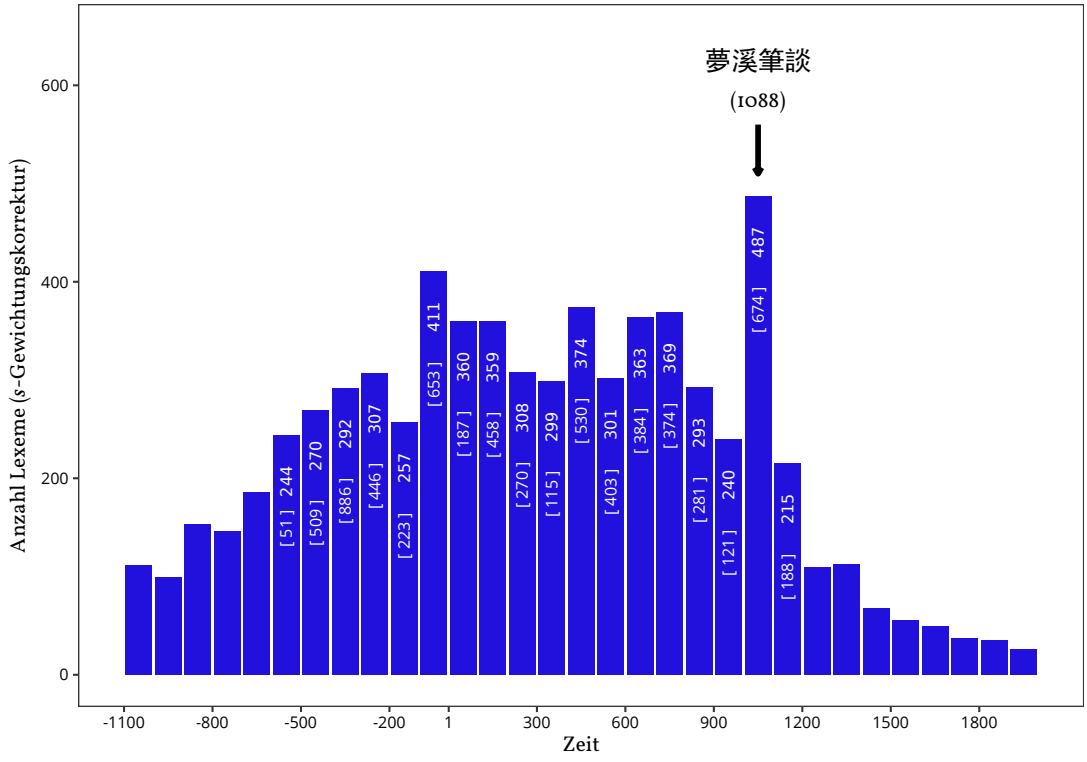


Abbildung 6.16 Neologismusprofil für das *MXBT* (*s*-Gewichtungskorrektur, ohne Korpusbelegstellen)

Welche Darstellung als Grundlage für eine philologische Interpretation vorzuziehen ist, mag von subjektiven Präferenzen bestimmt sein – für eine automatisierte Datierung funktioniert die einfache Annahme des konstanten Wortschatzwachstums jedoch minimal besser, wie in Abschnitt 6.2.5 (ab S. 197) gezeigt wird. Bei optischer Analyse des Profils sind eventuelle Peaks mit linearer Gewichtungskorrektur zudem gegebenenfalls leichter erkennbar.

Die Werte für den Faktor w_c werden gemäß der im *DHYDCD* aufgezeichneten Lexikalisierung pro Jahrhundert berechnet. Damit sind sie abhängig von der Zeichenlänge der Lexeme und der Veränderung ihrer Datierung durch die oben beschriebene Ergänzung früherer Belegstel-

115 Wie bereits in Kapitel 5.7.2, v. a. S. 146, verwende ich hierfür die Funktion `drm` aus dem *R*-Paket `drc`. Siehe Ritz 2016.

len. Tabelle 6.9 gibt die entsprechenden Werte von w_c für den gesamten Betrachtungszeitraum wieder. Die ersten vier Wertespalten beziehen sich auf 2–4 Zeichen-Lexeme ohne Berücksichtigung von zusätzlichen Korpusbelegstellen. Bei einer durchschnittlichen Lexikalisierung von 8.305 Wörtern pro Jahrhundert ergibt sich aus der naiven Annahme einer linearen Lexikalisierung der jeweilige Faktor aus der Anzahl der *types* V_c. Die Spalte $V_{c_{ideal}}$ gibt die mit der Annahme eines *s*-förmigen Wortschatzwachstums modellierte Lexikalisierung an, die nächste Spalte den daraus errechneten *s*-Faktor. Da dem Modell eine kumulative Betrachtung zugrunde liegt, wird der Wortschatzzuwachs als Differenz zum jeweils vorherigen Jahrhundert berechnet. Der erste Faktor s_{-1100} ist daher mit 1 angegeben. Für die aufgezeichnete Lexikalisierung nach Berücksichtigung der zusätzlichen Korpusbelegstellen sind ebenfalls die Werte von V_c und w_c, jeweils für die Betrachtung von 2–4 und 2–3-Zeichen Lexemen angegeben.[116]

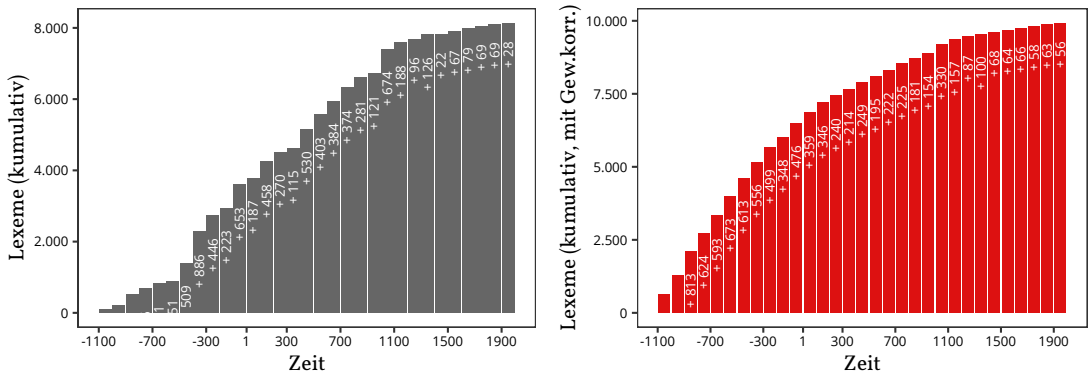
Tabelle 6.9 Lexikalisierung und Gewichtungskorrekturfaktoren nach Jahrhundert

Zeitraum	2–4 Z., ohne Korpusbelege				2–4 Z., m. Belegen		2–3 Z., m. Belegen	
Jh.	# *types* V_c	w_c	$V_{c_{ideal}}$	s_c	V_c	w_c	V_c	w_c
1100-1000 v. u. Z.	1.423	5,836	1.423	1	1.612	5,165	1.509	5,100
1000-900 v. u. Z.	1.442	5,761	1.624	1,126	1.627	5,115	1.523	5,053
900–800 v. u. Z.	3.082	2,695	1.968	0,639	4.099	2,030	3.777	2,038
800–700 v. u. Z.	2.071	4,010	2.378	1,148	2.967	2,806	2.842	2,708
700–600 v. u. Z.	2.114	3,929	2.862	1,354	4.055	2,053	3.974	1,937
600–500 v. u. Z.	632	13,133	3.430	5,424	766	10,861	708	10,871
500–400 v. u. Z.	6.889	1,206	4.089	0,594	8.023	1,037	7.449	1,033
400–300 v. u. Z.	13.224	0,628	4.844	0,366	18.293	0,455	17.178	0,448
300–200 v. u. Z.	7.425	1,119	5.697	0,767	9.197	0,905	8.678	0,887
200–100 v. u. Z.	5.312	1,563	6.641	1,250	6.363	1,308	6.084	1,265
100–1 v. u. Z.	11.399	0,729	7.663	0,672	13.457	0,618	12.680	0,607
1–100	4.332	1,917	8.739	2,017	6.972	1,194	6.691	1,150
100–200	10.993	0,756	9.834	0,895	16.335	0,510	15.640	0,492
200–300	9.336	0,890	10.902	1,168	11.954	0,696	11.401	0,675
300–400	4.453	1,865	11.886	2,669	3.382	2,461	3.209	2,399
400–500	17.693	0,469	12.725	0,719	18.510	0,450	17.632	0,436
500–600	17.193	0,483	13.362	0,777	16.836	0,494	16.116	0,478
600–700	14.350	0,579	13.749	0,958	11.995	0,694	11.109	0,693
700–800	13.773	0,603	13.853	1,006	10.149	0,820	9.514	0,809
800–900	12.854	0,646	13.666	1,063	10.094	0,825	9.521	0,808
900–1000	6.513	1,275	13.205	2,027	8.191	1,016	7.554	1,019
1000–1100	16.960	0,490	12.504	0,737	13.028	0,639	12.098	0,636
1100–1200	9.947	0,835	11.617	1,168	7.531	1,105	6.729	1,144
1200–1300	9.161	0,907	10.604	1,158	7.976	1,044	7.036	1,094
1300–1400	10.466	0,794	9.523	0,910	9.999	0,832	9.192	0,837
1400–1500	2.682	3,097	8.429	3,143	1.999	4,163	2.656	2,897
1500–1600	8.706	0,954	7.365	0,846	6.631	1,255	8.455	0,910
1600–1700	9.938	0,836	6.363	0,640	8.055	1,033	7.062	1,090
1700–1800	9.939	0,836	5.444	0,548	7.435	1,119	5.605	1,373
1800–1900	9.076	0,915	4.619	0,509	6.561	1,269	3.552	2,167
1900–2000	4.081	2,035	3.892	0,953	3.925	2,121	1.413	5,446
Ø V_c	8.305				8.323		7.696	

116 Letztere ist für die Verwendung von *n*-Gramm Datensätzen, deren *n*-Gramm Raum auf 3 limitiert ist, unverzichtbar. Siehe Abschnitt 6.2.5, ab S. 197, bzw. Kapitel 4.2, S. 66.

Kumulative Darstellung

In einer kumulativen Darstellung (Abb. 6.17) ließe sich unter optimalen Bedingungen ein zunächst logarithmischer Anstieg beobachten, der ab dem Jahrhundert, aus dem der untersuchte Text stammt, stark abflacht. Da die Lexikalisierung pro Jahrhundert allerdings deutlich schlechter erkennbar ist und für ein echtes Abflachen zu viele *false positives* vorhanden sind, ist die distinktive Darstellung vorzuziehen.



(a) Ohne Gewichtungskorrektur

(b) Mit Gewichtungskorrektur

Abbildung 6.17 Kumulative Neologismusprofile des *Meng xi bi tan* (ohne Korpusbelegstellen)

6.2.2 Temporale Textprofile: Erweiterung um Namen und Zeitausdrücke

Neben Wörtern bzw. Lexemen enthalten Texte oft zusätzliche zeitlich konnotierte Zeichenfolgen. Dazu zählen *temporal expressions* – Erwähnungen bestimmter Jahre oder Monate, sowie Personennamen ab einer Länge von drei Zeichen,[117] die mit der Lebensspanne der genannten Person eine temporale Dimension gewinnen.[118] Zu diesem Zweck werden Informationen zu Personen aus der *CBDB* geladen,[119] sowie auf Basis der *DDBC* nach *temporal expressions* gesucht. Wie in Kapitel 4.8 diskutiert, werden Zeitangaben in schriftsprachlichen chinesischen Texten gewöhnlich in Form von Regierungsdevisen und Jahresangaben gemacht, gegebenenfalls gefolgt von Monats- und Datumsangaben im lunisolaren Kalender.[120]

Die Neologismusprofile lassen sich so zu einem *temporalen Textprofil* erweitern, das die Informationen zusammenfassend darstellt. Theoretisch denkbar ist zudem die Nutzung von Ortsnamen, die ebenfalls in der *CBDB* auf die früheste den Herausgebern vorliegende Nennung datiert sind. Dass diese Daten mit Vorsicht zu genießen sind, zeigt Abb. 6.18. Gut ein Drittel der insgesamt 102 Übereinstimmungen mit unterschiedlichen Ortsnamen sind dem 12.–16. Jh. zugeordnet.

Auch unter den Personennamen finden sich – trotz des Ausschlusses uneindeutiger Namen – einzelne *false positives*, die ebenfalls dem 12.–16. Jh. zugeordnet sind. Eine ein-eindeutige Zuordnung von Namen zu bestimmten Personen ist nicht zuverlässig möglich, da Texte andere

117 Namen mit zwei Zeichen weisen ein sehr hohes Ambiguitätspotenzial auf. Siehe Kapitel 4.7, ab S. 97.
118 Siehe auch Abschnitt 6.1.1, S. 159. Es werden dabei nur Namen berücksichtigt, die in der *CBDB* nur einer einzigen Person zugeordnet sind.
119 Siehe Kapitel 4.7, ab S. 97.
120 Siehe Kapitel 4.8, ab S. 103.

Personen gleichen Namens erwähnen können. Außerdem können Zeichenfolgen, die für Namen verwendet werden, in derselben Kombination auch in lexikalisierten Bedeutungen auftreten.[121]

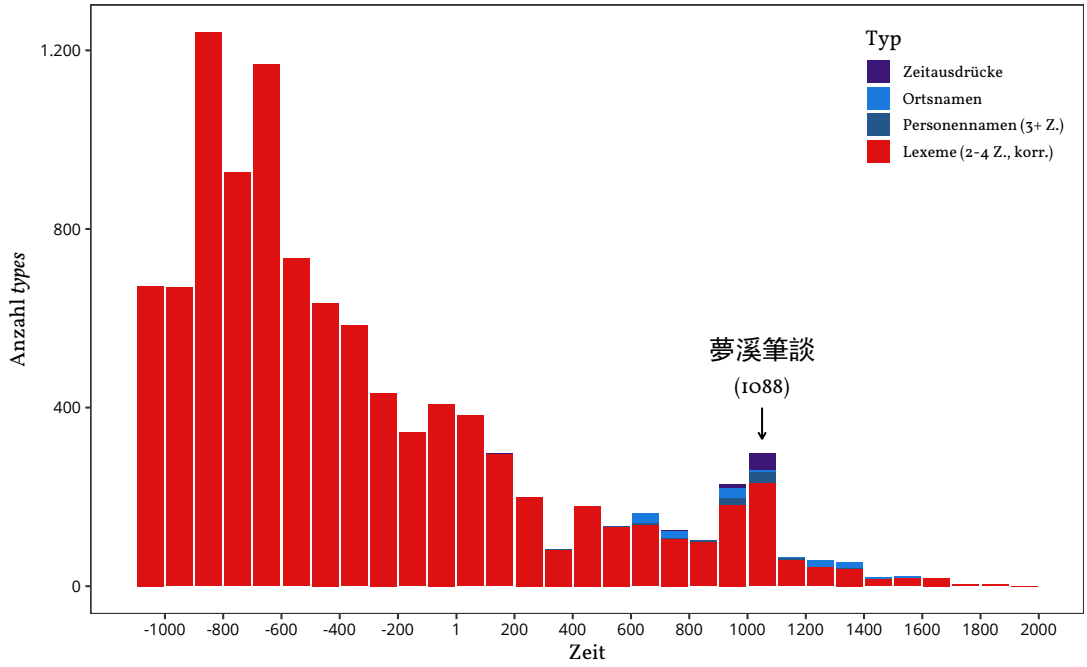


Abbildung 6.18 Temporales Textprofil für das *Meng xi bi tan*

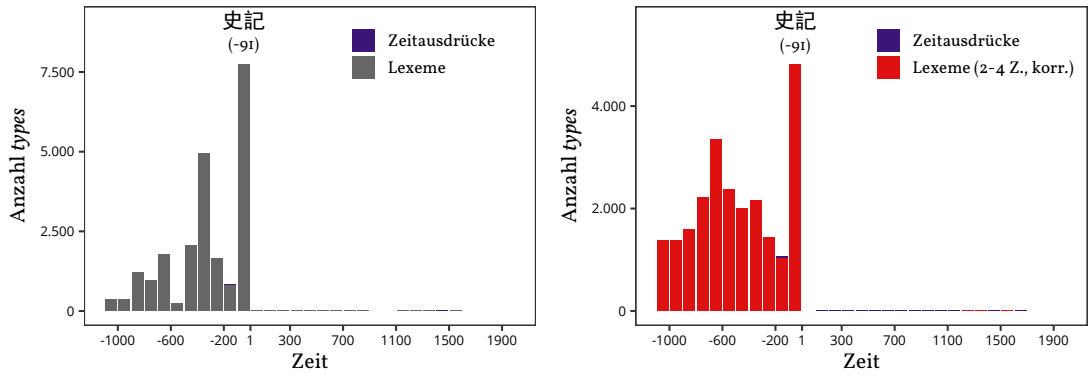
6.2.3 Interpretation temporaler Textprofile

Für den bereits gezeigten Fall des *MXBT* lässt sich aus dem temporalen Textprofil mit dem 11. Jh. bereits graphisch ein wahrscheinlicher Zeitraum der Entstehung des Textes ablesen. Er enthält einen großen Anteil älteres Vokabular, der dann zum Jahrhundert der Textgenese hin abnimmt. Zudem ist wieder ein größerer Anteil an Vokabular aus der Zeit der Textentstehung enthalten, sowie Namen von Zeitgenoss:innen (Abb. 6.18). Es kommt uns zugute, dass das *MXBT* selbst von den Kompilator:innen des *DHYDCD* gerne als *Locus classicus* herangezogen wurde.[122] Zudem wird im Text auf rezente historische Ereignisse Bezug genommen, so dass die chronologische Zuordnung von *temporal expressions* bekräftigt wird.[123]

Einfacher wäre die Interpretation bei Vollständigkeit der historischen Lexemdatenbank: Der untersuchte Text könnte in der Regel dem spätesten Jahrhundert zugeordnet werden, aus dem Lexeme nachgewiesen sind. Durch Betrachtung von zwei *zhengshi*-Texten aus dem Trainingskorpus[124] lassen sich diese Optimalbedingungen simulieren. *Shiji* 史記 ist auf die Lebensspanne von

121 Siehe dazu die Erläuterungen und Beispiele in Kapitel 4.7, ab S. 97.
122 Von 8.129 betrachteten Lexemen werden 359 dem 11. Jh. zugeordnet, bei 239 davon ist das *MXBT* selbst als *Locus classicus* angegeben.
123 Insgesamt 113 unterschiedliche Angaben werden erkannt, davon 54 mit Angaben zum 11. Jh, z. B. *Yuanfeng wu nian* 元豐五年 („das fünfte Jahr [i. e. 1082] der Regierungsdevise *Yuanfeng* [1078–1085] von Kaiser Shenzong 神宗 der Song-Dynastie (reg. 1067–1085“)
124 Siehe dazu Kapitel 5.5.4, ab S. 134.

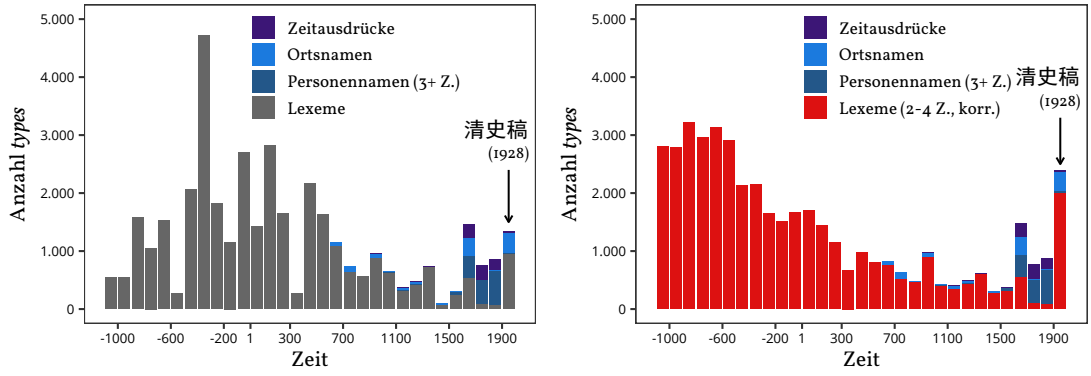
Sima Qian 司馬遷 (ca. 145–90 v. u. Z.) datierbar,[125] das 1928 veröffentlichte *Qingshi Gao* 清史稿 behandelt die Geschichte der Qing-Dynastie (清, 1644–1912) und ist in einem an das *Shiji* angelehnten Stil verfasst.



(a) *Shiji* **(b)** *Shiji*, mit Gewichtungskorrektur

Abbildung 6.19 Temporale Profile des *Shiji* 史記

Bei Verwendung der zusätzlichen Belegstellen sind im Profil des *Shiji* kaum *types* sichtbar, die erst nach dem 1. Jh. v. u. Z. belegt sind. Eine Spitze, die noch deutlicher auf dieses Jahrhundert hinweist, ist wenig überraschend, da der Text selbst häufig als Belegstelle herangezogen wird.[126]



(a) *Qingshi gao* **(b)** *Qingshi gao*, mit Gewichtungskorrektur

Abbildung 6.20 Temporale Profile des *Qingshi gao* 清史稿

Für das *Qingshi gao* sind etliche Lexeme sichtbar, die über den gesamten Beobachtungszeitraum, bis ins 20. Jh., datiert werden. Im Profil ohne Gewichtungskorrektur (Abb. 6.20a) macht sich an den starken Schwankungen in der nachgewiesenen Anzahl an Lexemen erneut das bereits an-

125 Das *Shiji* wurde von Sima Tan 司馬談 (gest. 110 v. u. Z.) begonnen und dann von Sima Qian fertiggestellt. Auch wenn begründete Zweifel an der Authentizität einzelner Kapitel bestehen, bei denen es sich um spätere Rekonstruktionen handeln könnte, sollte der Text im Wesentlichen zu Lebzeiten Sima Qians entstanden sein. Siehe Anthony François Paulus Hulsewé 1993b: „Shih chi 史記". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 405–406.

126 Siehe Kapitel 5.7, ab S. 138.

gesprochene *Bias* bemerkbar. Der hohe Anteil an Vorkommen von Personennamen und *temporal expressions* für den Zeitraum zwischen 1600–1900 spiegelt zudem den Inhalt des Texts wider.

Selbst mit kontinuierlicher Ergänzung der Lexemdatenbank durch frühere Belegstellen ist ein Zustand, in dem für *alle* Lexeme die früheste Belegstelle korrekt und genau datiert ist, quasi unerreichbar. In der Praxis kann die Interpretation zudem auch dann komplexer und aufwändiger sein, wenn Texte in einem „altertümelnden Stil" abgefasst sind, der nicht nur syntaktisch die Entstehungszeit des Textes verschleiert, sondern auch zeitgenössisches Vokabular bewusst vermeidet. Dies kann entweder durch bewusste Fälschung geschehen,[127] oder ist den Anforderungen oder Gepflogenheiten bestimmter Textgattungen, stilistischen Trends oder Vorlieben von Autor:innen geschuldet.

6.2.4 Das *Zhongjing* 忠經 als Anwendungsbeispiel

Ein schriftsprachlicher Text, dessen Entstehungszeit von Sinolog:innen diskutiert wird, ist das *Zhongjing* 忠經 („Klassiker der Loyalität").[128] Er wird traditionell Ma Rong 馬融 (79–166) zugeschrieben, was eine Datierung in die Han 漢-Zeit impliziert.[129] Die Autorschaft gilt jedoch als widerlegt, da Zitate aus Alttext-Teilen des *Shangshu* 尚書 (*Guwen Shangshu* 古文尚書) enthalten sind. Diese Textteile entstanden wahrscheinlich erst Anfang des 4. Jahrhunderts.[130]

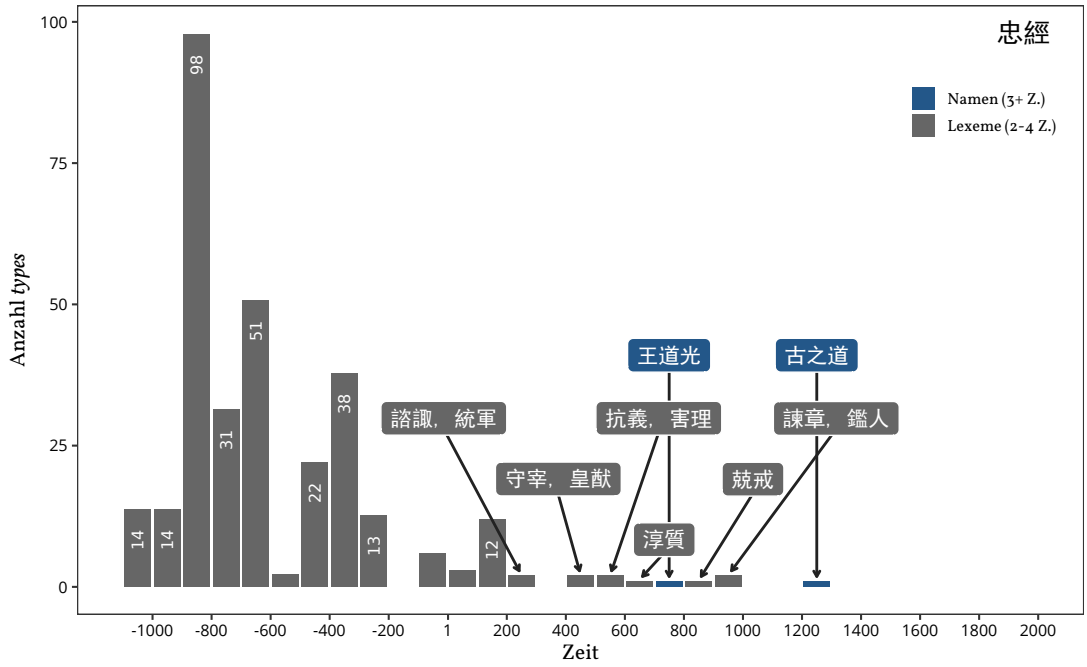


Abbildung 6.21 Temporales Profil des *Zhongjing* 忠經, ohne Gewichtungskorrektur

127 Siehe dazu auch Kapitel 3, S. 37.
128 Siehe auch Kapitel 3, ab S. 35. Den Hinweis, das *Zhongjing* in diesen Kontext zu stellen, verdanke ich Kai Vogelsang.
129 Siehe Suwald 2008, S. 7.
130 Siehe Nylan 2001, S. 134; zitiert in Suwald 2008, S. 70, siehe auch Kapitel 3, S. 39.

Die tatsächliche Entstehungszeit des *Zhongjing* bleibt Gegenstand von Spekulationen, die sich auf einen Zeitraum zwischen etwa dem Jahr 320 und dem Beginn der Song 宋-Dynastie (960–1279) erstrecken Aus jener Zeit stammen die ältesten schriftlich überlieferten Belege für die Existenz des *Zhongjing*.[131] Bedenkt man, dass die entlarvenden Zitate theoretisch auch nachträglich zu der heute vorliegenden Textfassung hinzugefügt worden sein können, bleibt sogar die traditionelle Zuschreibung im Bereich des Möglichen.

Betrachten wir das temporale Textprofil einer digitalen Ausgabe (Abb. 6.21).[132] Acht Lexeme und zwei „Namen" datieren später als das 3. Jh. Letztere sind mit einem Blick auf den Kontext schnell als *false positives* entlarvt: Wang Daoguang 王道光[133] (gest. 751) und Gu Zhidao 古之道[134] (früheste Belegstelle von 1294) kommen beide nicht als Namen, sondern als Folgen von Zeichen mit lexikalisierten Bedeutungen vor. Zeichenkombinationen, die erst nach der Song-Zeit nachgewiesen sind, enthält der Text nicht, was Suwalds Annahme, dass eine Entstehung des Textes nach 1040 sehr unwahrscheinlich ist,[135] bekräftigt.

Die acht Lexeme, die in den Zeitraum zwischen dem 4. und 10. Jh. datiert sind (Tabelle 6.10), können nun mit überschaubarem Aufwand geprüft werden, etwa durch Suche nach weiteren Belegstellen. Der Verdacht auf eine spätere Datierung lässt sich somit erhärten oder abschwächen. Eine derartige Vorgehensweise entspricht im Wesentlichen auch derjenigen Suwalds, die ebenfalls – allerdings ohne entsprechende Datenbank – mithilfe der vorkommenden Zeichenfolgen der Entstehungszeit des *Zhongjing* auf den Grund geht. Dabei kann sie zusätzlich Ausdrücke untersuchen, die im *DHYDCD* nicht lexikalisiert sind, z. B. *Zhou Kong zhi cai* 周孔之才.[136]

Tabelle 6.10 2–4-Zeichen-Kombinationen im *Zhongjing* mit Nachweisen *nach* dem 3. Jh.

#	Lexem	belegt in:[137]	datiert auf:	Vork.	Kontext
1	*jiànzhāng* 諫章	*Jiu Tang shu* 舊唐書	945	1	…書》云「旌別淑忒」，其是謂乎忠諫章第十五忠臣之事君也，莫先於諫，…
2	*jiànrén* 鑑人	*Jiu Tang shu* 舊唐書	945	1	…後從諫則聖。」證應章第十六惟天鑑人，善惡必應。善莫大於作忠，惡莫…
3	*jīngjiè* 兢戒	*Shang Jiang shilang qi* 上蔣侍郎啟	*ca.* 860–866	1	…之，天下盡忠，以奉上也。是以兢兢戒慎，日增其明，祿賢官能，式歉大…
4	*chúnzhì* 淳質	*Sui shu* 隋書	617	1	…，則人不爭。故得人心和平，天下淳質，樂其生，保其壽，優遊聖德，以…
5	*kàngyì* 抗義	*Wei shu* 魏書	*ca.* 551–554	1	…則非忠臣。夫諫，始於順辭，中於抗義，終於死節，以成君休，以寧社稷…
6	*hàilǐ* 害理	*Nanqi shu* 南齊書	*ca.* 509–537	1	…審則分。君子去其私，正其色，不害理以傷物，不憚勢以舉任。惟善是與…
7	*shòuzǎi* 守宰	*Hou Han shu* 後漢書	*ca.* 445	2	…詩》云「靖共爾位，好事正直。」守宰章第五在官惟明，蒞事惟平，立身… …觀乎子，則人愛之，如愛其親，蓋守宰之忠也。《詩》云「愷悌君子，民…
8	*huángyóu* 皇猷	*Song shu* 宋書	*ca.* 492–493	1	…大化，惠澤長久，萬民咸懷。故得皇猷丕丕，行於四方，揚於後代，以保…

131 Eine ausführliche Diskussion findet sich in Suwald 2008, S. 71–77.

132 *Zhongjing* 忠經. Unkommentierte Ausgabe auf *WikiSource*. URL: https://zh.wikisource.org/zh-hant/%E5%BF%A0%E7%B6%93 (besucht am 30. 03. 2019), Diese Version des Textes hat eine Länge von 2.499 Zeichen mit insgesamt 6.287 2–4-Gramm-*types*, von denen 326 im *DHYDCD* lexikalisiert sind. Zu 313 davon liegen chronologische Daten vor.

133 „忠臣之事君也，莫先於諫，下能言之，上能聽之，則王道光矣。" ebd., 辨忠章十四.

134 揚聖章第十三 ebd., „不足則補之，聖明則揚之，古之道也。".

135 Siehe Suwald 2008, S. 80, S. 82.

136 Siehe ebd., v. a. S. 78–80.

1. *jianzhang* 諫章 ist ein *false positive*, das durch die fehlende Segmentierung entsteht: *zhong jian* 忠諫 ist der Titel des 15. Kapitels.
2. *jianren* 鑑人 ist im Kontext als Verb-Objekt-Konstruktion zu verstehen.[138]
3. Bei *jingjie* 兢戒 handelt es sich ebenfalls um ein *false positive*.[139] *jingjing* 兢兢 ist bereits im *Shijing* 詩經 belegt,[140] *jie shen* 戒慎 im *Liji* 禮記.[141]
4. *chunzhi* 淳質 („rein und natürlich")[142] ist erst im *Sui shu* 隋書 nachgewiesen und kann als Indiz für eine Entstehung des *Zhongjing* deutlich nach der Han-Zeit gewertet werden.[143]
5. Dasselbe gilt für *kangyi* 抗義 („Einspruch")[144], das erst mit dem *Wei shu* 魏書 belegt ist.
6. Für die Kombination *hai li* 害理 (etwa: „Ordnungen zerstören")[145] ist das *Nanqi shu* 南齊書 als früheste Textstelle angegeben, was als weiterer Hinweis für eine spätere Entstehung gewertet werden kann.
7. *shouzai* 守宰 (etwa: „Gebietsverwalter")[146] ist der Titel des fünften *Zhongjing*-Kapitels. Da der Begriff mit dem *Hou Han shu* 後漢書 belegt ist, kann davon ausgegangen werden, dass er während der Han-Zeit bereits Verwendung fand.
8. *huangyou* 皇猷 („kaiserliche Vorhaben")[147] ist – mit Ausnahme des *Zhongjing* selbst – ebenfalls erst im *Song shu* 宋書 belegt.

Da auch das *Zhongjing* selbst als Belegquelle für ein Lexem angegeben sein kann, oder es aufgrund der traditionellen Datierung von den Herausgeber:innen des *DHYDCD* fälschlich als *Locus classicus* angenommen worden sein kann, müssen auch diejenigen Lexeme beleuchtet werden, die mit dem *Zhongjing* belegt sind (Tabelle 6.11).[148]

1. Für *bingzhi* 秉職 „an der Pflicht festhalten"[149] lässt sich mithilfe des *Chinese Text Project* eine Han-zeitliche Belegstelle im Text *San lüe* 三略 finden.[150]
2. *jingzhi* 敬職 („die Ämter mit Ehrerbietigkeit versehen")[151] lässt sich mit dem ebenfalls Han-zeitlichen *Qian fu lun* 潛夫論 belegen.[152]
3. *qianyun* 潛運 ist erst in einem Text von 魏收 Wei Shou (506–572) mit dem Titel *Xi Liang wen* 檄梁文 belegt. Die Textstelle wird im Tang-zeitlichen *Yiwen leiju* 藝文類聚 (erschienen 624) wiedergegeben.[153] Suwald bemerkt, dass der gesamte Ausdruck *chen mou qian yun* 沉謀潛

137 Angaben aus der diachronen Lexemdatenbank.
138 „惟天鑑人，善惡必應。" – „Nur der Himmel überblickt den Menschen, gut und böse werden gewiß beantwortet." Übs. aus Suwald 2008, S. 249.
139 Die Textstelle im *Zhongjing* lautet „是以兢兢戒慎，日增其明。" 2019, Kapitel 2 聖君章; „Wer deshalb auf vorsichtige Weise achtsam und aufmerksam ist, dessen Klarheit wird sich täglich steigern." Suwald 2008, S. 200.
140 *DHYDCD*, 兢兢.
141 *DHYDCD*, 戒慎; Siehe auch Suwald 2008, S. 200.
142 Suwald 2008, S. 78, siehe auch S. 238.
143 Vgl. auch ebd., S. 78.
144 Ebd.
145 Ebd., S. 126.
146 Ebd., S. 88.
147 Ebd., S. 78.
148 Im *DHYDCD* wird der Text Ma Rong 馬融 zugeschrieben. Sechs Lexeme werden aufgrunddessen in der Graphik dem 2. Jh. zugeordnet.
149 Suwald 2008, S. 78.
150 Huang Shigong 黃石公 ca. 100–9 v. Chr. *San lüe* 三略. Hrsg. von Donald Sturgeon. ctext.org.
151 Suwald 2008, S. 226.
152 Wang Fu 王符 ca. 102–167 v. Chr. *Qian Fu Lun* 潛夫論. Hrsg. von Donald Sturgeon. ctext.org.
153 Ouyang Xun 歐陽詢 et. al. 0624: *Yiwen leiju* 藝文類聚. Hrsg. von Donald Sturgeon. ctext.org.

運 („tiefliegende Pläne und verborgene Wendungen“) erst wieder in einem mingzeitlichen Text zu finden ist.[154]

4. Ebenfalls im *Yiwen leiju* ist die Phrase *zhigong wusi* 至公無私 („höchstes Allgemeinwohl und keine Eigensucht“[155]) zu finden. Zitiert wird aus dem deutlich älteren Text *[Shengxian] gaoshi zhuan [zan]* [聖賢] 高士傳 [贊] von Ji Kang 嵇康 (223–262)[156].
5. Der Begriff *zhongchen* 冢臣, „herausragender Untertan“[157] ist Titel des dritten *Zhongjing*-Kapitels. Die Herausgeber des *DHYDCD* finden erst im Qing-zeitlichen *Diaoqiao zhuang ge* 雕橋莊歌 eine weitere Belegstelle.[158]
6. Während die erste Textstelle mit *zhongneng* 忠能 eine Subjekt-Verb-Konstruktion („Die Loyalität vermag es…“) und damit ein *false positive* ist, kann die zweite Textstelle als Lexem (etwa: „[seine] Loyalitätskompetenz“)[159] gelesen werden. Eine weitere Belegstelle findet sich im *Yin Wen zi*, für dessen überlieferte Fassung aber eine spätere Datierung als die Han-Zeit angenommen wird.[160]

Tabelle 6.11 2–4-Zeichen-Kombinationen im *Zhongjing* mit inzestuösen Belegstellen

#	Lexem	belegt in:	datiert auf:[162]	Vork.	Kontext
1	*bǐngzhí* 秉職	*Zhongjing* 忠經	166	1	…行其政，居則思其道，動則有儀。秉職不回，言事無憚，苟利社稷，則不…
2	*jìngzhí* 敬職	~ ~	~	1	…以之而克則無怨，夫如是，則天下敬職，萬邦以寧。《詩》云「載馳載驅…
3	*qiányùn* 潛運	~ ~	~	1	…色直辭，臨難死節而已矣在乎潘謀潛運，正己安人，任賢以爲理，端委而…
4	*zhìgōngwúsī* 至公無私	~ ~	~	1	…所覆，莫大乎忠。忠者，中也，至公無私。天無私，四時行地無私，萬物…
5	*zhǒngchén* 冢臣	~ ~	~	2	…詩》云「昭事上帝，聿懷多福。」冢臣章第三爲臣事君，忠之本也，本立… …臣事君，忠之本也，本立而化成。冢臣於君，可謂一體，下行而上信，故…
6	*zhōngnéng* 忠能	~ ~	~	2	…心之謂矣。爲國之本，何莫由忠。忠能固君臣，安社稷，感天地，動神明… …者備矣，然後可以理人。君子盡其忠能，以行其政令，而不理者，未之聞…

Mit der obigen Analyse kann weder die Han-zeitliche Entstehung des Textes wider-, noch eine spätere Textgenese belegt werden. Die im Text gefundenen Lexeme liefern uns aber Hinweise, die aus lexikographischer Sicht für eine Entstehung des Textes nach der Han- und spätestens wäh-

154 Siehe Suwald 2008, S. 78.
155 Siehe auch ebd., S. 195. *gong* 公 und *si* werden bereits bei Han Fei 韓非 als Antonyme präsentiert.
156 Siehe Zhu Jinxiong 朱錦雄 2013: „*Lun* Ji *Kang ‚Shengxian gaoshi zhuanzan‘ zhong de ‚gao shi‘ fanxing* 論嵇康《聖賢高士傳贊》中的「高士」範型 (‚Diskussion des Begriffs *gaoshi* in Ji Kangs *Biographien heiliger gaoshi (etwa: untadeliger Menschen)*‘)“. In: 國立臺北教育大學語文集刊 *Guo li Taibei daxue yuwen jikan (Journal of Language and Literature Studies)* 24, S. 233–260, S. 236.
157 Suwald 2008, S. 202. Suwald vermutet eine Anspielung auf den *Zhou*-zeitlichen Begriff *zhongzai* 冢宰.
158 Im *DHYDCD* wird der Begriff schlicht als *dachen* 大臣 („Minister / hoher Beamter“) übersetzt. Siehe *DHYDCD*, 冢臣.
159 Suwald übersetzt hier mit „Loyalität und Fähigkeiten“ Suwald 2008, S. 209.
160 Yin Wen 尹文: *Yin Wen Zi* 尹文子. Hrsg. von Donald Sturgeon. URL: https://ctext.org/yin-wen-zi (besucht am 17. 02. 2020).
162 Vgl. *CBDB*, `text_data`.

rend der Tang-Zeit sprechen.[163] Um solche Schlussfolgerungen zu stützen bzw. zu entkräften, sind zusätzliche philologische, bzw. „sinologischere" Methoden erforderlich. Suwald betrachtet daher ausführlich den Inhalt des *Zhongjing*, die Verwendung tabuisierter Zeichen,[164] sowie Sui- (隨, 581–618) und Tang-zeitliche (唐, 618–907) Literaturkataloge – in denen das *Zhongjing* allerdings nicht aufgeführt wird.[165]

Durch die chronologische Darstellung der in einem Text enthaltenen Lexeme können temporale Textprofile Hinweise auf sprachliche Anachronismen liefern und so helfen, Fälschungen aufzudecken, die mit dem Ziel geschaffen wurden, „älter" zu erscheinen. Sowohl die Erwähnung von Ereignissen, die sich erst nach dem angeblichen Verfassen eines Werkes abgespielt haben, sowie die Verwendung von Wörtern, die erst später gebraucht wurden, sind Hinweise auf eine Fälschung, da ein Text in der Regel keine Lexeme enthalten kann, die neuer sind als der Text selbst. Die Tatsache, dass offensichtliche Anachronismen für Fälscher:innen relativ leicht zu vermeiden sind, spricht allerdings gegen ein erfolgreiches Entlarven von Fälschungen mit dieser Methodik. Es kann lohnender sein, auf die „most trivial details", die „involuntary signs"[166] zu achten. Dazu zählen bei Texten die Worthäufigkeiten bzw. die am häufigsten verwendeten Wörter, sowie die Häufigkeit der Verwendung bestimmter Worttypen.[167]

Je vollständiger und genauer die historische Lexemdatenbank ist, die den temporalen Textprofilen zugrunde liegt, desto besser können von der Software Anachronismen aufgespürt werden, während dies für menschliche Leser:innen nur mit Expert:innenwissen und akribischer Recherche machbar ist. Einzelne Wörter können so „clear dating implications"[168] mit sich bringen, jedoch sollte auch die Gesamtheit der Sprache eines Texts kritisch betrachtet werden, denn „viewed in isolation, an individual word generally cannot yield decisive dating implications."[169] Die umfassende Ergänzung der Zitate aus dem *DHYDCD* um frühere Belegstellen aus einem relativ kleinen Textkorpus verdeutlicht zudem, dass eine diachrone Lexemdatenbank niemals vollständig sein kann.[170]

Temporale Textprofile können beim Anfangsverdacht einer Fälschung helfen, weitere Indizien zu finden, die aber einzeln geprüft werden müssen. In vielen Fällen wird es sich schlicht um eine frühere Belegstelle für das gefundene Lexem, oder um eine semantisch abweichende Zeichenfolge handeln, die zuvor nicht in der Datenbank erfasst war.

Limitationen ergeben sich weiterhin durch die geringe Detailtiefe von 100 Jahren und vor allem durch die vor und während der Han-Zeit noch ungenaueren historischen Lexikalisierungsdaten. Für Probleme wie die Klärung der Autorschaft von Textteilen, z. B. der Kapitel 81 bis 120 des *Hong lou meng* 紅樓夢,[171] oder die Unterscheidung von Alttext- und Neutext-Versionen

163 Tatsächliche Song-zeitliche Lexeme konnten in dem Text nicht nachgewiesen werden. Begrenzt man die Analyse auf die Lexikologie, könnte es sich beim *Zhongjing* theoretisch immer noch um einen Han-zeitlichen Text handeln, der *Locus classicus* für die oben untersuchten Lexeme ist.

164 Siehe dazu auch Kapitel 4.3, S. 69.

165 Siehe Suwald 2008, S. 72–77. Auch das ist allenfalls ein Indiz, kein Beweis, dass der Text erst in der Song-Zeit entstanden ist.

166 Carlo Ginzburg 1989: *Clues, Myths, and the Historical Method*. Baltimore & London: Johns Hopkins University Press, S. 97, 118; zitiert in Allison et al. 2011, S. 24.

167 Siehe Allison et al. 2011, S. 2, S. 24. Allison et al. sprechen von *language action types* wie z. B. *FirstPerson*, für die von ihnen verwendete Software *DocuScape* definierte linguistische Kategorien.

168 Leonard Neidorf 2014: „Lexical Evidence for the Relative Chronology of Old English Poetry". In: *SELIM* 20, S. 7–48, S. 11.

169 Ebd., S. 35.

170 Siehe dazu Kapitel 5.5.4, ab S. 134.

171 Siehe z. B. Hu Xianfeng, Wang Yang und Wu Qiang 2014: „Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber". In: *Advances in Adaptive Data Analysis* 6. doi: 10.1142/S1793536914500125, S. 17–18. Die

des *Shangshu*, reichen die zur Erzeugung der Profile verfügbaren Daten nicht aus. Methoden aus der Stilo(chrono)metrie oder Verfahren wie die *PCA* sind zur Bearbeitung solcher Fragen besser geeignet – solange passende Vergleichstexte bzw. Trainingsdaten vorliegen.[172]

6.2.5 Automatisierte Datierung mit temporalen Textprofilen

In den Abschnitten 6.2.3 und 6.2.4 wurden einzelne temporale Textprofile mit sinologischem Vorwissen und unter Heranziehen zusätzlicher Quellen gedeutet. Sie können jedoch auch für eine automatisierte Schätzung des Zeitraums der Textgenese eingesetzt werden. In diesem Abschnitt erarbeite ich anhand eines Trainingsdatensatzes von 432 Texten aus dem *difangzhi* 地方誌 Korpus (*DFZ*)[173] eine Herangehensweise für die automatisierte Interpretation temporaler Textprofile. Der bereits in Kapitel 6.1 verwendete Testdatensatz aus 216 *DFZ* wird erneut zur Evaluierung verwendet. Die beschriebene Methodik wird anschließend zusätzlich mit Texten aus dem *XXSKQS*-Datensatz und den *zhengshi* 正史 erprobt.[174] In Tabelle 6.12 sind die Ergebnisse der durchgeführten Experimente zusammenfassend dargestellt.[175] Zur Einschätzung der Ergebnisse dienen auch hier die *Accuracy*, also der Anteil der dem korrekten Jahrhundert zugeordneten Texte und der *mean average error* (*MAE*) in Jahren. Der *mean error* D_{mean} für die Zuordnung eines Texts ist als Mittelwert der Differenz zwischen Anfang und Ende des datierten *chronon c* und dem in den Metadaten angegebenen Jahr der Veröffentlichung definiert. Bei einer Granularität der Datierung von 100 Jahren ergibt sich daraus ein Mindestwert von $D_{mean} = 50$ für einen korrekt datierten Text.[176]

Betrachtung von Lexemen

Im Optimalszenario einer vollständigen diachronen Lexemdatenbank und einer fehlerfreien Segmentierung des zu datierenden Texts ließe dieser sich in der Regel dem spätesten Jahrhundert zuordnen, aus welchem noch Lexem-*types* vorkommen. Im Hinblick auf die tatsächlichen frühesten Belegstellen aller Lexem-*types* eines unbekannten Texts muss aber von einer unvollständigen Datenbank ausgegangen werden. Zudem führt die vereinfachte *n*-Gramm-Segmentierung zu weiteren *false positives*. Texte können daher einen variablen Anteil an Lexemen enthalten, die später eingeordnet sind als der zu datierende Text. Dieser nimmt tendenziell mit dem Alter des zu datierenden Texts zu (Abb. 6.22).

Autoren argumentieren anhand einer *Support Vector Machine* (*SVM*)-Klassifizierung der einzelnen Kapitel, dass nicht nur die Kapitel 81–120, sondern wahrscheinlich auch Kapitel 67 nicht von Cao Xueqin 曹雪芹 (gest. 1763/4) stammen.

172 Siehe Kapitel 3.1, S. 40.
173 *DFZ*, siehe auch Kapitel 4.2, S. 66.
174 Siehe auch Kapitel 6.1, S. 175 u. 172.
175 Siehe S. 207.
176 Siehe Kapitel 6.1, S. 157.

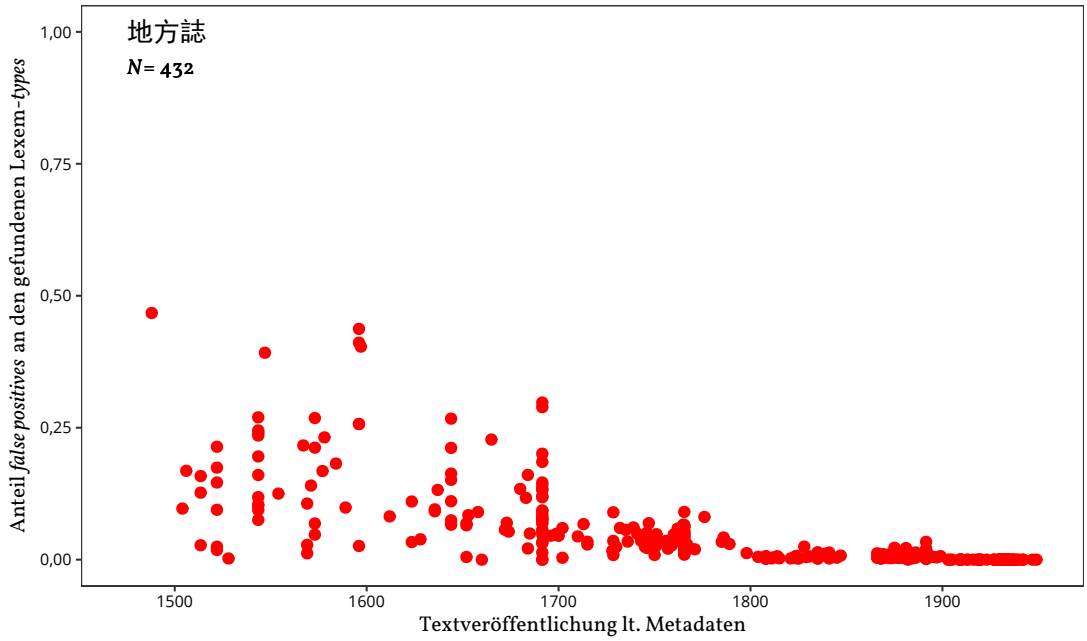


Abbildung 6.22 Anteil *false positives* (zu neu datierte Lexeme) nach Textveröffentlichung

Eine automatisierte Analyse von Profilen undatierter Texte kann sich daher nicht auf den erwarteten Anteil an *false positives* stützen. Zielführender ist es, den Anteil der Lexeme zu kalibrieren, der üblicherweise z. B. noch jeweils aus den Jahrhunderten vor und zur tatsächlichen Datierung festgestellt werden kann. Die Matrix in Abb. 6.23 zeigt anhand derselben Texte die Korrelationen zwischen der Gesamtzahl festgestellter Lexem-*types* und den *types*, die dem Jahrhundert der jeweiligen Textentstehung bzw. dem vorangegangenen Jahrhundert lexikographisch zugeordnet sind. Diese Berechnungen werden ohne, mit und mit *s*-Gewichtungskorrektur durchgeführt.

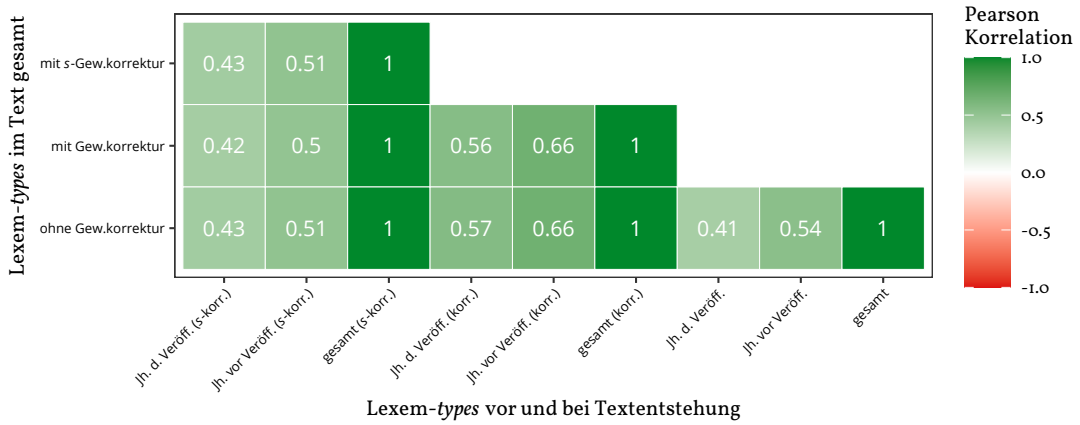


Abbildung 6.23 Korrelationsmatrix: *types* vor und zur Veröffentlichung und Gesamtanzahl *types*

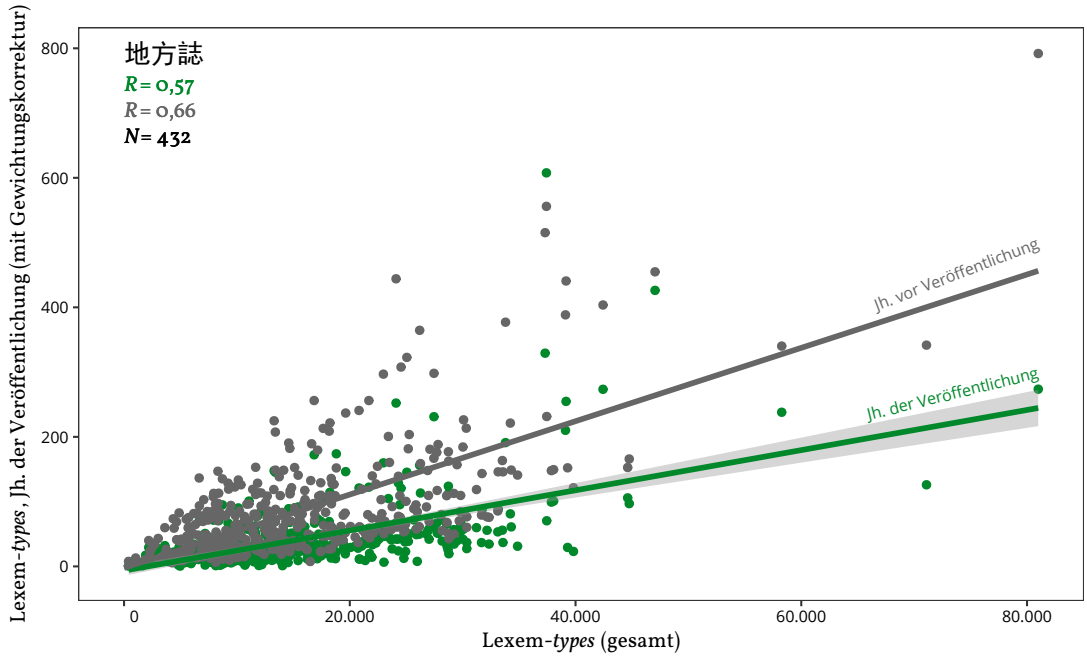


Abbildung 6.24 Korrelation Lexem-*types* zur und vor Veröffentlichung und Gesamtanzahl *types*

Die stärkste Korrelation R[177] von 0,66 ergibt sich dabei zunächst zwischen der Anzahl *types*, die bei Anwendung der linearen Korrektur dem Jahrhundert vor der Entstehung des Textes zugeordnet sind mit der Gesamtzahl der festgestellten *types*. Ohne bzw. mit *s*-Gewichtungskorrektur ist die Korrelation schwächer. Abb. 6.24 zeigt die Korrelation zwischen der Gesamtzahl der in den Texten enthaltenen 2–3-Zeichen Lexem-*types* T (*x*-Achse) und der Anzahl der *types*, die den Jahrhunderten vor und zur Veröffentlichung des jeweiligen Textes zugeordnet sind (*y*-Achse). An den unterschiedlichen Steigungen der Regressionsgeraden zeigt sich erneut, dass die pro Jahrhundert zugeordneten *types* in Richtung des Jahrhunderts der Textentstehung (grün) typischerweise abnehmend verlaufen.[178] Mittels der Funktion dieser Regressionsgeraden kann nun für einen Text mit T Lexem-*types* die für das Jahrhundert der Textentstehung erwartete Anzahl Lexem-*types* projiziert werden. Experimente mit den genannten Regressionsmodellen ergeben, dass ungeachtet der schwächeren Korrelation diese Herangehensweise zielführender ist, als die für das Jahrhundert vor der Entstehung des Textes erwarteten Lexeme zu projizieren.
Daraus resultiert folgende Herangehensweise zur automatisierten Datierung.

1. Die Anzahl der mit Gewichtungskorrektur für das Jahrhundert der Textentstehung c erwarteten *types* $t_{proj.}$ wird abhängig von der Gesamtzahl der gefundenen *types* T berechnet. Bei Verwendung der Lexemdatenbank mit zusätzlichen Korpusbelegen aus *zhengshi*,

177 Der Pearson-Korrelationskoeffizient R misst, wie gut die Messwerte zweier Merkmale in einem linearen Modell miteinander korrelieren.Dabei steht der Wert 1 oder –1 für eine perfekte Abhängigkeit zwischen den Merkmalen, ist der Wert 0, sind sie unkorreliert. Der Korrelationskoeffizient wird aus der Summe der quadrierten Standardabweichungen berechnet. Siehe z. B. Ludwig Fahrmeir et al. 2013: *Regression – Models, Methods and Applications*. Berlin & Heidelberg: Springer, S. 287.

178 Vgl. auch die Einzeldarstellungen in Abb. 6.18, S. 190 bzw. 6.25, S. 201.

Loewe und *DFZ*,[179] einer linearen Gewichtungskorrektur und bei Berücksichtigung von 2–3-Zeichen Lexem *types* ergibt sich:

$$t_{proj.} = 0,003 \times T - 6,513$$

Änderungen an der Betrachtung, also z. B. eine Erweiterung des *n*-Gramm-Raums auf 1–3- oder 2–4-Gramm-*types*,[180] sowie Anpassungen an oder Erweiterungen der Datenbank, erfordern eine Neuberechnung der Regression. Ohne Gewichtungskorrektur bzw. mit *s*-Gewichtungskorrektur ergeben sich ebenfalls eigene Werte für Steigung und Achsenabschnitt.

2. Aus dem Neologismusprofil mit Gewichtungskorrektur wird zunächst das Jahrhundert *c* mit der geringsten Differenz in der Menge zugeordneter *types* t_c zum errechneten Wert $t_{proj.}$ ausgewählt. Da wegen des negativen Achsenabschnitts bei Texten mit weniger als 2.000 *types* der Wert von $t_{proj.}$ unter 0 sinkt, wird als zusätzliche Bedingung ein Mindestwert von 5 festgesetzt. Andernfalls könnten einzelne *false* positives einen starken Einfluss auf das Datierungsergebnis nehmen.
3. Da die Profile mit (linearer) Korrektur der Gewichtung üblicherweise abnehmend verlaufen, werden anschließend die „benachbarten" Jahrhunderte betrachtet. Falls für die beiden vorangegangenen und nachfolgenden Jahrhunderte eine gegenläufige Tendenz zu beobachten ist, wird das Ergebnis *c* auf das späteste Jahrhundert korrigiert, für das $t_{proj.}$ bzw. 5 überschritten wird. Dies gilt also, wenn für eines der beiden vorherigen Jahrhunderte weniger, oder für eines der beiden späteren Jahrhunderte mehr Lexeme gemessen wurden als t_c.
4. Unter Berücksichtigung der bisherigen Erkenntnisse über die Möglichkeit von *Peaks* im Jahrhundert der Textentstehung, wird der Text überdies älter datiert, wenn für das vorangehende Jahrhundert ($c - 100$) deutlich mehr *types* nachgewiesen sind.[181]

Zur Veranschaulichung sei das temporale Profil eines Texts aus dem *DFZ*-Korpus gezeigt (Abb. 6.25). Die Veröffentlichung dieser *Guide fu zhi* 歸德府志 (*Chronik der Präfektur Guide*, heute Shangqiu 商丘, Henan 河南) wird mit 1754 angegeben.[182] Gemäß der Projektionsfunktion werden bei 17.643 im Datensatz festgestellten 2–3-Zeichen Lexem-*types* für das Jahrhundert der Entstehung des Textes 48,2 *types* erwartet. Der Wert mit der geringsten Differenz davon, 20,7, ist dem 18. Jh. zugeordnet, in dem der Text tatsächlich entstanden ist. Eine Umdatierung wegen eines gegen die Intuition verlaufenden Profils findet nicht statt. 82 *types* sind dem 17. Jh. zugeordnet, dem 19. Jh. nur 11,9. Ein *Peak* im 17. im Vergleich zum 18. Jh. besteht ebenfalls nicht.

Mit dem beschriebenen Algorithmus wird mit linearer Gewichtungskorrektur und 2–3-Gramm Lexem-*types*, der Testdatensatz aus 216 *Difangzhi* datiert. Für 47,2 % der Texte kann das Jahrhundert der Veröffentlichung bestimmt werden – bei einer durchschnittlichen Abweichung

179 Siehe Kapitel 5.5.4, ab S. 134. Verwendet man die zusätzlichen Belegstellen nicht, wird eine nahezu perfekte Korrelation mit $R^2 = 0,92$ zwischen Lexem-*types* (mit Gewichtungskorrektur) zum Jahrhundert der Veröffentlichung des Textes und der Gesamtzahl *types* erzielt – für Datierungszwecke ist dies allerdings dennoch nicht zuträglich, da kaum Diskrepanz zu „benachbarten" Jahrhunderten besteht.

180 Da nur 1–3-Gramme der *DFZ* vorliegen sind keine Experimente mit 2–4-Gramm-Daten möglich.

181 Bei den gewählten Parametern wird für „deutlich mehr" hier die vierfache Anzahl angenommen. Der beschriebene Algorithmus zum automatisierten „Lesen" der Neologismusprofile basiert auf der Betrachtung zahlreicher Profilverläufe und stellt lediglich eine von zahlreichen Möglichkeiten dar.

182 *DFZ*, # 0c87f43d7392c0589fbfb491e5165af9.

von 83,9 Jahren. Ohne Gewichtungskorrektur kann eine minimal bessere *Accuracy* von 48,6 % erreicht werden, der *MAE* erhöht sich jedoch auf 99,5 Jahre.[183]

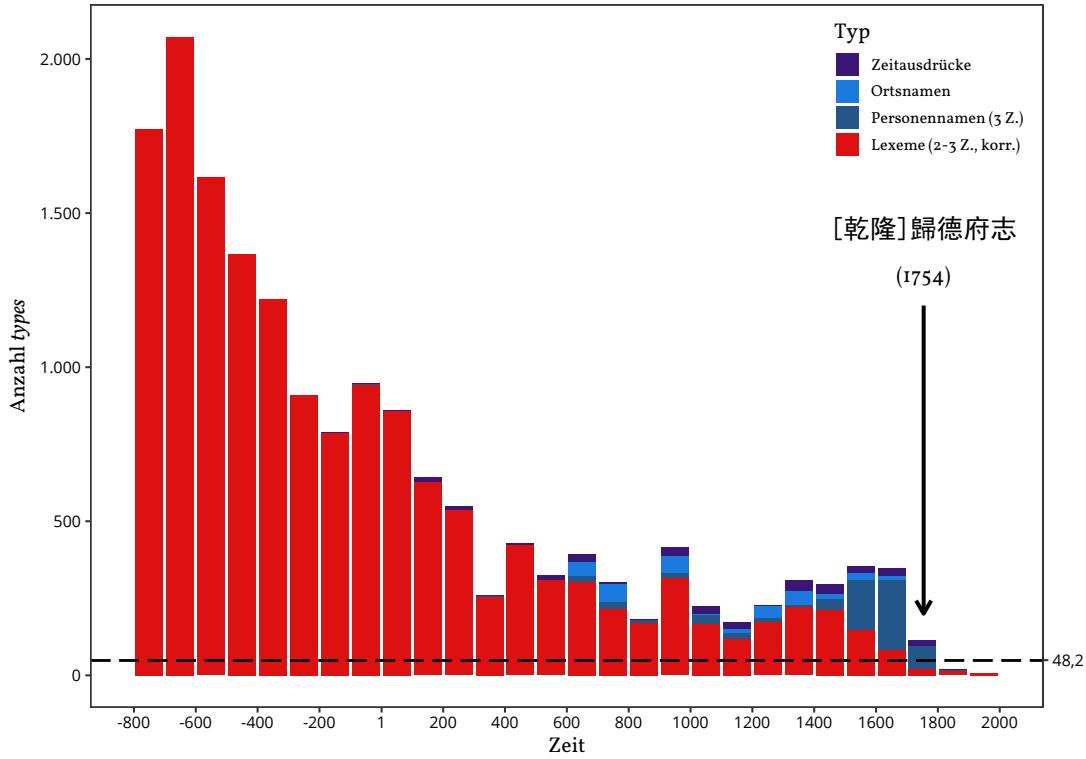


Abbildung 6.25 Temporales Profil des *Guide fu zhi* 歸德府志 von 1754

Berücksichtigung von Personennamen

Gerade in stilistisch alten Texten, die kaum zeitgenössisches Vokabular enthalten, können Personennamen wichtige Indizien für die Datierung liefern. Eine nur aufgrund der festgestellten Lexeme „zu alte" Datierung kann gegebenenfalls korrigiert werden, wenn Namen von Personen mit späteren biographischen Daten im Text genannt werden.[184] Zur Reduzierung von Ambiguitäten eignen sich dafür Namen mit einer Länge von mindestens drei Zeichen. Außerdem sollten diese zumindest in der *CBDB* eindeutig zugeordnet werden können, also nur eine einzige Person dieses Namens verzeichnet sein.[185] Trotz dieser Einschränkungen sind zwei Arten von *false positives* typisch. Personen gleichen Namens, über die kein Eintrag in der *CBDB* besteht, sowie Zeichenfolgen, die zufällig mit einem Namen übereinstimmen.[186] Ein Beispiel für ersteres aus dem *Guide fu zhi* ist Yang Zongji 楊宗稷, dessen biographische Daten in der *CBDB* mit 1865–1933

183 Siehe Abb. 6.27, S. 204; Tabelle 6.12, S. 207.
184 Aus der Nennung von Ortsnamen lassen sich bei der gegebenen Datenlage der *CBDB* kaum zuverlässige Erkenntnisse gewinnen.
185 Siehe Kapitel 4.7, ab S. 97.
186 Siehe auch Abschnitt 6.2.2, ab S. 189.

angegeben sind.[187] Im *Guide fu zhi* wird eine frühere Person desselben Namens als Teilnehmer an der Beamtenprüfung aufgelistet.[188]

Anders als bei Lexemen kann nicht davon ausgegangen, dass die Anzahl der jedem Jahrhundert zugeordneten Namen zum Jahrhundert der Textentstehung hin abnimmt. Texte können die Namen zahlreicher Zeitgenoss:innen nennen – gerade historiographische Texte können aber auch ausschließlich Personen aus früheren Jahrhunderten erwähnen.

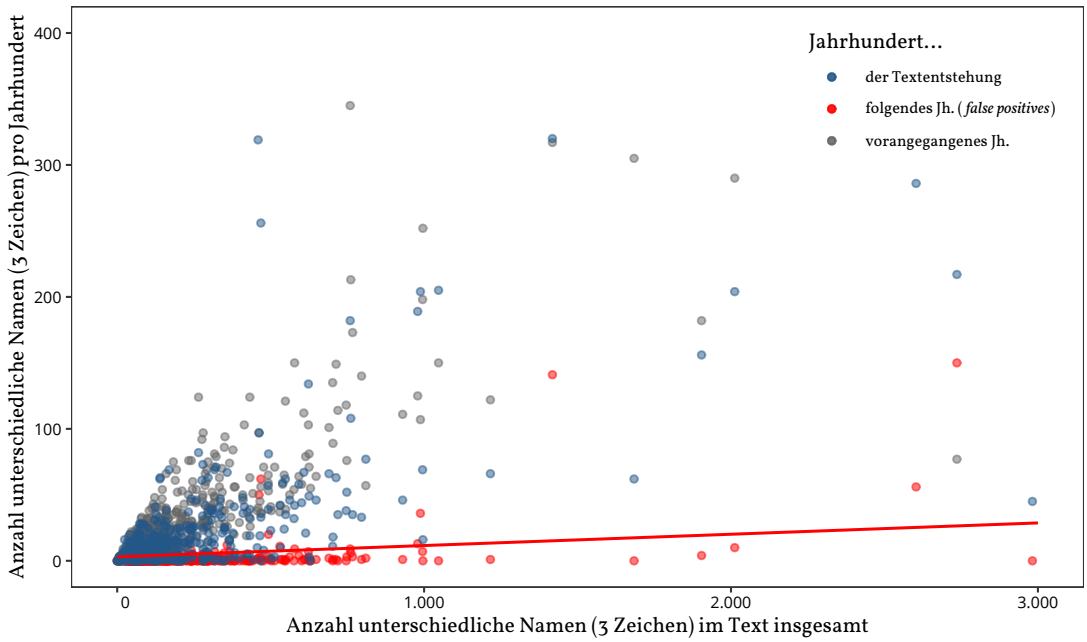


Abbildung 6.26 Namen in den Trainingsdaten

Abb. 6.26 zeigt die Anzahl unterschiedlicher 3-Zeichen Namen in den 432 Texten des Trainingsdatensatzes, die darin dem Jahrhundert der Veröffentlichung und den jeweils benachbarten Jahrhunderten zugeordnet sind in Relation zur Gesamtzahl der unterschiedlichen Namen *N* im Text. Insgesamt ist ein geringerer Anteil der Namen dem Jahrhundert der Textentstehung (blau) zugeordnet als dem vorangegangenen (grau). Dennoch lässt sich keine sinnvolle Abgrenzung vornehmen, da die Datenpunkte stark gestreut sind. Andererseits sind die Werte der *false positives* aus dem folgenden Jahrhundert (rot) deutlich niedriger. Im Gegensatz zu den Lexem-*types* sollte hier zudem kein Zusammenhang zwischen *false positives* und der Datierung des Textes (Abb. 6.22) bestehen.[189] Da unabhängig davon längere Texte potenziell eine größere Anzahl an *false positives* enthalten können, sollte ein Schwellenwert für die Anzahl von Namen, der eine Späterdatierung rechtfertigt, von der Textlänge abhängig gemacht werden. Eine Linearregression der für das Jahrhundert nach der Textentstehung festgestellten *false positives* auf *N* liefert zwar

187 *CBDB, ID* 76819. Die in der *CBDB* gemeinte Person war ein erfolgreicher *guqin* 古琴-Musiker.

188 Siehe Chen Yanglu 陳錫輅 und Cha Qichang 查岐昌, Hrsg. 2016 [1754]: *[Qianlong] Guide fu zhi 36 juan [乾隆] 歸德府志 36 卷 ([Qianlong] Chronik der Präfektur Guide, 36 juan)*. Online-Datenbank Diaolong 雕龍 / *Zhongguo Difang zhi* 中國地方誌, via CrossAsia. Nagoya 名古屋 & Taipeh 台北: Kaixi MS 日本凱希多媒體 & tts 大鐸資訊, S. 113–115.

189 Vgl. auch Kapitel 4.7, S. 102.

eine sehr schwache Korrelation, die Steigung kann aber zur textspezifischen Festsetzung eines Schwellenwerts n_t genutzt werden. Anstatt des Achsenabschnitts der Regression (0,25) wird ein Mindestschwellenwert n_θ von 3 festgelegt, um willkürliche Späterdatierungen zu minimieren. Aus den Trainingsdaten ergibt sich damit die Funktion:

$$n_t = 0,0086 \times N + 3$$

Auf dieser Basis kann anstelle des zuvor bestimmten Jahrhunderts c das späteste Jahrhundert als Zeitstempel angenommen werden, für das die Anzahl zugeordneter Namen n_t überschreitet und die Mindestanzahl von 5 Lexem-*types* noch erreicht wird.

Zur Veranschaulichung sei erneut das *Guide fu zhi* herangezogen (Abb. 6.25). Die 2–3-Gramme des Textes weisen 648 Übereinstimmungen mit eindeutigen 3-Zeichen Namen aus der *CBDB* auf. Der Wert von n_t liegt also bei 8,6. 74 Namen sind dem 18. Jh. zugeordnet – dem vorher bereits vergebenen, korrekten Zeitstempel. 6 Namen sind dem 19. Jh. zugeordnet – in diesem Fall kommt es also nicht zur Vergabe eines späteren Zeitstempels. Sowohl dem 19., als auch dem 20. Jh. sind mehr als 5 Lexem-*types* zugeordnet. Wären einem der beiden Jahrhunderte also 9 oder mehr *false positive* Namen zugeordnet, käme es zu einer falschen Späterdatierung.

Mit der oben beschriebenen Vorgehensweise kann anhand der zusätzlichen *NER*-Informationen ein höherer Anteil der 216 *DFZ* (62,5 %) bei einem *MAE* von 72,1 Jahren dem Jahrhundert der Veröffentlichung zugeordnet werden (Abb. 6.27; Tabelle 6.12). Voraussetzung für den Erfolg dieser Herangehensweise bleibt die Erwähnung von Zeitgenoss:innen bzw. ein geringer zeitlicher Abstand zwischen erzählter Zeit und dem Verfassen des Textes. Die Auflistung z. B. von Teilnehmern an lokalen Beamtenprüfungen in einigen *DFZ* schafft dafür eine gute Ausgangssituation, die für andere Textgattungen so nicht erwartet werden kann.

Betrachtung von *temporal expressions*

In Texten erkannte *temporal expressions* können – falls vorhanden – ebenfalls für die zeitliche Einordnung von Texten genutzt werden. Je vollständiger ein solcher Ausdruck ist, desto zuverlässiger verweist er eindeutig auf ein Jahr bzw. sogar ein bestimmtes Datum.[190] In einem Datensatz mit 1–3-Gramm-Häufigkeiten ist die Erkennung solcher Ausdrücke, die typischerweise 4–12 Zeichen lang sind, stark eingeschränkt. Möglich ist aber die Erkennung von Regierungsdevisen (meist zwei Zeichen), gefolgt von einer Ziffer. Bei Eindeutigkeit der Bezeichnung einer so erkannten Regierungsdevise kann ein solcher Ausdruck von drei Zeichen einem Jahr zugeordnet werden.[191] Um Falschzuordnungen durch auftretende *false positives* zu begrenzen, wird ein Schwellenwert von $t_\theta = 4$ unterschiedlichen Vorkommen (*types*) von Zeitausdrücken festgelegt.[192]

In einem Test mit den 216 *DFZ* können 88 % der Texte bei einem *MAE* von 57,5 Jahren korrekt zugeordnet werden, indem sie jeweils dem spätesten Zeitraum mit mindestens vier entsprechenden *temporal expressions* zugeordnet werden.[193] 4,6 % der Texte werden aufgrund von *false*

190 Zur Erkennung von *temporal expressions* in schriftsprachlichen Texten siehe Kapitel 4.8, ab S. 103.
191 Siehe Kapitel 4.8, ab S. 103.
192 t_θ wurde auf Grundlage der Trainingsdaten optimiert. 3,5 % der Texte enthalten 4 oder mehr *temporal expression false positives*, die dem Jahrhundert nach der Veröffentlichung zugeordnet sind. Mit einem niedrigeren Schwellenwert würden mehr Texte fälschlich später datiert, bei einem höheren Wert von t_θ wiederum deutlich weniger Texte noch dem Jahrhundert der Veröffentlichung zugeordnet.
193 Auch hier können *false positives* auftreten, vgl. auch Kapitel 4.8, S. 103.

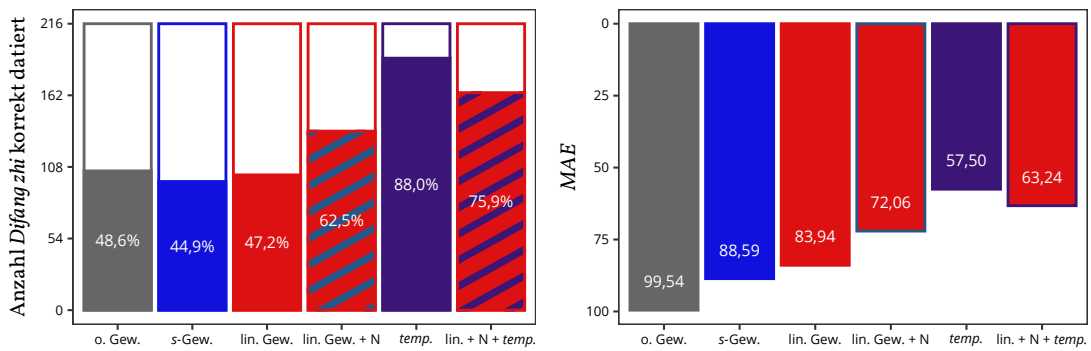
positives zu spät, 6,5 % zu früh eingeordnet. Zwei Texte (0,9 %) enthalten zu wenige *temporal expressions*.

Im Beispiel des *Guide fu zhi* (Abb. 6.25, S. 201) werden 20 unterschiedliche Jahresangaben erkannt, die dem 18. Jh. zugeordnet sind und in die Regierungszeiten der Kaiser Yongzheng 雍正 (reg. 1709–1722) und Qianlong 乾隆 (reg. 1733–1796) fallen. Dem 19. Jh. ist ein *false positive* zugeordnet, dem 20. Jh. keines.

Für den *DFZ*-Datensatz ist diese primitive Herangehensweise den bisher betrachteten überlegen, da sie als historiographische Texte einen hohen Anteil an *temporal expressions* enthalten. Die sehr hohe *Accuracy* hängt auch damit zusammen, dass Texte mit einem Abstand von mehr als 50 Jahren zwischen Veröffentlichung und erzählter Zeit von Trainings- und Testdaten ausgeschlossen sind, um störende Effekte durch spätere Editionen eigentlich älterer Texte zu vermeiden.[194] Da durch die Angabe der Regierungsdevise – anders als bei westlichen Texten – nur ein Zeitpunkt in der Vergangenheit, der Gegenwart oder der nahen Zukunft angegeben werden kann, eignet sich diese Herangehensweise aber grundsätzlich auch zur Datierung von anderen Textsorten. Dabei ist – ähnlich wie bei Personennamen – eine kritische Überprüfung der erkannten *temporal expressions* erforderlich.

Anstelle einer bloßen Betrachtung von *temporal expressions* können diese natürlich auch ergänzend zur Zuordnung auf Grundlage von Lexemen und Namen betrachtet werden. Dabei werden Texte später datiert, wenn einem späteren als dem bisher datierten Jahrhundert 4 oder mehr *temporal expressions* zugeordnet sind. Eine zu späte Datierung auf Basis von Namen oder Lexemen wird dabei nicht korrigiert. Andernfalls würden Texte primär auf die Zeit datiert, über die darin berichtet wird. Für den *DFZ*-Testdatensatz fallen die Ergebnisse mit einer *Accuracy* von 75,9 und einem *MAE* von 63,2 Jahren dann etwas schlechter aus als bei reiner Betrachtung temporaler Ausdrücke.

Ergebnisse



(a) Anteil richtig datierter Texte

(b) Durchschnittliche Abweichung in Jahren *MAE*

Abbildung 6.27 Performance profilbasierter Datierung, *Difangzhi*, 2–3 Zeichen-Lexeme

194 Siehe dazu Abschnitt 6.1.1, S. 158.

In Abb. 6.27 werden die Ergebnisse der beschriebenen Profildatierungen gegenübergestellt. Bei einer reinen Betrachtung von 2–3-Zeichen Lexemen mit und ohne Gewichtungskorrektur, sowie mit der in Abschnitt 6.2.1 eingeführten *s*-Gewichtungskorrektur zeigt sich, dass ein Ausgleich des *HYDCD-Bias* sich grundsätzlich positiv auf den *MAE* auswirkt. Die stark vereinfachende Annahme, dass der Wortschatz in jedem Jahrhundert gleich stark wächst (lineare Gewichtungskorrektur) führt dabei zu den besseren Ergebnissen.[195] Bei ergänzender Verwendung von Personennamen weicht die vorausgesagte Veröffentlichung dabei nur bei 6,4 % der untersuchten *Difangzhi* um mehr als ein Jahrhundert ab, die maximale Abweichung beträgt 277 Jahre. Diese Ergebnisse sind – ohne jede Berücksichtigung von Worthäufigkeiten – durchaus vergleichbar mit denen bei Verwendung von genrespezifischen statistischen Sprachmodellen, obwohl hier ein deutlich längerer Datierungszeitraum von 700 v. u. Z. bis ins 20. Jh. berücksichtigt wird.[196] Bei reiner Betrachtung von *temporal expressions* können die mit Abstand besten Ergebnisse erzielt werden. Sie übertreffen für diese besondere Textgattung sogar eine kombinierte Betrachtung von Lexemen, Namen und Zeitausdrücken.

Im vergleichbaren Ergebnis bei Verwendung eines statistischen Sprachmodells mit *NLLR* werden 59,7 % der Texte korrekt datiert. Der maximale Fehler liegt dann bei 306 Jahren, bei 7,4 % der Texte ist die Abweichung über 100 Jahre. Werden zusätzlich *temporal expressions* mit *NLLR*TE* betrachtet, können 64,4 % der Texte korrekt datiert werden.[197] Da bei der statistischen Datierung der *DFZ* mit einer *chronon*-Dauer von 50 Jahren gearbeitet wird, beträgt der minimale Fehler einer korrekten Datierung E_{min} bei einer korrekten Datierung nur 25 Jahre. Der *MAE* ist daher mit 41,5 bzw. 40,3 Jahren deutlich kleiner.

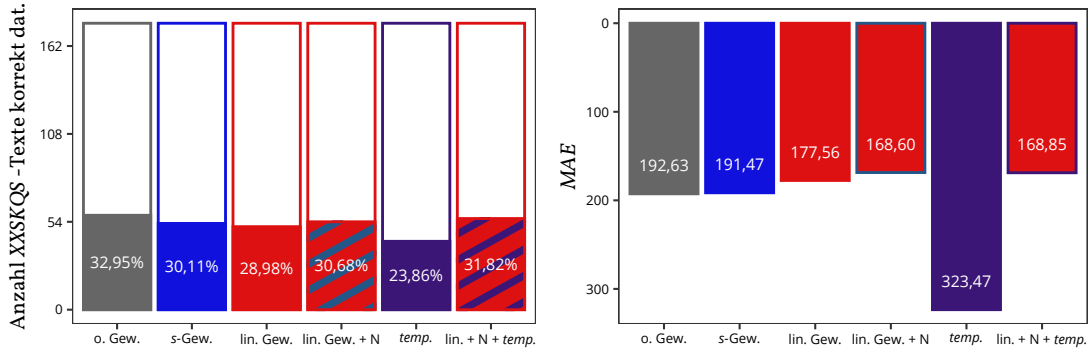
Abgesehen von der Zugänglichkeit für eine philologische Interpretation brauchen temporale Textprofile auch für die automatisierte Datierung den Vergleich mit statistischen Sprachmodellen nicht zu scheuen. Die Regression auf die Anzahl der für das Jahrhundert der Entstehung erwartbaren Lexem-*types* profitiert allerdings ebenfalls von spezifischen Trainingsdaten.

Experimente mit weiteren Korpora

Experimente mit weiteren Testdatensätzen sollen die Eignung der oben beschriebene Methodik für einen erweiterten Testzeitraum und andere Textgenres prüfen. Die Zuordnung von 176 Texten aus den *Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*)[198] zeigt, dass Texte anhand temporaler Textprofile auch dann ungefähr chronologisch eingeordnet werden können, wenn kein passendes Trainingskorpus verwendet wird (Abb. 6.28).[199] Durch die kombinierte Betrachtung von Lexemen, Namen und Zeitausdrücken kann noch eine *Accuracy* von 31,8 erreicht werden – beinah so viel, wie mit einem spezifisch trainierten statistischen Sprachmodell.[200] Der *MAE* liegt dabei zwar mit 168 Jahren deutlich höher, es wird hier aber innerhalb eines deutlich längeren Zeitraums und doppelter *chronon*-Länge datiert.[201] Aufgrund des Stils und vor allem Inhalts einzelner Texte des *XXSKQS*-Korpus kommt es zu Falschdatierungen mit einer Abweichung von bis

195 Erkenntnisse aus der Sprachwandelforschung und die aus dem *HYDCD* extrahierten Daten sprechen allerdings eher für ein *s*-förmiges Wortschatzwachstum, das dem Piotrowski-Gesetz folgt. Siehe dazu Kapitel 2.1 (ab S. 14) und 5.7.2 (ab S. 142).
196 Vgl. Kapitel 6.1, ab S. 156. Die verglichenen *SLM* umfassen den Zeitraum von 1475–1925.
197 Siehe Kapitel 6.1.1, Tabelle 6.1, S. 165, Beobachtungen #4 und #9.
198 Siehe S. 171.
199 Siehe S. 207.
200 Siehe Abschnitt 6.1.2, Tabelle 6.4, S. 171. Mit *SLM*s wurde mit *CS* und *tf-idf* eine *Accuracy* von 33,5 % erreicht.
201 Mit Sprachmodellen konnte mit demselben Datensatz ein *MAE* von 81 Jahren erzielt werden, wobei nur in *chronons* zwischen 1475 und 1925 klassifiziert wurde. Siehe S. 171.

zu 1.773 Jahren. So sind beispielsweise in dem laut Metadaten im Jahr 1823 veröffentlichte *Kaifang shuo* 開方說 des Mathematikers Li Rui 李銳 (1769–1817) keine Lexeme enthalten, die erst nach dem 2. Jh. nachgewiesen sind.[202]

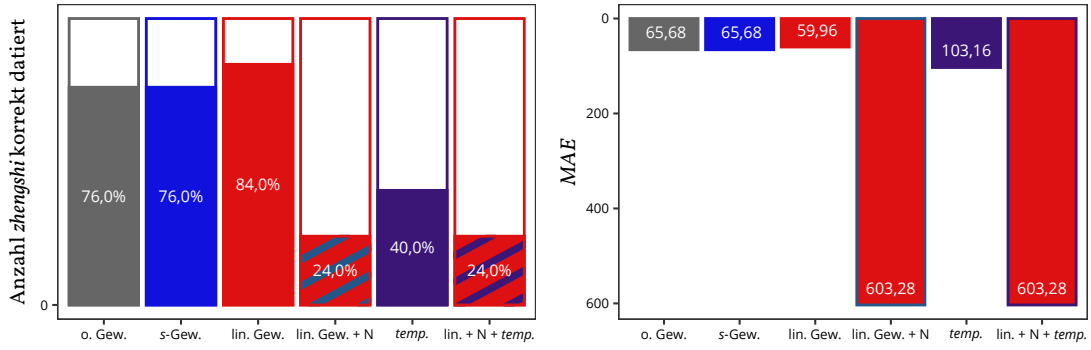


(a) Anteil richtig datierter Texte **(b)** Durchschnittliche Abweichung in Jahren *MAE*

Abbildung 6.28 Performance profilbasierter Datierung, *XXSKQS*, 2–3 Zeichen-Lexeme

Bei alleiniger Betrachtung von *temporal expressions* können nur 23,9 % der Texte dem richtigen Jahrhundert zugeordnet werden, da die untersuchten Texte mehrheitlich keine rezenten oder gar keine Zeitangaben enthalten.

Ein anderes Bild ergibt die Anwendung der Profildatierung auf die Dynastiegeschichten (Abb. 6.29). Dieses Korpus aus nur 25 Texten deckt einen Zeitraum von 91 v. u. Z. bis 1928 ab. Anders als die Datensätze der *DFZ*- und des *XXSKQS* stehen sie als Volltext zur Verfügung, was eine genauere Erkennung von *temporal expressions* ermöglicht.[203]



(a) Anteil richtig datierter Texte **(b)** Durchschnittliche Abweichung in Jahren *MAE*

Abbildung 6.29 Performance Profildatierung, *zhengshi*, 2–3 Zeichen-Lexeme / *temporal expressions*

202 Vgl. Li Rui 李銳 und Li Yingnan 黎应南 2000 [1823]: *Kaifang shuo* 開方說. Online-Datenbank Diaolong 雕龍 / *Xuxiu Siku quan shu* 續修四庫全書, via CrossAsia. Nagoya 名古屋 & Taipeh 台北: Kaixi MS 日本凱希多媒體 & tts 大鐸資訊, Weitere Beispiele für qing-zeitliche Texte, die keine oder kaum zeitgenössische Zeichenkombinationen enthalten, werden weiter unten ab S. 209, sowie in Abschnitt 6.1.3, ab S. 174 diskutiert.

203 Siehe dazu auch Kapitel 2.3, ab S. 20, sowie Kapitel 4.2, S. 65.

Bei der Betrachtung von Lexemen werden mit und ohne Gewichtungskorrektur sehr gute Ergebnisse erzielt, die bei Verwendung der linearen Korrektur am besten sind. Dies ist allerdings nicht repräsentativ, denn durch die intensive Nutzung dieses Korpus bei der Kompilation des *HYDCD* und das zusätzlich damit durchgeführte Training der Lexemdatenbank werden hier Optimalbedingungen ermöglicht,[204] die einer Identität von Trainings- und Testdatensatz ähneln.

Durch die Berücksichtigung von Namen werden die Ergebnisse deutlich verschlechtert. 76 % der Texte werden aufgrund von *false positives*, also Zeichensequenzen, die zufällig identisch mit Namen sind, sowie späteren Personen gleichen Namens, zu spät datiert.[205] Auch die Verwendung von *temporal expressions* ist weniger erfolgreich als bei den *DFZ*. Da ein größerer Zeitraum zwischen erzählter Zeit und Kompilation der Texte liegen kann,[206] werden die Texte häufig zu früh datiert. Die *Liao shi* 遼史 wurde z. B. 1343 fertiggestellt, die späteste darin erkannte *temporal expression* ist aber dem 11. Jh. zugeordnet.[207] Eine wegen *false positives* zu späte Datierung ist unwahrscheinlich, aber ebenfalls möglich. [208]

Zusammenfassung und Einschränkungen

Tabelle 6.12 Ergebnisüberblick der Datierungsexperimente aus 6.2.5

Korpus	Methode	*A* (%)	*MAE* (Jahre)	E_{max} (Jahre)	zu alt dat. (%)	zu neu (%)
DFZ	ohne Gewichtungskorrektur	48,6	99,5	2.815	38	13,4
	s-Gewichtungskorrektur	44,9	88,6	315,5	43,5	11,6
	lineare Gewichtungskorrektur	47,2	83,9	277	33,8	19
	+ Namen	62,5	72,1	277	17,6	19,9
	4+ *temporal expressions*	88	57,5	481	7,4	4,6
	kombiniert (linear + Namen + *temp.*)	75,9	63,2	277	1,9	2,2
XXSKQS	ohne Gewichtungskorrektur	33	192,6	2.890	38,6	28,4
	s-Gewichtungskorrektur	30,1	191,5	1.346	42	27,8
	lineare Gewichtungskorrektur	29	177,6	1.773	32,4	38,6
	+ Namen	30,7	168,6	1.773	29,5	39,8
	4+ *temporal expressions*	23,9	323,5	1.938	72,7	3,4
	kombiniert (linear + Namen + *temp.*)	31,8	168,9	1.773	28,4	39,8
zhengshi	ohne Gewichtungskorrektur	76	65,7	205	16	8
	s-Gewichtungskorrektur	76	65,7	205	16	8
	lineare Gewichtungskorrektur	84	60	211	12	4
	+ Namen	24	603,3	1.540	0	76
	4+ *temporal expressions*	40	103,2	340	56	4
	kombiniert (linear + Namen + *temp.*)	24	603	1.540	0	76

Die Ergebnisse der Experimente mit automatischer Datierung auf der Grundlage von temporalen Profilen werden in Tabelle 6.12 zusammengefasst. Die mit einem Datensatz von 432 *DFZ*

204 Siehe Kapitel 5.7.4, S. 150; siehe auch 5.5.4, S. 134.
205 Siehe dazu auch Kapitel 4.7, ab S. 97, sowie Abschnitt 6.2.2, ab S. 189.
206 Bei Auswahl der Texte aus dem *DFZ*-Korpus wurden Texte mit einem Abstand von über 50 Jahren zwischen Veröffentlichung und erzählter Zeit ausgeschlossen, um die Aufnahme späterer Auflagen zu minimieren.
207 Die Liao 遼 herrschten von 916–1125.
208 Das *Han shu* 漢書 enthält einige Zeitangaben mit den Äranamen *jianshi* 建始 (32–28 v. u. Z.). In der *DDBC* ist *jianshi* nur für die spätere Yan (*Hou Yan* 後燕, 384–409), eines der Sechzehn Reiche, verzeichnet. Das *Han shu* wird daher auf das 5. statt auf das 2. Jh. datiert.

trainierte automatisierte Analyse der temporalen Profile funktioniert grundsätzlich für alle drei Testkorpora zur ungefähren zeitlichen Einordnung der Texte. Erwartungsgemäß können die *zhengshi*, die bereits zur Erweiterung der zugrundliegenden Lexemdatenbank analysiert wurden, am besten zugeordnet werden. Für den *DFZ*-Testdatensatz können ebenfalls gute Ergebnisse erzielt werden. Selbst von den sehr heterogenen *XXSKQS*-Testdaten kann noch knapp ein Drittel korrekt zugeordnet werden, obwohl weder die Lexemdatenbank, noch der Profildatierungsalgorithmus mit diesem Datensatz trainiert wurden.

Werden nur Lexeme analysiert, zeigt sich anhand der *DFZ* und *XXSKQS*, dass Texte tendenziell eher zu alt als zu neu datiert werden. Durch Berücksichtigung von Namen lässt sich diese Tendenz ausgleichen. Eine Späterdatierung auf Basis von erkannten Personennamen kann jedoch nur funktionieren, wenn Namen von Zeitgenoss:innen der Verfasser:innen darin erwähnt werden. Bei den *zhengshi* führt das zu einer Überkompensierung und einer deutlich verschlechterten Performance, da in großer Zahl *false positives* auftreten. Angesichts der für das 1. Jahrtausend in der *CBDB* nur spärlich vorhandenen biographischen Daten und der fehlenden Möglichkeit einer zuverlässigen Tokenisierung bzw. *NER*, bedürfte die Betrachtung von Namen grundsätzlich einer Anpassung auf Spezifika des zu datierenden Textgenres.

Die Verwendung von *temporal expressions* erweist sich diesbezüglich als deutlich robuster – allerdings nur solange entsprechende Ausdrücke überhaupt vorkommen.

Mithilfe der Gewichtungskorrektur können für die *DFZ* extreme Abweichungen der Zeitstempel von der tatsächlichen Datierung vollständig verhindert werden. Bei einzelnen Texten aus dem *XXSKQS*-Datensatz kommt es allerdings auch damit noch zu massiven Fehldatierungen. Die Ursache hierfür sind Texte, die keine oder nur sehr wenige zeitgenössische Lexeme enthalten. Dies ist besonders problematisch, wenn ein Text weder zeitgenössische Personen, noch rezente Ereignisse referenziert. Durch die Beschränkung der Analyse auf Lexem-*types* mit 2–3 Zeichen und ein-eindeutige 3-Zeichen Namen aus der *CBDB* werden die verfügbaren Informationen zusätzlich reduziert.

Wie sehr sich solche schriftsprachlichen Texte allen Bemühungen um eine computerlinguistische Datierung entziehen können, sei am Beispiel des Qing-zeitlichen *Yuan shan* 原善 aus den *XXSKQS*-Testdaten veranschaulicht.

Die in den *XXSKQS* enthaltene Ausgabe dieses philosophischen Texts von Dai Zhen 戴震 (1724–1777) ist in den Metadaten auf das Jahr 1796 datiert.[209] Mit der oben beschriebenen Methodik würde er bestenfalls auf das 6. Jh. datiert, also um etwa 1.200 Jahre zu früh. Das temporale Profil des Textes (Abb. 6.30, ohne Gewichtungskorrektur) zeigt, wie diese Fehleinschätzung zustande kommt: Es sind keine zeitgenössischen Lexeme nachweisbar. Zudem werden im Text weder Namen von in der *CBDB* verzeichneten Personen genannt, noch lassen sich in den 2–3-Grammen des Textes *temporal expressions* feststellen. Auch die „neuesten" Lexem-*types* im Text (*hongju* 閎鉅, *cuanjue* 竄絕, *mingmei* 明昧 usw.) sind bereits in Ming-zeitlichen Texten nachgewiesen – ihre Anzahl ist aber so gering, dass es sich – ohne weitere Prüfungen – eben auch um in der Datenbank zu spät belegte Einträge handeln könnte. Dass das Profil bereits ab der Han-Zeit stark abflacht ist für einen Text aus dem 18. Jh. eher ungewöhnlich. Eine Ursache dafür ist Dai Zhens Argumentationsweise, der sich zur Darlegung seiner konfuzianisch geprägten

209 *XXSKQS*, # 9a0b80531e87b3dbfa56dfa1f5e7c3e4. Eine gedruckte Fassung des Textes existierte aber schon mindestens 1777. Siehe Cheng Chung-ying 成中英 2019 [1971]: *Tai Chen's Inquiry into Goodness: A Translation of the Yuan Shan, With an Introductory Essay*. Honolulu: University of Hawai'i Press, S. 50.

Standpunkte zahlreichen Zitaten aus Texten der klassischen Periode bedient, u. a. *Yijing* 易經, *Mengzi* 孟子 und *Zuo zhuan* 左傳.[210]

Auch die eigenen Textpassagen schreibt Dai allerdings in einer klassischen Sprache, die mit den statistischen Modellen aus Kapitel 6.1 sogar dem 4. Jh. *v. u. Z.* zugerechnet wird.[211] Damit dürfte der Text gegen jeden Versuch einer rein linguistischen Datierung quasi resistent sein.

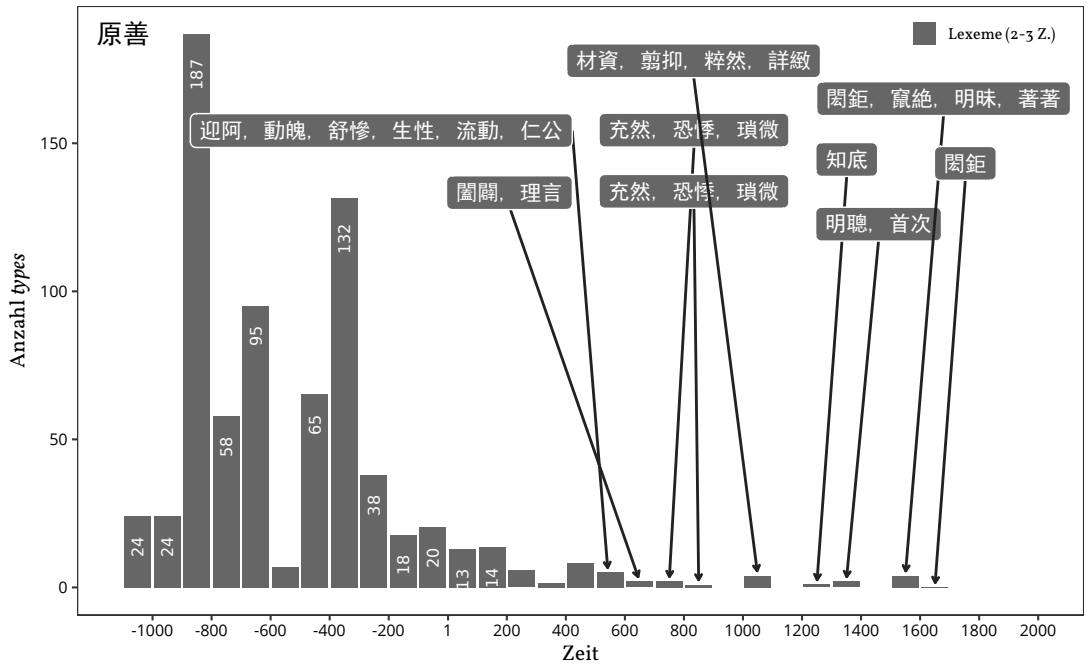


Abbildung 6.30 Temporales Profil des *Yuan shan* 原善

Auch wenn *Yuan shan* durch das Fehlen von Namen und temporalen Ausdrücken sicherlich ein Extrembeispiel darstellt, darf nicht vergessen werden, dass die Machart dieses Textes für die späte Kaiserzeit keineswegs außergewöhnlich ist. Das Beispiel erinnert an eine wichtige Limitation jeder linguistischen Textdatierung: „[T]he absence of any linguistic phenomenon in a book […] proves precious little."[212] Dass Harbsmeier diese Überlegung im Kontext der Datierung des *Lunyu* 論語 formuliert, deutet an, dass diese Problematik sich nicht auf die Datierung einiger spätkaiserzeitlicher Texte beschränkt, sondern für die gesamte Texttradition Relevanz hat.

Ob eine automatisierte Datierung anhand temporaler Profile erfolgreich sein kann, hängt stark von Inhalt und Stil des zu datierenden Textes ab. Das gilt ebenso für die Entscheidung, ob Lexeme, Namen und/oder temporale Ausdrücke betrachtet werden sollten. Unabhängig davon erlaubt die graphische Darstellung – anders als bei rein statistischen Methoden – immer noch eine den individuellen Besonderheiten eines Texts angepasste Interpretation.

210 Für eine ausführliche Darstellung siehe Cheng Chung-ying 成中英 2019 [1971], S. 50–51.
211 Siehe Abb. 6.5, S. 176. Datiert mit *NLLR* u. *NLLR*TE*, Modelle trainiert mit dem Zitatmaterial aus dem *DHYDCD*, vgl. S. 175.
212 Harbsmeier 2019, S. 207.

6.3 Datierung mit dem „durchschnittlichen Lexemalter“ von Texten

„History doesn't proceed in a linear way.“[213]
Francis Fukuyama

Mit den in Kapitel 6.1 und 6.2 vorgestellten Methoden werden Texte aufgrund ihrer Worthäufigkeiten bzw. der nachgewiesenen Lexeme stets vordefinierten Zeiträumen bzw. *chronons* zugeordnet. Intuitiv wird Zeit aber nicht als *kategorisch*, sondern als *kontinuierlich* wahrgenommen.[214]

Mit der auch in Kapitel 6.2 eingesetzten Datenbank lässt sich – wie bereits beschrieben – allen lexikalisierten Zeichenkombinationen eines Textes das (mittlere) Jahr ihrer ältesten Belegstelle zuordnen.[215] Daraus kann eine abstrahierte, absolute chronologische Messgröße für den Text berechnet werden – eine Art „durchschnittliches Wortentstehungsjahr“, im Folgenden durchschnittliches Jahr der Lexikalisierung, bzw. *Average Year of Lexicalization* (*AYL*). Ähnlich wie bei Überlegungen aus der Stylochronometrie, mit der Texte einer Autorin oder eines Autors per Regression auf sprachliche Merkmale datiert werden können,[216] soll untersucht werden, ob und wie mithilfe eines solchen *score* die Entstehungszeit von Texten auch als kontinuierliche Variable berechnet werden kann.

Das *AYL* als durchschnittliche mittlere Datierung *Y* der frühesten Belegstellen von *n* zu bewertenden Lexem-*types w* sei definiert als:

$$AYL = \frac{\sum_{w_1}^{w_n} Y}{n}$$

Zur Veranschaulichung wird die Berechnung des *AYL* für den Satz „昔者莊周夢為蝴蝶。“[217] erläutert. Enthalten sind darin 18 unterschiedliche 2–4-Gramm *types*, von denen drei im *DHYDCD* lexikalisiert und belegt sind:[218]

1. *hudie* 蝴蝶 – Die älteste angegebene Belegstelle stammt überraschenderweise aus einem Tang-zeitlichen Text, *Shi lin ji shi* 士林紀實.[219] Aus der *CBDB* sind die Lebensdaten des Autors Han Wo 韩偓 (842–914)[220] bekannt (siehe dazu auch 5.5.3, 132), so dass das Jahr 878 verwendet wird.[221]

213 Stephen Moss 2011: *Francis Fukuyama: 'Americans are not very good at nation-building'*. URL: https://www.theguardian.com/books/2011/may/23/francis-fukuyama-americans-not-good-nation-building (besucht am 15. 12. 2021).
214 Siehe auch Kapitel 3.3, S. 50.
215 Vgl. auch Kapitel 6.2, ab S. 179.
216 Siehe dazu Kapitel 3.1, S. 41.
217 „Einst träumte Zhuang Zhou, er sei ein Schmetterling.“ Zhuang Zhou 莊周 o. J. [ca. 3. Jh. v. u. Z.] *Zhuangzi* 莊子. *Digitale Ausgabe*. *Guoxue jingdian shuku* 國學經典書庫. Dongyang ligong daxue tushuguan 東洋理工大學圖書館, *juan* 2.14.
218 Die hier implizierten Verarbeitungsschritte sind auf S. 182 beschrieben, insb. Schritte 1–5.
219 *DHYDCD*, 蝴蝶.
220 Vgl. *CBDB*, S. ID 0094717.
221 Selbstverständlich läge mit dem Satz aus *Zhuangzi* 莊子 bereits eine deutlich frühere Belegstelle vor. Sehr wahrscheinlich ist der tatsächliche *Locus classicus* von *hudie* in noch älteren Texten zu finden. Vgl. George A. Kennedy 1964 [1955]: „The Butterfly Case, Part I“. In: *Selected Works of George A. Kennedy*. Hrsg. von Li Tien-yi. New Haven: Far Eastern Publications, S. 274–322, *passim*.

2. *xizhe* 昔者 – das *DHYDCD* gibt das *Yijing* 易經 als *Locus classicus* für *xizhe* an.[222] Der in der Datenbank enthaltene geschätzte ungefähre Entstehungszeitraum ist 850–800 v. .u. Z.,[223] es wird also der Wert -825 verwendet.
3. Zhuang Zhou *meng* 莊周夢 – dieser Ausdruck wird – wenig verwunderlich – mit dem *Zhuangzi* 莊子 belegt,[224] als mittleres Jahr des geschätzten Zeitraums der Entstehung wird hier 300 v. u. Z. angenommen.

Das *AYL* dieses einzelnen Satzes würde also berechnet als:

$$AYL = \frac{878 + (-825) + (-300)}{3} = 115\ v.\,u.\,Z.$$

Berechnet man das *AYL* für alle 2–4-Zeichen-Lexeme des gesamten *Zhuangzi* ergibt sich der Wert -155, für das deutlich ältere *Shangshu* 尚書 -459, für das Song 宋-zeitliche *Meng xi bi tan* 夢溪筆談 231. Das *AYL* kann keineswegs die Entstehungszeit eines Textes angeben, korreliert aber damit: Ältere Texte erhalten tendenziell niedrigere Werte. Mittels Linearregression kann dieser Zusammenhang zwischen *AYL* und Veröffentlichung zu datierender Texte formalisiert werden.

AYL und Textveröffentlichung – experimentelle Formalisierung

Die Eignung und optimale Verwendung des *AYL* zur Schätzung der Entstehungszeit von Texten muss experimentell angenähert werden, da zur Berechnung entweder der gesamte Wortschatz eines Textes, oder unterschiedliche Anteile oder Mengen seiner häufigsten Lexeme betrachtet werden können. Die so berechneten Werte werden mit dem tatsächlichen Jahr der Veröffentlichung von Texten eines bereits datierten Korpus korreliert. Je besser die erzielte Korrelation, desto besser der temporale Informationsgehalt des verwendeten Messwerts.

Zur Untersuchung dieses Zusammenhangs wird eine diachrone Textreihe benötigt, ein homogenes Korpus, das einen möglichst großen Zeitraum abdeckt. Hierfür bieten sich erneut die offiziellen Dynastiegeschichten (*zhengshi* 正史, siehe auch Kapitel 2.3, S. 20) an, deren Fertigstellung sich über den Zeitraum von 91 v. u. Z. bis 1928 erstreckt. Ein *Caveat* ist dabei – wie bereits in Kapitel 6.1 und 6.2 – die Überlagerung zweier temporaler Aspekte: Während die Texte – trotz ihrer strukturellen und stilistischen Anlehnung an das „Vorbild“ *Shiji* 史記 – sprachliche Merkmale aus der jeweiligen Zeit ihres Entstehens aufweisen und stilistische Einflüsse und Vorlieben der Autor:innen bzw. Kompilator:innen vorhanden sind,[225] enthalten sie doch in großem Umfang auch Lexeme, die spezifisch für den vorangegangenen, *erzählten* Zeitraum sind. Eine weitere Ungenauigkeit entsteht durch die Aufnahme von früherem Textmaterial. So wurde z. B. das *Hou Han shu* 後漢書 von Fan Ye 范曄 mehr als zweihundert Jahre *nach* Ende der Han-Dynastie zusammengestellt – allerdings größtenteils aus deutlich früher verfassten, tatsächlich Han-zeitlichen Dokumenten und Texten.[226] Eine Übersichtsdarstellung von inhaltlich abgedeckter Zeitperiode und Veröffentlichung der *zhengshi* findet sich in Kapitel 2.3.[227]

222 *DHYDCD*, 昔者.
223 Diese Angabe basiert auf den Ausführungen von Edward Shaughnessy, siehe Edward Shaughnessy 1993a: „*I ching* 易經 (*Chou I* 周易)“. In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 219.
224 *DHYDCD*, 莊周夢.
225 Siehe auch Kapitel 2.3, ab S. 20.
226 Siehe auch Kapitel 5.7.4. Zur Entstehungsgeschichte des *Hou Han shu* 後漢書 siehe Bielenstein 1954, S. 9–17.
227 Siehe Abb. 2.2, S. 22.

Abb. 6.31 zeigt die Korrelation zwischen *AYL* bei Berücksichtigung *aller* 2–4-Zeichen-Lexem-*types* (kurz $AYL_1^{2--4gram}$) und der Veröffentlichung der 25 *zhengshi* als Linearregression. Obwohl – gegeben durch die sehr unterschiedliche Länge der Korpustexte – die Anzahl der für die Berechnung des *AYL* berücksichtigten Lexeme bzw. Jahresangaben sehr unterschiedlich ist und zwischen 13.778 (*Chen shu* 陳書) und 73.174 Lexemen (*Song shi* 宋史) liegt, kann die Korrelation zwischen *AYL* und Veröffentlichung der Texte mit $R = 0,896$[228] bzw. $R^2 = 0,80$ als vielversprechend angesehen werden. Das *AIC* liegt bei 344,4.[229] Da das Jahr der Textgenese die später im Modell zu errechnende Variable ist, wird es hier auf der *y*-Achse dargestellt.

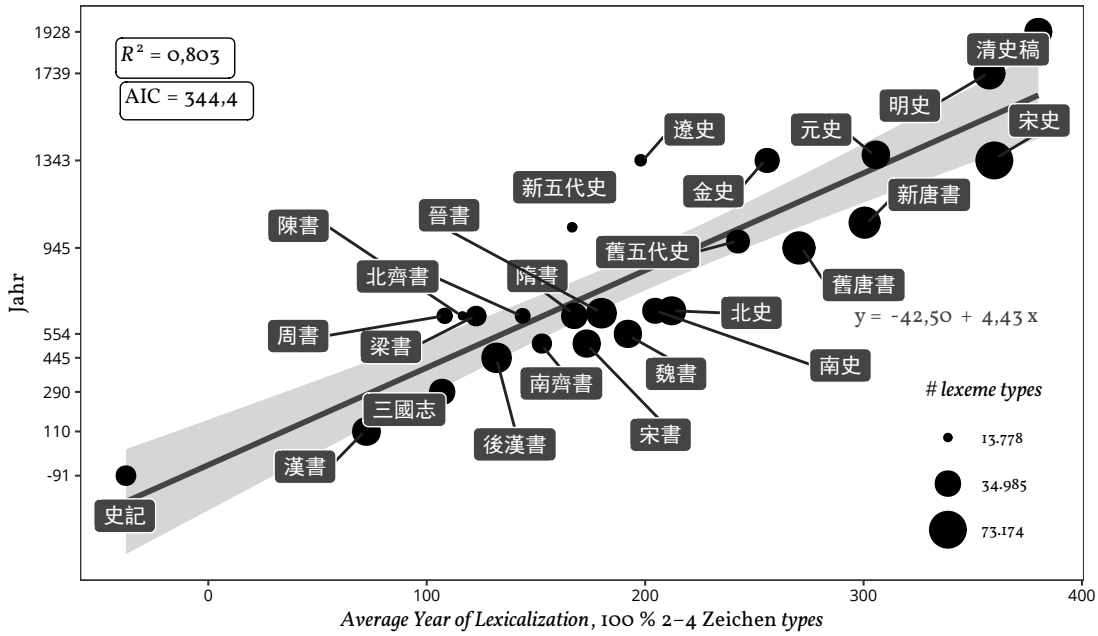


Abbildung 6.31 Korrelation Veröffentlichung *zhengshi*, *AYL* bei 100 % 2–4 Zeichen *types*

Korpus-Beobachtungen mit einer *PCA* legen nahe, dass nicht durch Betrachtung *aller*, sondern nur der *häufigsten types* eine höhere temporale Aussagekraft erzielt werden kann.[230] Zur Op-

228 Der Pearson-Korrelationskoeffizienten *R* misst, wie gut die Messwerte zweier Merkmale in einem linearen Modell miteinander korrelieren. Dabei steht der Wert 1 oder -1 für eine perfekte Abhängigkeit zwischen den Merkmalen, ist der Wert 0, sind sie unkorreliert. Der Korrelationskoeffizient wird aus der Summe der quadrierten Standardabweichungen berechnet. Siehe z. B. Fahrmeir et al. 2013, S. 287.

229 Akaike's *Information Criterion* (AIC, Akaikes Informationskriterium) ist ein Kriterium zur Auswahl statistischer Modelle, das 1973 von Hirotsugu Akaike vorgeschlagen wurde und das die Varianz der Residuen mehrerer Modelle vergleicht. Dabei ist das Modell mit dem geringsten *AIC* zu bevorzugen. Es eignet sich daher nur zum direkten, relativen Vergleich ähnlicher Modelle, da es kein absolutes Maß für die Qualität eines statistischen Modells darstellt. Siehe Jan deLeeuw 1992: „Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Hrsg. von Samuel Kotz und Norman L. Johnson. Bd. I: Foundations and Basic Theory. New York: Springer, S. 599–609, S. 607; bzw. die ursprüngliche Veröffentlichung: Akaike Hirotsugu 赤池弘次 1992 [1973]: „Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Hrsg. von Samuel Kotz und Norman L. Johnson. Bd. I: Foundations and Basic Theory. New York: Springer, S. 610–624.

230 Siehe auch Kapitel 3.1, S. 41; siehe auch T. Schalmey 2021, S. 254–255.

timierung der Korrelation kann das *AYL* für die *zhengshi* experimentell mit unterschiedlichen Parametern berechnet werden.[231] Folgende Optionen werden hierfür erörtert:

1. **Gewichtung**. Lassen sich durch die Berücksichtigung von Worthäufigkeiten bessere Ergebnisse erzielen, oder sollte lediglich betrachtet werden, *welche types* in einem Text vorkommen? Das *AYL* kann hierzu mit der Worthäufigkeit gewichtet werden (*Frequency Weighted Average Year of Lexicalization*, *FWAYL*).
2. **Gewichtungskorrektur**. Mit der bereits in Abschnitt 6.2.1 (ab S. 185) verwendeten *Gewichtungskorrektur* kann zudem ein *Standardized Average Year of Lexicalization* (*SAYL*) berechnet werden. Das durch die Kompilation des *DHYDCD* gegebene *Bias*[232] kann so reduziert werden, gleichzeitig findet aber auch eine Verzerrung der Daten statt.
3. Berücksichtigung von **Einzelzeichen**: Die Betrachtung von 2–4 Gramm-Lexemen im Vergleich zur Betrachtung von 1–4 Gramm-Lexemen. Dass für einen Teil der im *DHYDCD* lexikalisierten Einzelzeichen sehr frühe Belegstellen vorhanden sind, während spätestens ab dem Mittelalter nur noch wenig neue Schriftzeichen lexikalisiert werden[233] und die Ergebnisse von Kapitel 6.1 und 6.2 sprechen dafür, Vorkommen von Einzelzeichen aufgrund ihrer geringeren temporalen Entropie zu vernachlässigen.
4. **Anteil oder Anzahl?** Sollen *alle* Lexeme eines Textes, oder nur die häufigsten Lexeme für die Berechnung des *AYL* berücksichtigt werden? Sollte dafür stets dieselbe *Anzahl* an *types* berücksichtigt werden, oder ein festgelegter *Anteil*?
5. In diesem Kontext sollte zudem der Umgang mit der stark unterschiedlichen Länge der Korpustexte evaluiert werden. Während das *Chen shu* als kürzester Text „nur“ knapp 200.000 Schriftzeichen umfasst, hat die Geschichte der Song-Dynastie (*Song shi*) deutlich über 4 Mio. Zeichen.
6. Ist es zweckmäßig, **Interpunktion** in Texten beizubehalten, da sie teilweise Informationen zu Wortgrenzen enthält?

Gewichtung mit Worthäufigkeiten

Die in den Kapiteln 6.1 und 6.2 gemachten Beobachtungen implizieren, dass aufgrund der stilistischen Rigidität einiger schriftsprachlicher Textgattungen eine chronologische Einord-

231 Solche Experimente sind nicht unüblich. Vgl. z. B. Matthew L. Jockers 2013: *Macroanalysis: Digital Methods and Literary History*. Topics in The Digital Humanities. Urbana, Chicago und Springfield: University of Illinois Press, S. 63–104. Jockers zeigt in seinem Buch, wie durch die Betrachtung unterschiedlicher Messungen von (teils einzelnen) Worthäufigkeiten von Texten verschiedene Charakteristika sichtbar gemacht werden können. Er stellt fest, dass unterschiedliche Merkmale und Signale sich jeweils mehr oder weniger eignen, um Genres, Autoren, Geschlecht der Autorin bzw. des Autors, die Entstehungszeit von Texten oder die Texte selbst mit computerlinguistischen Mitteln voneinander zu unterscheiden. Vierthaler experimentiert mit dem *zhengshi*-Korpus, um die beste Metrik zur Genreunterscheidung von spätkaiserzeitlichen chinesischen Texten zu finden. Siehe Vierthaler 2016a, S. 8; Vierthaler bezieht sich dabei auf einen Beitrag von Christof Schöch. Ihn treibt die Frage um, welche Messwerte (z. B. bei unterschiedlicher Länge der Worthäufigkeitslisten) zu verwenden sind, um Autoren oder Genres im Rahmen einer *PCA* jeweils besser unterscheiden zu können. Christof Schöch 2012: „Author or genre? Assessing the quality of cluster analysis graphs in two-dimensional classification problems“. In: *The Dragonfly's Gaze: Computational analysis of literary texts*. URL: https://dragonfly.hypotheses.org/148 (besucht am 30. 12. 2018), Er kommt zu dem Ergebnis, dass sinnvoll ist, 750 oder mehr der häufigsten Wörter zu betrachten, um französische Stücke aus dem 17. Jh. nach Genre zu clustern. Dabei ließen sich Signale für Autorschaft und Signale für das Genre allerdings nicht klar trennen – dieses Beispiel zeigt, wie spezifisch am Ende eines solchen Experiments der Erkenntnisgewinn sein kann und dass am Ende für jeden unterschiedlichen Fall eigene Experimente durchgeführt werden sollten.

232 Siehe dazu Kapitel 5.7.2, ab S. 142.

233 Siehe dazu auch Kapitel 5.7, ab S. 138 bzw. Abb. 5.12, S. 147.

nung besser auf Basis der vorkommenden Lexeme, als über deren Häufigkeit funktioniert. Andererseits lassen sich über einen längeren Zeitraum hinweg durchaus Veränderungen z. B. von Häufigkeiten wichtiger Funktionswörter beobachten.[234] Auch wurde gezeigt, dass diese Veränderungen für die zeitliche Zuordnung von klassischen chinesischen Texten eine Rolle spielen können.[235] Es sollte daher geprüft werden, ob sich durch Einbeziehung ihrer Häufigkeit im jeweiligen Text als Gewicht für die einzelnen *types* eine stärkere Korrelation erzielen lässt. Hierzu wird das *Frequency Weighted Average Year of Lexicalization FWAYL* berechnet als:

$$FWAYL = \frac{\sum_{w_1}^{w_n} Y\times \mid t \mid}{\mid T \mid}$$

D. h. die Summe aller „Lexikalisierungsjahre" *Y* der enthaltenen Wort-*types w*, jeweils multipliziert mit der Häufigkeit des jeweiligen *type*s im untersuchten Text, geteilt durch die Gesamtanzahl *T* der *tokens t*. Die erzielte Korrelation zum Erscheinungsjahr der Korpustexte bei Berücksichtigung aller 2–4 Zeichen-Lexeme ist immer noch sehr gut, erreicht mit $R = 0,864$ bzw. $R^2 = 0,746$ aber nicht die des ungewichteten Modells (Tabelle 6.13, S. 215). Es bleibt für die *AYL*-Datierung eines Textes entscheidender, *welche* Wörter (häufig) darin vorkommen, als *wie häufig* diese Wörter im zu datierenden Text enthalten sind.

Eine mögliche Erklärung für die schwächere Korrelation ist, dass die *zhengshi* genretypische Lexeme aufweisen, die in allen Texten häufig sind und es zum Teil über den gesamten Betrachtungszeitraum (1. Jh. v. u. Z. bis 20. Jh.) auch bleiben.[236] Bei Berücksichtigung der Häufigkeit werden also auch zahlreiche *types* stark gewichtet, deren Häufigkeit relativ konstant ist. Hinzu kommt die Tatsache, dass alte Wörter tendenziell häufiger sind.[237]

Gewichtungskorrektur

Das *Bias*, das durch die Auswahl der Einträge und Belegstellen im *DHYDCD* entsteht, führt zu einer ungleichen Gewichtung der Lexikalisierungszeiträume.[238] Um den Effekt dieser ungewollten Gewichtung auf die Berechnung des *AYL* zu reduzieren, kann ein gewichtungskorrigiertes durchschnittliches Wortentstehungsjahr (*Standardized Average Year of Lexicalization*, *SAYL*) berechnet werden:

$$SAYL = \frac{\sum_{c_1}^{c_n} |V_c| \times g_c \times (c + 50)}{\sum_{c_1}^{c_n} |V_c| \times g_c}$$

234 Siehe dazu Kapitel 2.3, ab S. 20.
235 Siehe Kapitel 3.1, S. 41. Siehe auch Yamada Takahito 山田崇仁 2004.
236 Von 1.000 der häufigsten 2–4-Zeichen-Lexeme des ältesten (*Shiji* 史記), des neuesten (*Qing shi gao* 清史稿), sowie eines mittleren Korpustextes (*Jiu Tang shu* 舊唐書) treten 12 % in allen drei Texten auf, in *Shiji* und *Jiu Tang shu* sogar 30 %, *Jiu Tang shu* und *Qing shi gao* 25 %. Betrachtet man 1–4-Zeichen-Lexeme ergibt sich sogar eine Übereinstimmung von 50,3 % in allen drei Texten. Dass eine hohe Ähnlichkeit von Wortschatz und -häufigkeit innerhalb des *zhengshi* Korpus besteht, hat auch die Studie von Vierthaler bereits eindrücklich gezeigt. Siehe Vierthaler 2016a, z. B. S. 26. In einer *Principal Component Analysis* der 1.000 häufigsten Zeichen bilden die Texte eindeutige *Cluster* abseits von anderen untersuchten Genres.
237 Siehe auch Kapitel 6.2, S. 181.
238 Siehe Kapitel 6.2.1, ab S. 185.

Die Anzahl der jedem Jahrhundert *c* zugeordneten *types* ($|V|$), multipliziert mit dem Gewicht *g* des jeweiligen Jahrhunderts und dem mittleren (50.) Jahr, geteilt durch die genau so gewichtete Gesamtanzahl an *types*. Zur Gewichtung kann sowohl ein linear gleichförmiges, als auch ein *s*-förmiges Wortschatzwachstum angenommen werden.[239] In beiden Fällen lässt sich keine Verbesserung der Korrelation erzielen, bei Verwendung der linearen Gewichtungskorrektur verschlechtert sie sich leicht auf $R = 0,859$ (Tabelle 6.13), mit *s*-Gewichtungskorrektur bleibt sie mit $R = 0,895$ nahezu identisch (Tabelle 6.13). Zur Berechnung des *SAYL* müssen die Daten pro Jahrhundert aggregiert werden, um das ebenfalls nur auf Jahrhunderte genau berechnete *Bias* zur Gewichtung verwenden zu können. Es ist zu vermuten, dass die so entstehende Unschärfe eventuelle positive Effekte der Gewichtungskorrektur wieder ausgleicht.

Eine geringfügig kleinere Maximalabweichung im *SAYL* Modell mit *s*-Gewichtungskorrektur, rechtfertigt kaum die gegenüber dem *AYL* erhöhte Komplexität.[240]

Tabelle 6.13 Vergleich linearer Modelle

Modell	R	R^2	*AIC*	Δ_{max}	Regress.gerade
Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,896	0,803	344,4	510	$y = -42 + 4,4x$
Frequency Weighted Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,864	0,746	350,7	593	$y = 630 + 4,9x$
Standardized Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,859	0,738	351,5	585	$y = 1200 + 4,5x$
s-Standardized Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,895	0,8	344,7	489	$y = -210 + 4,5x$

Berücksichtigung von Einzelzeichen und Anteil verwendeter Lexeme

In den bisherigen *AYL*-Berechnungen wurden nur Lexem-Einträge mit 2–4 Zeichen Länge berücksichtigt. Verwendet man zusätzlich einzelne Zeichen, die im *DHYDCD* ebenfalls mit datierbaren Belegen aufgeführt sind (*dan zi tiaomu* 單字条目), stehen mehr verwertbare Daten zur Verfügung. Das *AYL* kann zudem auch dann berechnet werden, wenn ein Text kaum 2–4-Zeichen-Lexeme enthält. Gegen die Verwendung sprechen bisherige Erkenntnisse über die deutlich geringere temporale Entropie von Einzelzeichen. Ein großer Teil des heute verwendeten, standardisierten Zeichenrepertoires ist bereits sehr früh nachweisbar, während nach der Han-Zeit nur noch wenige Zeichen hinzu kommen.[241]

Um die Verwendung von Einzelzeichen für die Berechnung des *AYL* zu evaluieren, lohnt es sich, zeitgleich zwei weitere Aspekte zu betrachten. Wenn bestenfalls nur ein zu bestimmender **Anteil** der *häufigsten* Lexem-*types* verwendet werden sollte, kann dieser – abhängig von der Länge der zu berücksichtigenden Lexem-*types* – unterschiedlich ausfallen. Da die Korpustexte überdies unterschiedlich lang sind, führt aber ein fixer *Anteil* an Lexemen zu stark unterschied-

239 Erläuterungen und Berechnung der Gewichte siehe unter Abschnitt 6.2.1, ab S. 185.

240 Eine Diskussion zum Vorzug einfacherer Modelle im Sinne von Ockhams Rasiermesser findet sich z. B. in Malcolm Forster und Elliot Sober 1994: „How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions“. In: *The British Journal for the Philosophy of Science* 45.1, S. 1–35.

241 Siehe dazu Kapitel 5.7, v. a. S. 147. Eine völlig andere Situation ergibt sich, wenn man zu bestimmten Zeiten oder Orten verwendete Zeichenvarianten (*yiti zi* 異體字 u. ä.) betrachtet, wie sie in großen Zeichenwörterbüchern wie dem *Hanyu da zidian* 漢語大字典 (*Großes Lexikon chinesischer Schriftzeichen*) gesammelt werden. Vgl. dazu auch die etwas irreführende Darstellung in Best und Zhu Jinyang 2006, S. 208–209.

lichen Grundmengen für die Berechnung. Parallel sollte also die Verwendung einer fixen **Anzahl** an *types* geprüft werden.

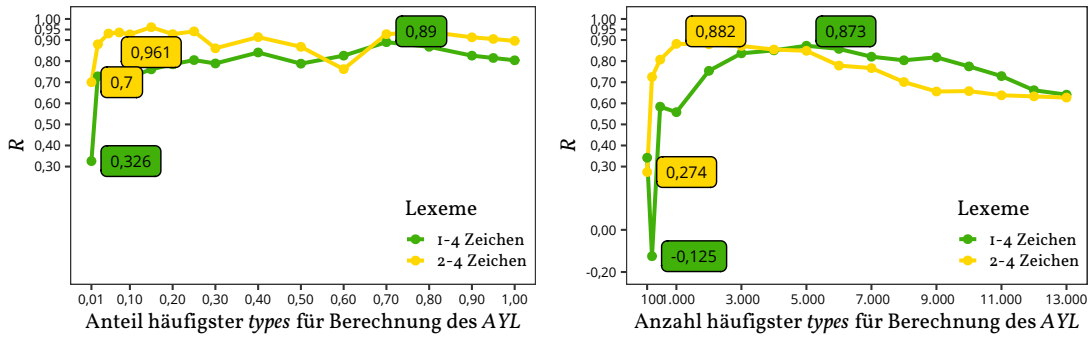


Abbildung 6.32 Vergleich linearer Modelle, *AYL* mit fixem *Anteil* und *Anzahl* von 1–4 und 2–4 Zeichen-Lexemen

Das *AYL* wird für unterschiedliche Anteile (1 %, 2,5 %, 5 %, 10 %, 15 %, 20 %, 30 %, … 90 %, 95 %, 100 %) und Mengen (100, 500, 1.000, 2.000, … 13.000[242]) an jeweils häufigsten Lexemen neu berechnet und *R* für die Korrelation mit dem Jahr der Veröffentlichung bestimmt. Um einen kompakten Überblick zu ermöglichen, zeigt Abb. 6.32 die Veränderung dieser Korrelation für lineare Regressionsmodelle mit einem steigenden Anteil (links) bzw. einer zunehmenden Anzahl berücksichtigter *types*.[243]

Ab etwa 1.000 bzw. 2,5 % der *types* lässt sich für den Zeitraum von 91 v. u. Z. bis zur Veröffentlichung des *Qing shi gao* 清史稿 1928 ein starker linearer Zusammenhang zwischen *AYL* und dem Jahr der Veröffentlichung herstellen. Wie erwartet lassen sich insgesamt mit 2–4-Zeichen-Modellen (gelb) bessere Ergebnisse erzielen als mit 1–4 Zeichen (grün). Die Verwendung eines *relativen Anteils* an häufigsten Lexemen (links) scheint überdies etwas bessere Korrelationen zu ermöglichen, als dies bei einer festgelegten *Anzahl types* (rechts) der Fall ist.

Für beide Fälle ist ein Optimalbereich der Berechnungsgrundlage für die temporale Aussagekraft des *AYL* erkennbar. Die stärkste Korrelation zur Veröffentlichung des Textes ergibt sich hier mit 15 % der häufigsten 2–4 Gramm *types*. Dabei werden zur Berechnung des *AYL* für das *Chen shu* 2.066, im Fall der *Song shi* 10.976 Lexemdatierungen zugrunde gelegt. Trotz dieser Diskrepanz wird mit relativen Anteilen eine bessere Korrelation erzielt. Bei Verwendung einer absoluten Anzahl – z. B. 2.000 – häufigsten *types* werden bei Betrachtung kürzerer bzw. weniger diverser Texte ungleich mehr seltenere Lexeme berücksichtigt, was ursächlich für die Verschlechterung der Korrelation sein kann. Bei Verwendung eines größer werdenden Anteils seltener *types* nehmen die Korrelationen allgemein wieder ab. Bei Modellen mit einer festgelegten Anzahl an 1–4 Grammen ist derselbe Effekt mit einer leichten Verschiebung zu beobachten, da bei der Betrachtung von Einzelzeichen potenziell erst mit mehr Daten sehr seltene *types* berücksichtigt werden.[244]

242 Um eine für alle Texte identische Anzahl zu ermöglichen, stellt die mit 13.778 von allen Korpustexten geringste Anzahl *types* des *Chen shu* die Obergrenze für dieses Experiment dar.

243 Auf die explizite Darstellung der insgesamt 64 Regressionsmodelle wie in Abb. 6.31 (S. 212) wird hier bewusst verzichtet.

244 Da die Texte nicht linear segmentiert sind, werden Einzelzeichen hier in jedem Vorkommen gezählt und nicht nur dann, wenn sie als eigenständiges Wort vorkommen.

Zur Ermittlung eines fixen Anteils der häufigsten Lexeme eines Textes müssen aus seinen *n*-Grammen zuerst *alle* infrage kommenden Lexeme ermittelt werden. Um Rechenleistung einzusparen, kann gleichermaßen auch direkt mit den *n*-Gramm-Frequenzlisten gearbeitet werden.[245] Sehr starke Korrelationen (*R* von 0,955, 0,967, 0,96) lassen sich bei Verwendung eines relativen Anteils im Bereich von 1 %, 3 %, 5 % an häufigsten *2–4*-Gramm-*types* erzielen. Da bei kürzeren Texten die Menge der tatsächlich berücksichtigten Lexeme mit steigender Anzahl berücksichtigter *n*-Gramm *types* immer weniger zunimmt, lässt sich auch mit einer festgelegten Anzahl z. B. der 200.000 häufigsten 2–4-Gramm-*types* eine annähernd gleich starke Korrelation mit *R* = 0,952 beobachten.[246]

Eine inzestuöse Korrelation?

Zwei Faktoren tragen hier sicherlich zur starken Korrelation bei: Als dynastiespezifische Geschichtstexte enthalten die *zhengshi* zwangsläufig eine große Anzahl zeitspezifischer oder -typischer Lexeme. Ein großer Teil der 25 Dynastiegeschichten ist zudem im *DHYDCD* mit zahlreichen Belegstellen präsent.[247] Dadurch wird automatisch ein Teil der in eben diesen Texten gefundenen Lexeme genau dem Zeitraum zugerechnet, in den der Text datiert werden soll. Das impliziert eine problematische Identität von Test- und Trainingsdaten, deren Auswirkungen sich in der Praxis aber als marginal erweisen. Zur Veranschaulichung sei hier das „Paradebeispiel“ des *Hou Han shu* (*HHS*) angeführt: Die verwendete Ausgabe enthält 43.705 unterschiedliche Lexeme von 2–4 Zeichen Länge.

Um die tatsächliche Auswirkung der „selbstevidenten“ Lexeme auf die Berechnung des *AYL* und damit auf die Korrelation zu bewerten, können für die Berechnung die knapp 8.000 *types* (etwa 18 %), für die das *HHS* selbst als erste Belegstelle angegeben wird, weggelassen werden. Bei Berücksichtigung aller *types* würde das *AYL* tatsächlich stark beeinflusst. Da es sich bereits als zielführend erwiesen hat, nur die häufigsten *types* eines Texts zu betrachten, ist die Hebelwirkung der dann verbleibenden „inzestuösen“ *types* auf die Berechnung des *AYL de facto* allerdings überraschend gering.

Verwendet man z. B. 3 % Prozent der häufigsten 2–4-Gramme zur Berechnung des *AYL*, sind lediglich etwa 8 % der verwendeten Wort-*types* des *HHS* „selbstevident“.[248] Die übrigen *zhengshi* wurden von den Kompilator:innen des *DHYDCD* weniger häufig als *Locus classicus* herangezogen, so dass der Effekt noch geringer bzw. nahezu bedeutungslos wird.[249]

245 Da – im Gegensatz zu 2–*n*-Grammen – ein sehr hoher Anteil der vorkommenden Zeichen im *DHYDCD* lexikalisiert und belegt sind, kann mit einer geringeren Anzahl häufigster 1–*n*-Gramme eine gute Korrelation zum Erscheinungsjahr der Texte erreicht werden. Insgesamt können mit 2–4-Grammen aber bessere Korrelationen erzielt werden als mit 1–4-Grammen.

246 Bei einer Betrachtung von 2–4-Grammen haben die untersuchten Texte zwischen 349.232 und 6.614.725 unterschiedliche *n*-Gramm *types*. Werden für alle Korpustexte nur die häufigsten 1 % *n*-Gramm-*types* untersucht, sind etwa 16 % davon tatsächlich im *DHYDCD* lexikalisiert. Durchschnittlich 19.000 *n*-Gramm-*types* enthalten etwa 3.000 Lexem-*types* pro Korpustext. Die *n*-Gramm *type-token*-Relation (*TTR*, Diversifikationsquotient) der Texte liegt zwischen 145.795 und 211.005 *types* pro 100.000 Zeichen. So enthält die verwendete Ausgabe des *Shiji* 史記 569.000 Zeichen, die etwas über eine Million 2–4-Gramm-*types* bilden. Untersucht man die häufigsten 1 % bzw. 10.300 davon, sind davon wiederum 1.595 (15 %) im *DHYDCD* lexikalisiert und belegt.

247 Siehe Kapitel 5.7.4, ab S. 150. Insbesondere bei *Shiji* 史記, *Han shu* 漢書 und *Hou Han shu* 後漢書 ist Vorsicht geboten, da es sich um die am häufigsten im *DHYDCD* als *Locus classicus* angegebenen Texte handelt – auch einige andere Texte des *zhengshi*-Korpus sind aber häufig im *DHYDCD* zitiert.

248 660 von 8.079 Lexem-*types*, die chronologisch zugeordnet werden können.

249 z. B. sind bei Betrachtung der 3 % häufigsten 2–4-Gramme nur 132 von 11.000 erkannten Lexemen des *Jiu Tang shu* 舊唐書 mit eben diesem Text im *DHYDCD* belegt. Der Einfluss auf die Berechnung des *AYL* ist damit marginal. Bei anderen *zhengshi* fällt der Einfluss teilweise noch geringer aus.

Berechnet man die oben gezeigten 2–4-Gramm-Modelle für das *AYL* unter Ausschluss der jeweiligen *Locus classicus*-Lexeme erneut, ist tatsächlich nur eine marginale Verringerung von *R* um etwa 0,01 bis 0,03 zu beobachten. Wie zu erwarten ist die Auswirkung auf Modelle, die auf einem geringen prozentualen Anteil basieren besonders gering, z. B. von 0,967 auf 0,956 bei Verwendung von 3 % der häufigsten 2–4-Gramme. Mit 100.000 der häufigsten 2–4-Gramm-*types* verschlechtert sich der Korrelationskoeffizient *R* von 0,95 auf 0,926. Insofern kann Entwarnung gegeben werden: Der experimentelle Ausschluss von „selbstevidenten" Lexemen zeigt, dass die Verortung der *zhengshi* im *DHYDCD* Korrelationen insgesamt nur marginal begünstigt. Die Korrelationen zwischen *AYL* und der Entstehung von Texten bestehen auch unabhängig von diesem Einfluss. Da in einem „realen" Anwendungsfall der zu datierende Text möglicherweise weder bekannt, noch im *DHYDCD* als *Locus classicus* angegeben wäre, bleibt diese Art des Ausschlusses eine theoretische Überlegung.

Nutzt man hingegen die mit zusätzlichen Belegen aus dem *zhengshi*-Korpus erweiterte diachrone Lexemdatenbank zur Berechnung des *AYL*,[250] kann erwartungsgemäß eine weitere Verbesserung der Korrelation erreicht werden. Mit 1,5 % der häufigsten 2–4-Gramme der *zhengshi* wird ein theoretischer Korrelationskoeffizient von $R = 0{,}968$ erreicht. Für die Datierung der Dynastiegeschichten käme das endgültig einer Identität von Trainings- und Testdaten nahe – zum Zweck der Datierung anderer, nicht im *DHYDCD* verorteter Texte kann diese Erweiterung hingegen sinnvoll sein.

Umgang mit der unterschiedlichen Länge der Texte

Die Korrelation zwischen *AYL* und Veröffentlichung von Texten bei Verwendung eines fixen Anteils an häufigsten Lexemen scheint auch bei stark unterschiedlicher Länge der betrachteten Texte verhältnismäßig stabil zu sein. Zusätzlich kann geprüft werden, wie sich die Zerteilung der Korpustexte in gleich lange Abschnitte auswirkt.[251] Da die verwendete Version des *Chen shu* 176.000 und einige weitere Texte ebenfalls kaum mehr als 200.000 Zeichen aufweisen, wird hier mit Abschnitten zwischen 10.000 und maximal 150.000 Zeichen Länge experimentiert. Die Werte für *R* aus diesen Experimenten werden in Tabelle 6.14 aufgeführt.

Mit Abschnitten von 10.000 Zeichen können die Werte für das *AYL* der einzelnen Abschnitte sehr weit auseinander liegen. Würde die Entstehungszeit des *Qing shi gao* mithilfe einer Linearregression auf das jeweils niedrigste und höchste *AYL* der einzelnen Textpartitionen geschätzt, ergäbe sich ein Spielraum von über 2.000 Jahren, von ca. 360 bis zum Jahr 2450. Bei Verwendung längerer Textabschnitte nimmt diese Spannbreite ab und die Korrelation zu: Mit Abschnitten von 50.000 Zeichen kann ein Wert von $R = 0{,}835$ (100.000: 0,857, 150.000: 0,877) erreicht werden. Die einzelnen Abschnitte des *Qing shi gao* würden auf den Zeitraum zwischen 680 und 2280 datiert.

Wird anstatt der Einzelbeobachtungen für die jeweiligen Textpartitionen der *AYL*-Mittelwert bzw. Median für alle Abschnitte eines Textes berechnet, können stärkere Korrelationen auch schon mit kürzeren Abschnitten ab 10.000 Zeichen erzielt werden.[252] Ein sehr guter Wert von *R* kann bei Verwendung eines optimierten Anteils von ca. 5–10 % der häufigsten 2–4 Gramme und Abschnitten ab 50.000 Zeichen erzielt werden.

250 Siehe dazu Kapitel 5.5.4, ab S. 134.

251 Diese Vorgehensweise ist auch bei anderen quantitativen Untersuchungen üblich. Vgl. z. B. Binongo und Smith 1999, S. 448; zitiert in Vierthaler 2016a, S. 7–8. Vierthaler unterteilt seine Texte in Abschnitte von je 10.000 Zeichen. Der letzte Abschnitt wird dabei in aller Regel verworfen.

252 Mit dem Mittelwert aller Beobachtungen werden minimal bessere Korrelationen erzielt als mit dem Median.

Tabelle 6.14 Korrelationskoeffizient (*R*) für Anteile häufigster 2–4-Gramme aus 10.000–150.000-Zeichen Abschnitten (unter „all“ wird *R* bei Betrachtung aller Einzelbeobachtungen angegeben)

Abschn.	10.000 Zeichen			50.000 Zeichen			100.000 Zeichen			150.000 Zeichen		
% 2–4 Gramme	all	mean	med.	all	mean	med.	all	mean	med.	all	mean	med.
5 %	0,684	0,921	0,919	0,795	0,951	0,937	0,822	**0,962**	0,949	0,852	**0,959**	0,948
10 %	0,722	**0,935**	0,926	0,808	**0,955**	0,942	0,83	**0,957**	0,944	0,862	**0,954**	0,949
20 %	0,739	0,932	0,926	0,816	0,948	0,933	0,838	0,947	0,931	0,868	0,944	0,941
30 %	0,746	0,925	0,926	0,819	0,939	0,928	0,839	0,938	0,924	0,867	0,936	0,934
40 %	0,751	0,918	0,916	0,825	0,935	0,916	0,842	0,931	0,919	0,872	0,933	0,933
50 %	0,761	0,923	0,92	0,826	0,932	0,913	0,846	0,93	0,916	0,874	0,931	0,931
60 %	0,767	0,921	0,918	0,827	0,927	0,908	0,848	0,928	0,915	0,875	0,929	0,927
70 %	0,77	0,921	0,918	0,828	0,924	0,903	0,851	0,928	0,915	0,878	0,924	0,92
80 %	0,772	0,92	0,916	0,833	0,925	0,908	0,852	0,925	0,915	0,877	0,923	0,915
90 %	0,772	0,919	0,918	0,834	0,924	0,91	0,853	0,923	0,915	0,877	0,922	0,918
100 %	0,777	0,921	0,92	0,836	0,924	0,908	0,855	0,921	0,913	0,877	0,92	0,915

Insgesamt verschlechtert das Partitionieren der Texte bei Betrachtung der *AYL*-Mittelwerte für ausreichend lange Textabschnitte die Korrelation zwischen *AYL* und Textgenese nicht. Eine Verbesserung, die diese Maßnahme rechtfertigen würde, stellt sich aber ebenfalls nicht ein. Dass ein *R* von 0,935 bereits mit dem *AYL*-Mittelwert aller Partitionen von 10.000 Zeichen erreicht werden kann, deutet aber an, dass eine temporale Aussagekraft des *AYL* für die chronologische Einordnung von Texten bei dieser Textlänge bereits gegeben ist.

Löschung von Interpunktionszeichen

Die in chinesischen *DH*-Anwendungen verbreitete Entfernung der Interpunktionszeichen[253] wirkt sich kaum auf die Korrelation des *AYL* mit der Textgenese aus. Da Interpunktion zumindest an Satzenden *false positives*, die sonst durch die verwendete *n*-Gramm-Segmentierung entstehen, verhindert, sind Vorteile durch die Löschung der Interpunktion im gegebenen Kontext aber auch nicht erwartbar. Beim Vergleich von *n*-Gramm-Modellen mit und ohne Entfernung der Interpunktion lassen sich minimale Unterschiede bei der Korrelation feststellen. Diese sind vermutlich überwiegend auf eine eine Verschiebung des Optimalbereichs für den den Anteil der zu betrachtenden *types* zurückzuführen, da bei gleicher Länge der betrachteten *n*-Gramm-Häufigkeitslisten nach Entfernen der Interpunktion mehr vermeintliche Lexeme erkannt werden.[254] Die Entfernung der Interpunktion aus allen Texten mag für gemischte Korpora, d. h. solche, die sowohl Texte mit, als auch ohne Interpunktion enthalten, allerdings sinnvoll sein, um Verzerrungen vorzubeugen.

253 Siehe z. B. Vierthaler 2016a, S. 7.

254 Da etliche zuvor gebildete, häufige *n*-Gramme mit ，。 und 、etc. wegfallen, können nach Entfernung der Interpunktion in den (z. B.) 100.000 häufigsten *n*-Grammen mehr tatsächliche Lexeme ausgemacht werden. z. B. sind 8.522 der häufigsten 100.000 2–4-Gramme der interpungierten Version des *Shiji* 史記 im *DHYDCD* lexikalisiert. Nach Entfernung der Interpunktion sind es 9.350, da häufige 2–4-Zeichen Kombinationen mit Interpunktionszeichen wie „曰「“, „。」“, „也。“ oder „之。“ wegfallen und tatsächliche Wörter „nachrücken“.

Zusammenfassung

1. Die stärksten Korrelationen zwischen *AYL* und Entstehung von Texten lassen sich für das betrachtete *zhengshi*-Korpus mit der Berücksichtigung eines optimierten Anteils von etwa 15 % der häufigsten 2–4 Zeichen Lexem-*types* bzw. ca. 1–3 % der häufigsten 2–4-Gramme erzielen.
2. Statt eines festen Anteils häufigster Lexeme kann auch ein fester Anteil oder eine Anzahl an häufigsten *n*-Grammen festgesetzt werden.[255]
3. Die Korrelation verschlechtert sich sowohl bei der Betrachtung von zu vielen seltenen *types*, also auch dann, wenn zu wenige *types* berücksichtigt werden.
4. Vorkommende Einzelzeichen-*types* (Unigramme) eignen sich weniger zur Datierung der Texte als 2–4-Gramm-*types*, da die meisten Schriftzeichen bereits sehr früh lexikalisiert wurden.[256]
5. Die Gewichtung von Lexemen mit der Häufigkeit ihres Auftretens führt ebenfalls nicht zu stärkeren Korrelationen zwischen Textgenese und (*FW*)*AYL*. Dennoch ist die Worthäufigkeit ein sinnvolles Kriterium für die Auswahl der zu betrachtenden *types*.
6. Ein Ausgleich des *DHYDCD-Bias* durch eine Gewichtungskorrektur bietet einen Mehrwert für Visualisierungen,[257] ihr Nutzen für die Berechnung des (*S*)*AYL* wird aber durch die entstehende Unschärfe eliminiert.
7. Als Durchschnittswert ist das *AYL* robust gegenüber unterschiedlichen Textlängen bzw. Mengen an *types*, aus der es berechnet wird. Eine Partitionierung von Texten in Abschnitte gleicher Länge ist daher nicht unbedingt erforderlich. Mit dem arithmetischen Mittel der *AYL*-Berechnungen für gleich lange Textpartitionen ab einer Länge von 10.000 Zeichen kann ebenfalls eine starke Korrelation zur Entstehung der Texte festgestellt werden.
8. Die Entfernung von Interpunktionszeichen ist nur zu empfehlen, wenn Texte mit und ohne Interpunktion miteinander verglichen werden sollen.

6.3.1 Ein optimiertes *AYL*-Regressionsmodell

Nach Analyse der Parameter für die Berechnung des *AYL* kann ein lineares Regressionsmodell mit optimierter Korrelation zwischen *AYL* und Entstehung der Korpustexte bei Berücksichtigung von 15 % der häufigsten 2–4-Zeichen-Lexeme detaillierter betrachtet werden. Abb. 6.33 zeigt dieses Modell mit $R = 0{,}961$ und einem *AIC* von 320,7.

Steigung (4,22) und Achsenabschnitt (*Intercept* 639,8) sind beide hoch signifikant.[258] Mit der Regressionsgeraden des Modells (Abb. 6.33) kann das *AYL* der häufigsten Lexem-*types* auf die ungefähre Zeit der Entstehung $Y_{proj.}$ eines geeigneten Eingabetextes projiziert werden. Mit 15 % der häufigsten 2–4-Zeichen-Lexemen also: $Y_{proj.} = 4,22 \times AYL + 639,8$.

255 Da bei einer *n*-Gramm-Segmentierung die Häufigkeit der vorhandenen Lexem-*types* erst nach Ermitteln aller *n*-Gramm-*types* berechnet werden kann, ist die Betrachtung von Lexem-*types* mit Performanceeinbußen verbunden.

256 Siehe v. a. auch Abb. 5.11, S. 147.

257 Siehe dazu Abschnitt 6.2.1, ab S. 185.

258 Das Modell wird in *R* mit `lm` (*linear model*) berechnet. *t*: Die Nullhypothese besagt: *AYL* hat keinen Zusammenhang mit dem tatsächlichen Entstehungsjahr des Textes. Je niedriger *P* ist, desto geringer ist die Wahrscheinlichkeit, dass die Nullhypothese stimmt. Bei Werten kleiner als 0,05 kann man die Nullhypothese „kein Zusammenhang" in der Regel verwerfen. Die Signifikanzwerte *p* belaufen sich für das oben gezeigten Modell auf $2e^{-16}$ für den Achsenabschnitt und $2,47e^{-14}$ für die Steigung.

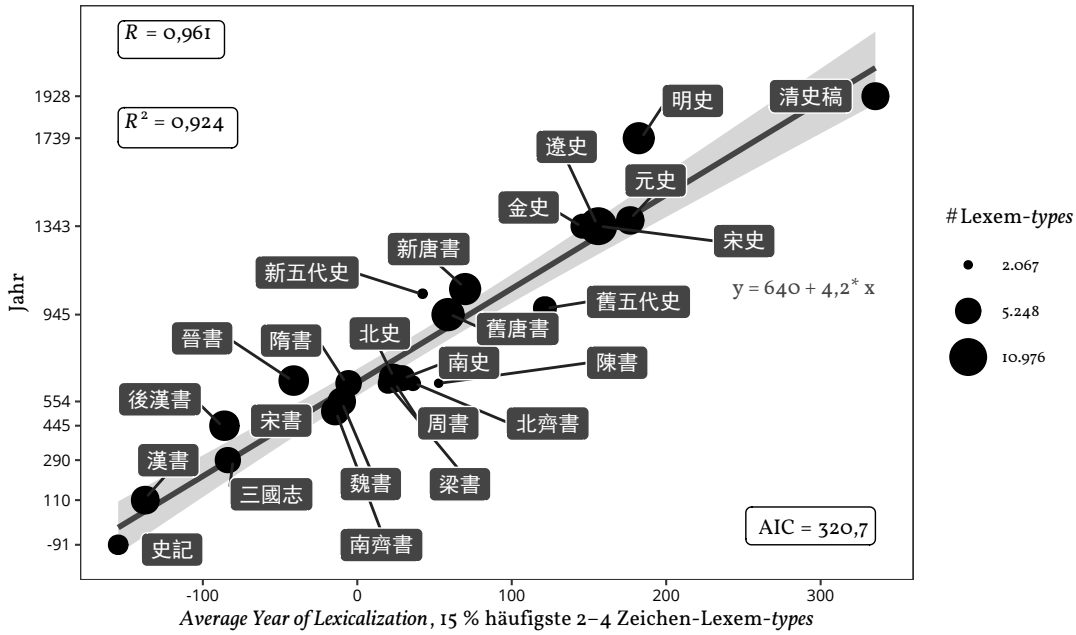


Abbildung 6.33 Korrelation Veröffentlichung *zhengshi*, *AYL* mit 15 % 2–4 Zeichen Lexem-*types*.

Der Standardfehler dieser Regression beträgt 136,6 [Jahre].[259]

Die Diagnoseplots in Abb. 6.34 ermöglichen ein genaueres Verständnis des Regressionsmodells.[260] Die Darstellung der Residuen (oben links) macht deutlich, welche Texte in dem verwendeten Modell am stärksten von der gewünschten Datierung abweichen. Die stärkste Abweichung vom Idealbild des Modells stellt dabei die *Ming shi* 明史 dar, die auf das Jahr 1409 datiert würde – 330 Jahre zu früh. Diese Datierung fällt aber in den im Text beschriebenen Zeitraum (1368–1644). Der erst 1739 veröffentlichte Text wurde zudem bereits 1645 in Auftrag gegeben und trotz der großen, teils herrschaftspolitisch bedingten Verzögerung naturgemäß aus Ming-zeitlichen Materialen kompiliert.[261] Im Hinblick darauf, dass die Geschichte der Ming also im Verhältnis zur beschriebenen Zeitperiode spät fertiggestellt wurde,[262] passt die Abweichung im Modell zu Inhalt und Textgeschichte. Dieser Erklärungsansatz passt auch zu den übrigen, deutlich zu alt datierten *zhengshi*: *Xin Wudai shi* 新五代史, *Jin shu* 晉書 und *Hou Han shu* 後漢書. Die in der Datierung um 221 Jahre abweichende *Xin Wudai shi* wurde von

259 Der *Residual standard error* gibt die Wurzel des durchschnittlichen Quadrats der Residuen an, d. h. hier die ungefähre durchschnittliche Abweichung des Alters eines zu datierenden Textes vom Idealbild der Regressionsgeraden.

260 *R* erzeugt mit der Funktion `plot` vier Diagnoseplots: *Residuals vs. Fitted* (Residuen gg. angepasste Werte), *Normal* Q-Q (Verteilung der Residuen), *Scale-Location* (ähnlich wie bei *Residuals vs Fitted* werden die Residuen gg. die angepassten Werte dargestellt, erstere werden allerdings normalisiert), sowie *Residuals vs Leverage* (Residuen und ihre Hebelwirkung auf das Modell anhand des Cook'schen Abstands). Die hier wiedergegebenen Plots wurden aus ästhetischen Gründen mit *ggplot2* erzeugt. Zur Erzeugung von Standard-Diagnoseplots mit `ggplot` siehe Raju Rimal 2014: *Diagnostic Plots using ggplot2*. Website. URL: `https://rpubs.com/therimalaya/43190` (besucht am 07. 11. 2018).

261 Siehe Thomas Wilson 1994: „Confucian Sectarianism and the Compilation of the Ming History“. In: *Late Imperial China* 15, S. 53–84. DOI: `10.1353/late.1994.0002`, S. 62; siehe auch Edward L. Farmer et al. 1994: *Ming History: An Introductory Guide to Research*. Ming Studies Research Series 3. Minneapolis: Center for Early Modern History, University of Minnesota, S. 72.

262 Siehe dazu auch Abb. 2.2, S. 22.

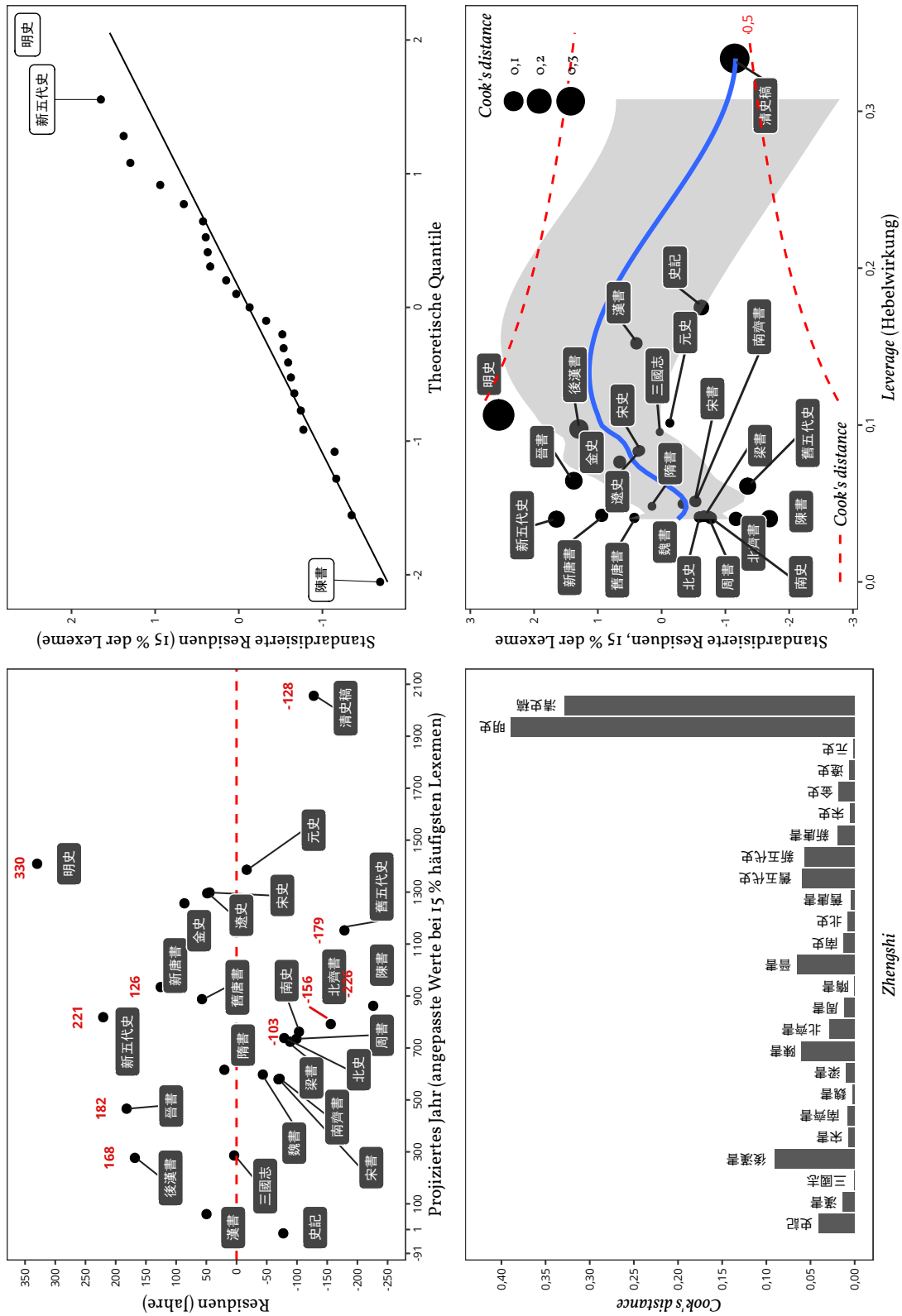


Abbildung 6.34 Diagnoseplots (Residuen, *Normal Q-Q*, *Cook´s distance*, *Leverage*)

Tabelle 6.15 Datierung unterschiedlicher Texte mit *AYL*, 15 % der häufigsten 2–4 Zeichen-Lexeme

Text	*AYL*-Datierung	tatsächliche Datierung	Δ_{min}
Dao de jing 道德經	-763	ca. –500—-400	-263
Zhuangzi 莊子	-116	ca. -400—200	+84
Zhongjing 忠經	81	ca. 320–960	-239
Wen xuan 文選	14	520–530 [ca. -300–500]	-
Zi zhi tong jian 資治通鑑	627	1084	-460
Meng xi bi tan 夢溪筆談	1282	1088	+194
Sanguo zhi yanyi 三國志演義	1333	1300–1400	-
Shui hu zhuan 水滸傳	2626	ca. 1320–1372	+1.254
Jin ping mei 金瓶梅	2785	ca. 1596–1610	+1.175
Hong lou meng 紅樓夢	2854	ca. 1750	ca. +1.104
Ru lin wai shi 儒林外史	2609	1750	+940
Xi you ji 西遊記	2266	1592	+674

Ouyang Xiu 歐陽修 (1007–1072) als konzisere Neufassung der *Jiu Wudai shi* 舊五代史 (damals 五代史) verfasst.[263] Ouyang Xiu, der auch einer der Verfasser des im Modell ebenfalls als „zu alt“ geschätzten *Xin Tang shu* 新唐書 ist,[264] orientierte sich dabei stilistisch stark an antiken Vorbildern.[265]

Auf der anderen Seite lassen sich im Modell stark *zu neu* datierte Texte nicht mit derselben Kohärenz erklären. *Qing shi gao* und *Jiu Wudai shi* sind sehr zeitnah nach Ende des beschriebenen Zeitraums entstanden, bei dem 226 Jahre zu neu datierten *Chen shu* liegen aber fast 50 Jahre zwischen dem Ende der Chen-Dynastie (557–589) und der Fertigstellung im Jahr 636, sehr ähnlich verhält es sich bei dem 156 Jahre zu neu datierten *Bei Qi shu* 北齊書. Beide Texte sind überdies deutlich kürzer und enthalten dadurch deutlich weniger *types* als der Durchschnitt des Korpus. Grundsätzlich sind die meisten Abweichungen normal verteilt (Abb. 6.34, oben rechts). *Shiji* und *Qing shi gao* haben als jeweils ältester und neuester Text den größten Einfluss auf Steigung und Achsenabschnitt der Regressionsgeraden, wohingegen die größten Residuen durch ihre Lage im mittleren Beobachtungsbereich verhältnismäßig wenig Einfluss auf das Modell nehmen können (Abb. 6.34, unten). Der Cook'sche Abstand aller Beobachtungen liegt unter 0,5, so dass keine zu starke Hebelwirkung auf das Modell besteht.

Auch wenn sich für die Abweichungen im gegebenen Regressionsmodell stimmige Erklärungen finden lassen, muss vermutet werden, dass die starke Korrelation zwischen *AYL* und Entstehung der zu datierenden Texte nur durch den ebenfalls stark temporal konnotierten Inhalt der *zhengshi* möglich wird. Zudem muss eine Überanpassung der berechneten Modelle auf das gegebene Textkorpus geprüft und infrage gestellt werden, ob das *AYL* überhaupt für die Datierung anderer Text(gattungen) infrage kommt. Aufgrund der geringen Größe des Korpus kann auch innerhalb der *zhengshi* keine sinnvolle Aufteilung in Test- und Trainingsdaten gemacht werden, die Grundvoraussetzung für eine solide Evaluierung der *AYL*-Projektion als Datierungsmethode wäre.

263 Siehe Davis 2004, S. xlvii.

264 Im Ggstz. zur *XWDS* war die Kompilation des *Xin Tang shu* allerdings keine private Unternehmung, sondern wurde gemeinsam mit Song Qi 宋祁 (998–1061) und weiteren Gelehrten auf Geheiß des Hofes kompiliert. Siehe auch Wilkinson 2000, S. 820.

265 Ouyang Xiu wird dabei eine „literary revolution to replace the awkward ‚contemporary‘ prose current in his day with the ‚classical‘ style of the Spring and Autumn period“ nachgesagt. Siehe Davis 2004, S. xliv.

Versucht man mit der ermittelten Funktion einige eklektisch ausgewählte Texte unterschiedlicher Genres zu datieren (Tabelle 6.15), ergibt sich ein sehr durchwachsenes Bild. Für klassische bzw. schriftsprachliche Texte lassen sich halbwegs sinnvolle Ergebnisse erzielen, die zutreffend sind oder im erwartbaren Rahmen abweichen. So deutet sich ein klarer, richtiger Unterschied zwischen den beiden daoistischen Klassikern *Zhuangzi* 莊子 und dem deutlich älteren *Dao de jing* 道德經 an. Das *Zhongjing* stammt zwar nicht – wie mittels *AYL*-Datierung eingeordnet – aus der Han-Zeit, doch entspricht dies der traditionellen Einordnung.[266] Das Anfang des 6. Jh. zusammengestellte *Wenxuan* 文選 wird zwar auf das Jahr 14 datiert, enthält jedoch Texte aus der Zeit von ca. 300 v. u. Z. bis zum Ende des 5. Jh. Das *Zi zhi tong jian* 資治通鑑 sollte als historiographischer Text besonders gut zu den verwendeten Trainingsdaten des *zhengshi*-Korpus passen, jedoch wird eine um mehr als 400 Jahre zu frühe Datierung projiziert. Eine Ursache dafür ist sicherlich, dass zwei zeitliche Aspekte hier untrennbar ineinander greifen. Zwar wurde der Text im 11. Jh. verfasst, das Vokabular ist aber geprägt durch den langen Zeitraum, über den geschrieben wird: vom 4. Jh. v. u. Z. bis zur zweiten Hälfte des 10. Jh. Wie die Beispiele der *xiaoshuo* 小說 aus dem 14.–18. Jh. zeigen (*Hong lou meng* 紅樓夢, *Shui hu zhuan* 水滸傳, *Jin ping mei* 金瓶梅, *Ru lin wai shi* 儒林外史 und *Xi you ji* 西遊記), werden umgangssprachlichere Texte um viele Jahrhunderte zu neu datiert. Lediglich die Datierung des Romans *Sanguo zhi yanyi* 三國志演義 wirkt zutreffend.[267] Auch hier wirkt sicherlich – diesmal positiv – ein Effekt der Zeit, über die geschrieben wird – das Ende der Han-Zeit und die Zeit der Drei Reiche (ca. 208–280) – auf das Datierungsergebnis ein.

Die Erweiterung des für die *AYL*-Funktion zugrunde gelegten Textkorpus um weite Teile des Loewe-Korpus[268] kann genutzt werden, um zu zeigen, dass für klassische und schriftsprachliche Texte auch dann ein guter Zusammenhang zwischen *AYL* und Textentstehung hergestellt werden kann, wenn keine Beschränkung auf historiographische Texte besteht. Abb. 6.35 zeigt die zugehörige Regressionsgerade.

Da das Loewe-Korpus zahlreiche deutlich kürzere Texte enthält, konnte hier die beste Korrelation mit $R = 0,93$ bei Betrachtung von 90 % der häufigsten 2–4-Zeichen Lexeme hergestellt werden.[269] Trotzdessen fällt der Standardfehler mit 221 [Jahren] deutlich höher aus als bei einem auf *zhengshi* spezialisierten Modell.

6.3.2 *AYL* mit unterschiedlichen Korpora

Die Erkenntnisse aus 6.3.1 suggerieren, dass die besten Ergebnisse erzielt werden können, wenn für unterschiedliche Textgattungen jeweils eigene Trainingsdaten verwendet werden. Um dies fundiert zu evaluieren, sind verschiedene größere, diachrone Textkorpora erforderlich, z. B. die Abteilungen des *Xu xiu si ku quan shu* 續修四庫全書.[270] Diese eignen sich – wie festgestellt werden musste – als Korpus für die Evaluation von Datierungsmethoden allerdings nur bedingt, da zahlreiche spätere Ausgaben antiker Texte enthalten sind.[271]

266 Die Datierung des *Zhongjing* 忠經 wird in Abschnitt 6.2.4, ab S. 6.2.4, diskutiert. Der Text wird traditionell einem Ma Rong 馬融 (79–166) zugeschrieben, dies gilt aber als widerlegt.

267 Hier wurden die sechs „klassische chinesische Romane" nach der Definition in Hsia Chih-tsing 夏志清 1968, ausgewählt. Zur Analyse wurden hier die *Project Gutenberg*-Versionen verwendet.

268 Das Loewe-Korpus enthält digitale Fassungen der Texte, die in der von Michael Loewe herausgegebenen Bibliographie *Early Chinese Texts* vorgestellt werden. Siehe Kapitel 4.2, siehe auch Loewe 1993; siehe auch T. Schalmey 2009, S. 104–106.

269 Die Gleichung der Regressionsgeraden lautet: $Y_{proj} = 2,57 \times AYL + 78,6$.

270 *XXSKQS*.

271 Siehe u. a. Kapitel 6.1.3, S. 174.

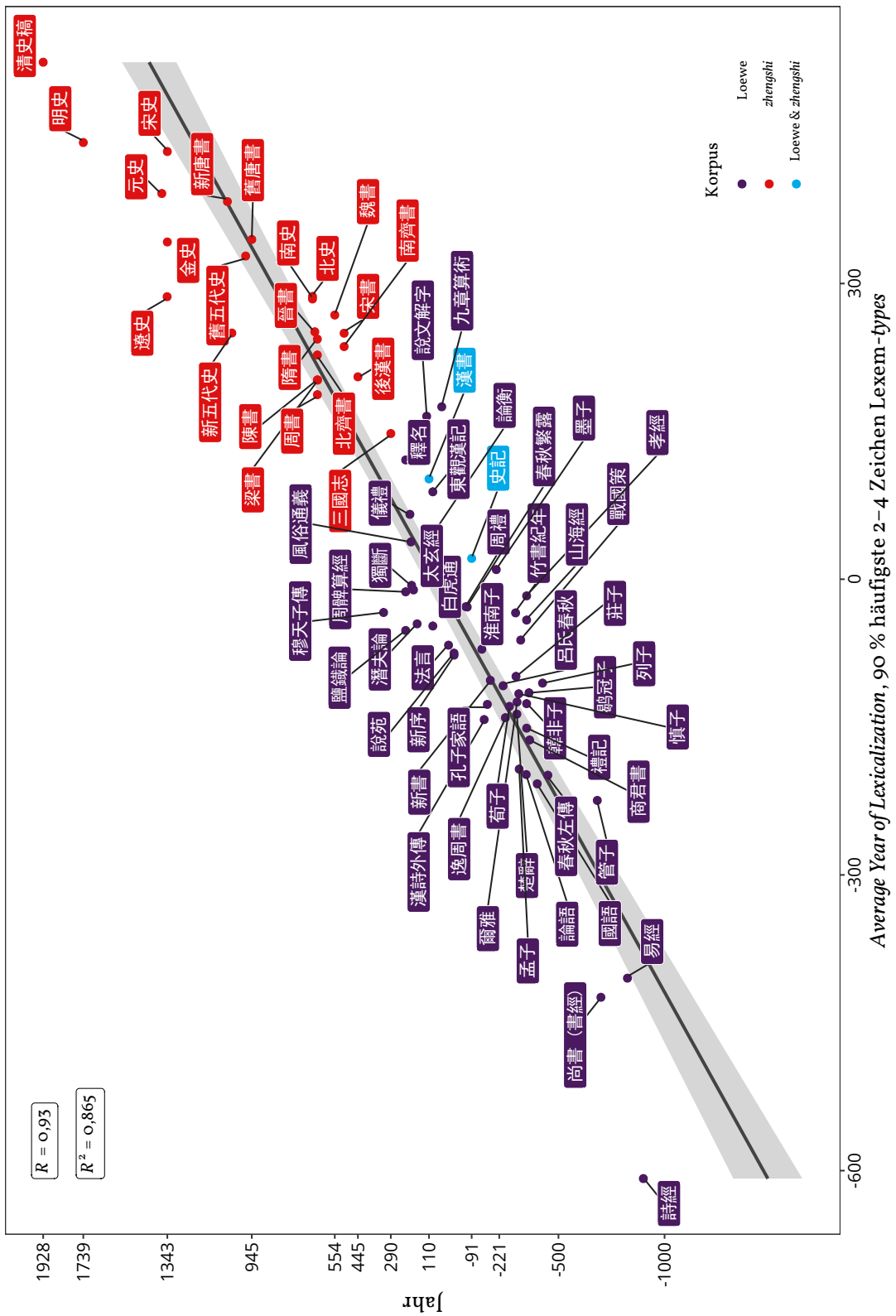


Abbildung 6.35 Korrelation Veröffentlichung Loewe / *zhengshi*, AYL mit 90 % 2–4 Zeichen Lexem-*types*

Obwohl – wie auch schon bei den *zhengshi* – die Problematik der Vermischung der Aspekte Entstehungszeit und Textinhalt besteht, lohnt sich aber – in Ermangelung geeigneterer Korpora – ein Blick auf die Verwendung des *AYL* im Kontext des *Difangzhi* 地方誌-Korpus[272] (*DFZ*).

Abb. 6.36b stellt ein wie in 6.3.1 optimiertes Regressionsmodell für 432 aus diesem Korpus zufällig ausgewählte Texte dar.[273] Da nur max. 3-Gramm-Daten zur Verfügung stehen, können nur Lexeme mit 2–3 Zeichen Länge betrachtet werden. Die stärkste Korrelation ergibt sich dabei mit ca. 70–90 % der häufigsten Lexem-*types* der Texte.[274] Im Gegensatz zu den *zhengshi* sind auch deutlich kürzere Texte mit teils nur wenigen hundert *types* enthalten. Bei Gegenüberstellung mit einem *zhengshi*-Modell (Abb. 6.36a) mit identischen Parametern, ergibt sich für die *DFZ* eine deutlich flachere Regressionsgerade mit höherem Achsenabschnitt. Als Referenz ist der relevante Abschnitt der Regressionsgeraden aus Abb. 6.36a ohne Datenpunkte in Abb. 6.36b erneut eingezeichnet.
Die gestrichelten grünen Linien geben in beiden Graphiken einen Toleranzbereich von 100 Jahren an. Datenpunkte innerhalb dieses Bereichs würden „richtig" datiert, wenn eine Genauigkeit von ±100 Jahren angestrebt wird.

Während für die *zhengshi* das lineare Modell mit sehr wenigen Datenpunkten und einem sehr langen Betrachtungszeitraum von 2.019 Jahren weiter sehr gut zu funktionieren scheint, ist in der rechten Graphik eine lineare Tendenz bei deutlich breiterer Streuung erkennbar. Bei einer Gruppe von *zhengshi*-Texten, die alle im Zeitraum zwischen 635 u. 659 herausgegeben wurden, ist allerdings eine ähnliche Streuung zu beobachten: *Liang shu* 梁書, *Chen shu* 陳書, *Bei Qi shu* 北齊書, *Zhou shu* 周書, *Sui shu* 隋書, *Jin shu* 晉書, *Nan shi* 南史 und *Bei shi* 北史 wurden in einem relativ kurzen Zeitraum von 25 Jahren während der Tang 唐 veröffentlicht. Die dem Modell angepassten Werte für die *AYL*-Datenpunkte in Abb. 6.36a für diese Texte ergeben einen Bereich von über 380 Jahren (501–883). Der Standardfehler der Regression des *zhengshi*-Modells beträgt dabei 171 [Jahre], für die *DFZ* ist er mit 95 trotz der breiten Streuung insgesamt deutlich niedriger. Die größten Abweichungen sind bei beiden Korpora sehr ähnlich: Die Geschichte der Liao (*Liao shi* 遼史) aus dem *zhengshi*-Korpus würde 364 Jahre zu früh auf das Jahr 980 datiert, die laut *DFZ*-Metadaten 1341 veröffentlichte Chronik der Präfektur Kunshan (*Kunshan qun zhi* 崑山郡志) 377 Jahre zu spät auf das Jahr 1718.

Bei mit dem *DFZ*-Korpus berechneten Modellen ist die Korrelation zwischen Veröffentlichung der Texte und *AYL* deutlich geringer, als dies bei den *zhengshi* der Fall war. Hierin spiegelt sich nicht nur die angesprochene Streuung mit deutlich mehr Texten aus einem deutlich kürzeren Betrachtungszeitraum wider, sondern auch die quasi nicht vorhandene Verortung der Texte im *DHYDCD* und eine hohe stilistische Rigidität der Textgattung.

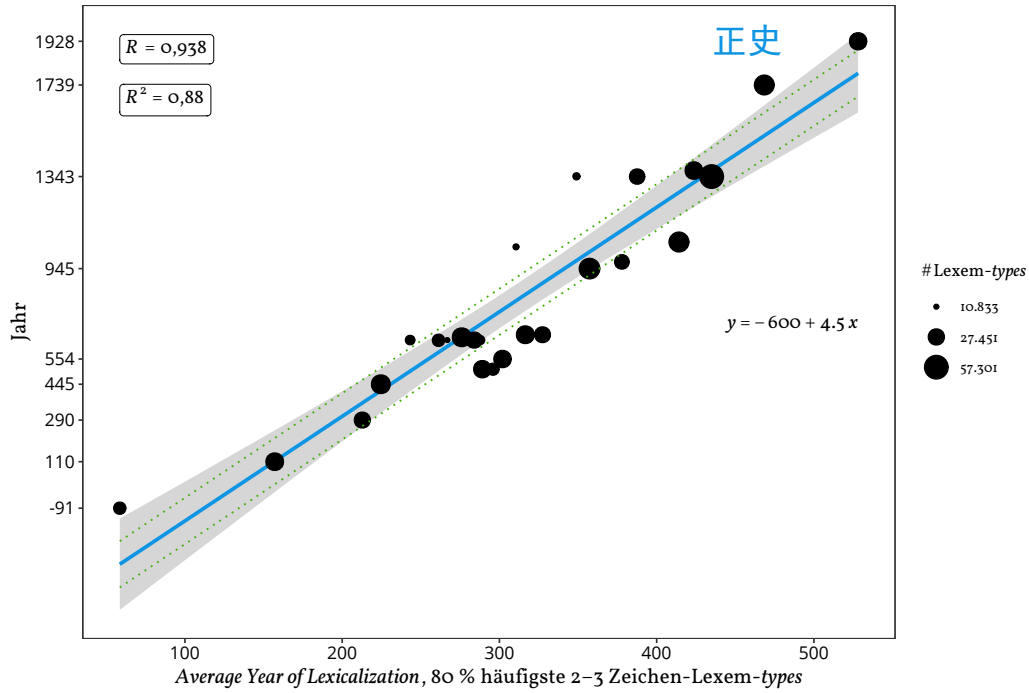
Die Größe des Korpus erlaubt es, das *AYL* als Datierungsmethode mit separaten Trainings- und Testdaten einem Praxistest zu unterziehen, der zumindest ansatzweise einen Vergleich mit den in den Kapiteln 6.1 und 6.2 vorgestellten Methoden ermöglicht.

Datiert man anhand der mit 432 *DFZ*-Texten trainierten Funktion aus Abb. 6.36b ($y = 1062{,}9 + 1{,}6 \times AYL$) die 216 Testdatensätze aus Kapitel 6.1.1, werden mit einem *MAE* von 94 Jahren bei einer

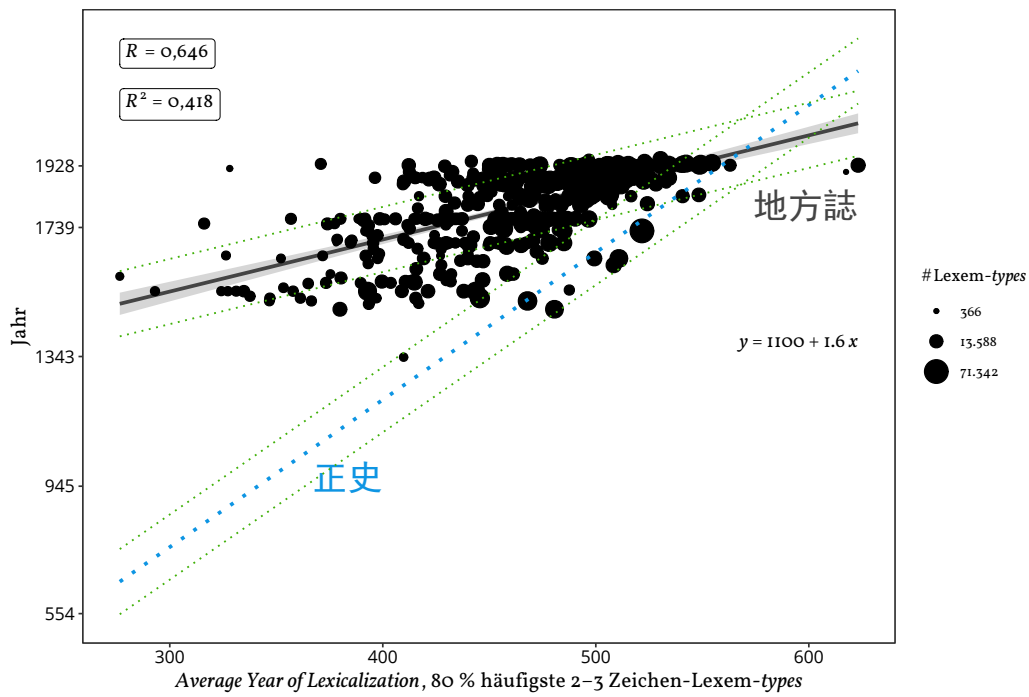
272 *DFZ*.

273 Um die Vergleichbarkeit der Ergebnisdaten zu erhöhen, wird der Testdatensatz aus Kapitel 6.1.1 mit 216 Texten ausgeschlossen.

274 Dass bei Verwendung von 15 % der häufigsten *types* ein schwaches *R* von nur 0,35 erreicht wird, ist sicherlich auf die deutlich geringere Länge einiger Texte zurückzuführen. Werden nur die Beobachtungen zu Texten mit min. 20.000 *types* berücksichtigt, erhöht sich der Korrelationskoeffizient *R* auf den – immer noch schwachen – Wert von 0,583.



(a) *zhengshi* 正史



(b) 432 *DFZ* 地方誌

Abbildung 6.36 Vergleich linearer Modelle, 80 % häufigste 2–3-Zeichen-Lexeme

Toleranz von ±100 Jahren 60,9 % der Texte korrekt datiert (Abb. 6.37), bei einer Toleranz von genau einem Jahrhundert (±50 Jahren) sind es immerhin noch 31 %.

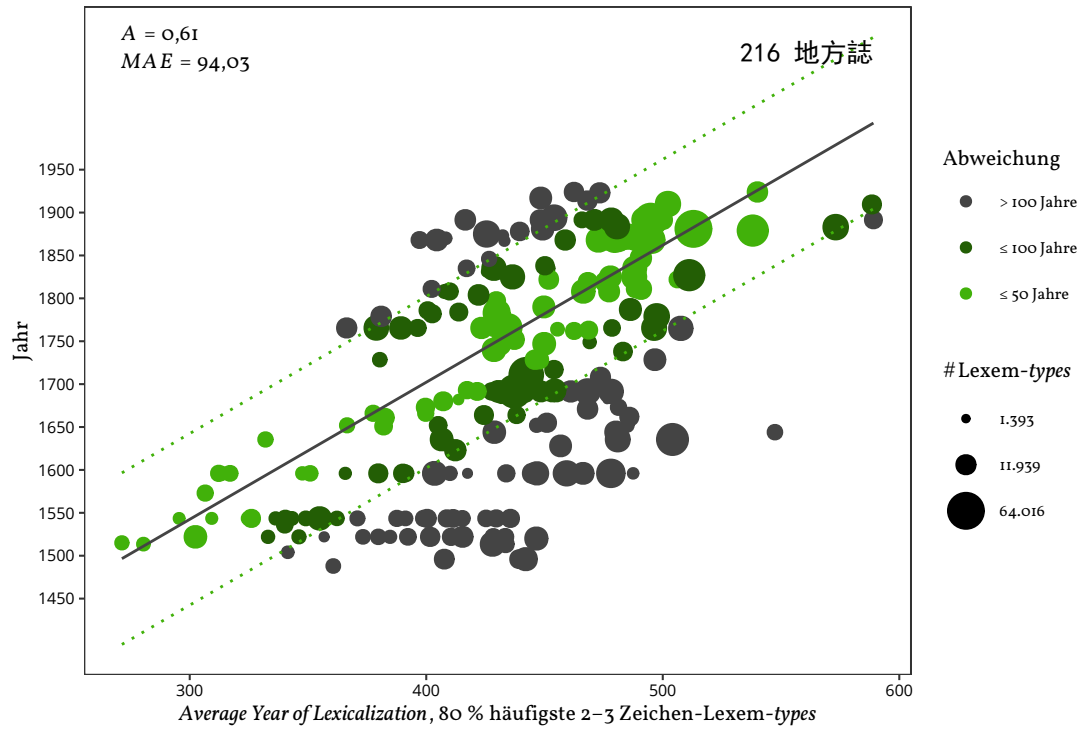


Abbildung 6.37 Datierungsergebnis *Difangzhi* mit 80 % häufigsten 2–3-Zeichen-Lexemen

Ordnet man dieselben *DFZ*-Texte unter Verwendung der Funktion aus dem *zhengshi*-Vergleichsmodell ein, würden bei einer durchschnittlichen Abweichung (*MAE*) von 360 Jahren nur 7,9 % der Texte bei einer Toleranz von ±100 Jahren korrekt zugeordnet.[275]

Für die Datierung von Texten anderer Genres bzw. außerhalb des Trainingskorpus ist ein lineares *AYL*-Modell also ungeeignet. Das gilt umso mehr, wenn mit schriftsprachlichen Trainingsdaten literarische bzw. v. a. umgangssprachliche Texte datiert werden. Das *AYL* kann ohne passende Trainingsdaten für das Chinesische also eher als eine Art relativer Zeit-Stil-Indikator eingesetzt werden. Die Verwendung mit dem Zweck der absoluten Textdatierung scheint nur sinnvoll, wenn Texte desselben Genres untereinander verglichen werden bzw. diachrone Trainingsdaten für genau diesen Texttypus vorliegen. Die Datierung mithilfe statistischer Sprachmodelle ist dann aber als etwas aussichtsreicher anzusehen.

275 Eine identische *Accuracy* bei einem *MAE* von 319 Jahren ergibt sich bei Anwendung der *zhengshi*-Funktion für die 432 Texte in Abb. 6.36b. Die korrekt zugeordneten Texte sind dabei zumeist verhältnismäßig lang, mit einer großen Anzahl an *types*. Da der Ausschluss kurzer Texte aber keinen wesentlichen Einfluss auf der Berechnung des *DFZ*-Modells hat, lassen sich daraus keine relevanten Schlüsse ziehen.

6.4 Untersuchte Datierungsmethoden im Überblick

In diesem Abschnitt werden die wesentlichen Unterschiede der in den Kapiteln 6.1, 6.2 und 6.3 untersuchten Methoden zur Textdatierung im Hinblick auf Potenzial und Limitationen bei der Betrachtung schriftsprachlicher chinesischer Texte zusammenfassend verglichen. Einige Ergebnisse werden – soweit möglich – anhand von *Accuracy* und *MAE* gegenübergestellt.

Bei der Datierung mithilfe statistischer Sprachmodelle (*Statistical Language Models*, *SLM*s) können Texte anhand unterschiedlicher Ähnlichkeitsmaße mit anderen Texten, oder mit aggregierten *chronons* verglichen werden. Beides funktioniert für schriftsprachliches Chinesisch grundsätzlich gut, wie in unterschiedlichen Experimenten gezeigt werden konnte. Von mehreren untersuchten Ähnlichkeitsmaßen hat sich – neben der *KLD* – die *NLLR* als am besten geeignet erwiesen. Zur Behandlung von *unseen events* hat sich ein einfaches *Smoothing* als am effizientesten herausgestellt. Dabei wird angenommen, dass ein *unseen event* weniger häufig ist als das seltenste *type* desjenigen *chronons* mit den meisten *types*.[276]

Zur Verwendung von *SLM*s sind für den gesamten Zeitraum, aus dem Texte datiert werden sollen, umfangreiche Trainingsdaten erforderlich. Die besten Ergebnisse werden dann erzielt, wenn die Trainings- und Testdaten aus demselben, stilistisch homogenen Korpus stammen, wie das bei den *difangzhi* 地方誌 (*DFZ*) der Fall ist. Da bis dato kaum entsprechende diachrone, schriftsprachliche Korpora digital vorliegen, ist die Nutzung von *SLM*s in der Praxis der linguistischen Datierung für schriftsprachliches Chinesisch zunächst nur eingeschränkt möglich. Eine – wenngleich grobe – Datierung mit genreübergreifenden Trainingsdaten ist jedoch möglich. Hierfür konnten die Belegstellen aus dem *DHYDCD* als Trainingsdaten verwendet werden.

Die in dieser Arbeit vorgestellte Datierung anhand von temporalen Textprofilen stützt sich im Wesentlichen auf diachrone Lexikalisierungsdaten, die in Kapitel 5.5[277] ebenfalls aus dem *DHYDCD* extrahiert wurden. Damit werden – grob vereinfacht – Lexeme, d. h. lexikalisierte 2–4-Zeichen-Kombinationen mit dem Jahr in Verbindung gebracht, in welchem ihr *Locus classicus* veröffentlicht wurde. Ein temporales Textprofil zeigt die Zuordnung aller in einem Text enthaltenen Lexeme zum Jahrhundert ihrer frühesten Verwendung als Balkendiagramm. Zusätzlich können Personen- und Ortsnamen, sowie temporale Ausdrücke in die Darstellung miteinbezogen werden. Um die zugrunde liegenden Daten für eine automatisierte zeitliche Einordnung des Textes zu nutzen, sind ebenfalls Trainingsdaten erforderlich. Die Abhängigkeit von einzelnen Genres oder Zeiträumen ist dabei aber deutlich geringer. Im Gegensatz zu den übrigen untersuchten Methoden erfordert die Interpretation der Textprofile nur geringe mathematische bzw. statistische Vorkenntnisse. Durch die transparente Darstellung ist die Nutzung der Profile für die Suche nach Hinweisen für die Textdatierung auch dann noch möglich, wenn ein statistischer Ansatz wenig aussichtsreich ist. Dies kann vor allem dann hilfreich sein, wenn der zu datierende Text sehr kurz ist, z. B. für Gedichte, oder bewusst in einem klassischen Stil verfasst wurde. Dieser lexikographische Ansatz spiegelt dabei eine traditionelle Methodik der historischen Linguistik wider – das Aufspüren von Anachronismen.

In Kapitel 6.3 wurde zuletzt damit experimentiert, die Lexikalisierungsdaten auf eine einzelne Messgröße, das durchschnittliche Jahr der Lexikalisierung der in einem Text enthaltenen Lexeme (*Average Year of Lexicalization*, *AYL*), zu reduzieren. Als Datengrundlage dienen auch hier die diachronen Lexikalisierungsdaten aus dem *DHYDCD*. Es konnte gezeigt werden, dass bei

276 Je nach Art und Umfang der Trainingsdaten sollte der *Smoothing*-Parameter λ zur Bestimmung der optimierten Häufigkeit des *unseen event* angepasst werden. Siehe dazu die Abschnitte 6.1.1, ab S. 164, sowie 6.1.3, ab S. 171.

277 Siehe ab S. 120.

Betrachtung einer homogenen, diachronen Textreihe ein linearer Zusammenhang zwischen der Entstehung der Texte und dem *AYL* besteht. Für seine Berechnung muss ein zu optimierender Anteil der häufigsten, im Text enthaltenen 2–4-Zeichen-Lexeme betrachtet werden. Eine starke Abhängigkeit besteht dabei erneut zu den stilistischen Eigenschaften der zu datierenden Texte. Für eine praktische Nutzung ist das Vorhandensein passender Trainingsdaten daher zwingend erforderlich. Die Anzahl der dafür erforderlichen Texte sollte – anders als bei der Verwendung von *SLM*s – allerdings überschaubar sein. Mit einer Linearregression kann anhand dieser Trainingsdaten eine Funktion berechnet werden, die das *AYL* auf die geschätzte Entstehung des Textes projiziert. Eine stabile Korrelation zwischen Textentstehung und *AYL* kann aber nur für längere Texte ab etwa 10.000 Zeichen (ca. 10 Seiten)[278] erreicht werden. Sind Genre und grobe Entstehungszeit eines zu datierenden Textes unbekannt oder stehen keine ausreichenden Trainingsdaten zur Verfügung, kann das *AYL* lediglich als grober Zeit-Stil-Indikator im direkten Vergleich zwischen ähnlichen Texten betrachtet werden.

Bei allen oben erläuterten Methoden wird von der Datierung von *Plain Text* bzw. *n*-Gramm Häufigkeitslisten ausgegangen. In allen Fällen sollte geprüft werden, ob die verwendeten Ausgaben Kommentare enthalten – auch wenn diese stilistisch nicht unbedingt moderner sind als der zu datierende Haupttext, können sie einerseits einen bedeutend größeren Anteil einnehmen als der eigentliche Text, andererseits suchen die Kommentator:innen oft den Text mit zeitgenössische(re)n Ausdrücken zu erklären, was grundsätzlich äußerst problematisch für den Versuch einer rein linguistischen bzw. lexikographischen Datierung ist.[279] Für eine philologische Untersuchung hingegen kann die Kommentartradition durchaus zur Beleuchtung datierungsrelevanter Aspekte beitragen, da sie Aufschluss über die Überlieferungsgeschichte geben kann. Auch Kommentare können dabei natürlich – wie ein Text selbst – eine spätere Fälschung sein.

Dass alle beschriebenen Datierungsmethoden anhand des *DFZ* Korpus getestet wurden, erlaubt einen groben Vergleich der Ergebnisse. In Tabelle 6.16 sind einige der in den Kapiteln 6.1, 6.2 und 6.3 durchgeführten Experimente gegenübergestellt. Dafür werden jeweils Ergebnisse mit der besten erzielten *Accuracy* (A_{max}) ausgewählt und der zugehörige *mean average error* (*MAE*) angegeben.
Auf den ersten Blick sind die Ergebnisse, die sich mit *SLM*- und lexikographischer Datierung erzielen lassen, sehr ähnlich. In beiden Fällen kann eine *Accuracy* von ca. 50 % erreicht und die Texte mit einem *MAE* von ca. 50–80 Jahren datiert werden. Während *SLM*s dafür auch Einzelzeichen und Worthäufigkeiten betrachten, werden bei der Erstellung von temporalen Profilen nur Lexeme mit einer Mindestlänge von 2 Zeichen berücksichtigt, unabhängig von ihrer Häufigkeit im untersuchten Text.

Eine hohe *Accuracy* konnte mit *SLM*s nur dann erreicht werden, wenn spezifische Trainingsdaten zur Erzeugung der Modelle verwendet wurden. Auch der Algorithmus für die Datierung mittels temporalen Textprofilen basiert zwar auf Beobachtungen an Trainingsdaten aus demselben Korpus, jedoch ist die Datierung hier offen über den Zeitraum von 700 v. u. Z. bis ins 20. Jh. angelegt. Bei der Datierung mit *SLM*s ist der Ergebnisraum auf den Zeitraum der Trainingsdaten begrenzt, von 1475–1925. Bei 17 überlappenden 50-Jahre-*chronons* ergibt sich eine

278 Zum Vergleich werden Seiten im Format A4 mit einer Schriftgröße von 12 pt herangezogen.
279 Während in gedruckten bzw. formatierten Ausgaben oder strukturierten Dateiformaten Kommentare in der Regel z. B. durch eine kleinere Schriftgröße bzw. entsprechende Markierungen klar erkennbar sind, ist die Trennung von Kommentaren und Haupttext in *Plain Text*-Formaten erschwert.

Tabelle 6.16 Vergleich der in Kapitel 6 vorgestellten Methoden anhand des *DFZ*-Korpus

Methode	A_{max}	*MAE*	Details	*T*	Training
Statistical Language Models					
50 Jahre *Chronon-SLM*	60,6	40,3	◆ *NLLR*, 1–2 Zeichen-Lexeme & *temporal expressions*	1475–1925	spezifisch, *DFZ*
50 Jahre *Chronon-SLM*	64,4	42,5	◆ *NLLR* * *TE*, 1–2 Zeichen-Lexeme & *temporal expressions*	1475–1925	spezifisch, *DFZ*
50 Jahre *Dokumenten-SLM*	46,3	59,3	◆ *NLLR*, 1–2 Zeichen-Lexeme & *temporal expressions*	1475–1925	spezifisch, *DFZ*
100 Jahre *chronon-SLM*	19,9	136	◆ *CS* * *tf-idf*, 1–2 Gramme	ca. 700 v. u. Z.–2000	unspezifisch, *DHYDCD*
100 Jahre *chronon-SLM*	16,2	140,6	◆ *NLLR* * *TE*, 1–2 Zeichen-Lexeme & *temporal expressions*	ca. 700 v. u. Z.–2000	unspezifisch, *DHYDCD*
Lexikographisch / datenbankgestützt					
100 Jahre Neologismusprofil, korrigierte Gewichtung	47,2	83,9	■ 2–3 Zeichen-Lexeme	ca. 700 v. u. Z.–2000	spezifisch
100 Jahre Temporales Profil, korrigierte Gewichtung	62,5	72,1	■■ 2–3 Zeichen-Lexeme, 3 Z. Namen,	ca. 700 v. u. Z.–2000	spezifisch
100 Jahre Temporales Profil, korrigierte Gewichtung	75,9	63,2	■■■ 2–3 Zeichen-Lexeme, 3 Z. Namen, *temporal expr.*	ca. 700 v. u. Z.–2000	spezifisch
Newest dates in text	88	57,5	■ 4+ *temporal expressions*	ca. 220 v. u. Z.–1912	nein
***Average Year of Lexicalization* / datenbankgestützt**					
$AYL_{0,8}^{2-3}$, Toleranz ±50 Jahre	31	94	▲ 80 % 2–3 Zeichen Lexeme; kontinuierlich	∞/ ca. 100 v. u. Z.–4000	spezifisch, 432 *DFZ*

Baseline-Wahrscheinlichkeit von etwa 11 %, mit der ein Zufallsgenerator die Texte korrekt zuordnen würde, bei temporalen Textprofilen wären es – ohne Überlappungen – lediglich 3,7 %. Andererseits kann dabei – auch aufgrund der Ungenauigkeit der auf den *attestations* im *DHYDCD* aufgebauten diachronen Lexemdatenbank – bislang nur eine Genauigkeit von 100 Jahren angestrebt werden.

Wird der Ergebnisraum für die *SLM*-Datierung ebenfalls auf 700 v. u. Z. bis ins 20. Jh. erweitert, indem Sprachmodelle mit 100-Jahre-*chronons* aus den *DHYDCD attestations* erzeugt werden, kann bei einer *Baseline* von etwa 4 % noch eine *Accuracy* von knapp 20 % erreicht werden, mit einem immer noch beachtlichen *MAE* von 136 Jahren. Ein überwiegender Teil der *DFZ* kann also auch ohne spezifische Trainingsdaten grob korrekt eingeordnet werden. Insgesamt erweist sich der Einsatz temporaler Profile bei der Datierung von Texten mit unspezifischen Trainingsdaten aber als robuster.[280]

Einen Sonderfall stellt die Datierung der *DFZ* auf Basis der spätesten enthaltenen *temporal expressions* dar. Eine Datierung, die sich allein auf Zeitangaben im Text stützt, kann nur dann erfolgreich sein, wenn der zu datierende Text überhaupt konkrete Zeitangaben enthält, vor allem aber dürfen erzählte Zeit und Veröffentlichung nur unweit auseinanderliegen. Diese Bedingungen sind für die Lokalchroniken auch deshalb optimal erfüllt, da eine entsprechende Einschränkung bei der Auswahl von Test- und Trainingsdaten vorgenommen wurde.[281] Es muss davon ausgegangen werden, dass für kaum eine andere Textgattung auf diesem Weg vergleichbare Ergebnisse erzielt werden können. In Verbindung mit temporalen Textprofilen bietet die Analyse vorhandener Zeitausdrücke aber immer einen Mehrwert.

Die Ergebnisse der zeitlichen Einordnung mithilfe des *AYL* sind – trotz des spezifischen Trainings – eindeutig als schwächer zu bewerten. Das Potenzial dieser Herangehensweise liegt also eher in der relativen Einordnung von Texten einer diachronen Reihe, bzw. kann eine Datierungsfunktion auch dann trainiert bzw. optimiert werden, wenn Trainingsdaten nur in geringem Umfang zur Verfügung stehen.

Während die chronologische Einordnung schriftsprachlicher chinesischer Texte mit allen drei untersuchten Methoden grundsätzlich funktioniert, zeigen sich außerhalb der mit dem *DFZ*-Korpus durchgeführten Experimente immer wieder Limitationen einer linguistischen Datierung. Gerade statistische Methoden sind besonders anfällig für eine zu frühe Datierung von Texten, die in einem streng altertümlichen Stil (*guwen* 古文) verfasst sind, dessen Grammatik und Wortwahl für viele Textgattungen bis zum Ende der Kaiserzeit Vorbilcharakter genoss bzw. in Ausnahmefällen immer noch genießt.[282] Auch eine lexikographische Methodik kann nur dann Abhilfe schaffen, wenn der zu datierende Text überhaupt zeitgenössische Lexeme enthält – und die zugrundeliegende Datenbasis ausreichend umfangreich und genau ist.

6.4.1 *VisualTime — user interface* für Datierungsmethoden

Um die wesentlichen hier vorgestellten Methoden zur Datierung (schriftsprachlicher) chinesischer Texte für Sinolog:innen mit geringen computerlinguistischen Vorkenntnissen nutzbar zu machen, wird mit *VisualTime* ein *user interface* im Browser bereitgestellt.[283]

280 Siehe Abschnitt 6.2.5, ab S. 197, v. a. Tabelle 6.12, S. 207.

281 Siehe dazu Abschnitt 6.1.1, S. 158. Ziel dieser Einschränkung ist der Ausschluss von späteren Editionen bzw. die Vermeidung potenzieller Verzerrungen durch diese.

282 Vgl. z. B. Tai und M. K. M. Chan 1999, S. 229.

283 Eine aktuelle Version kann unter `https://visualtime.schalmey.de` getestet werden.

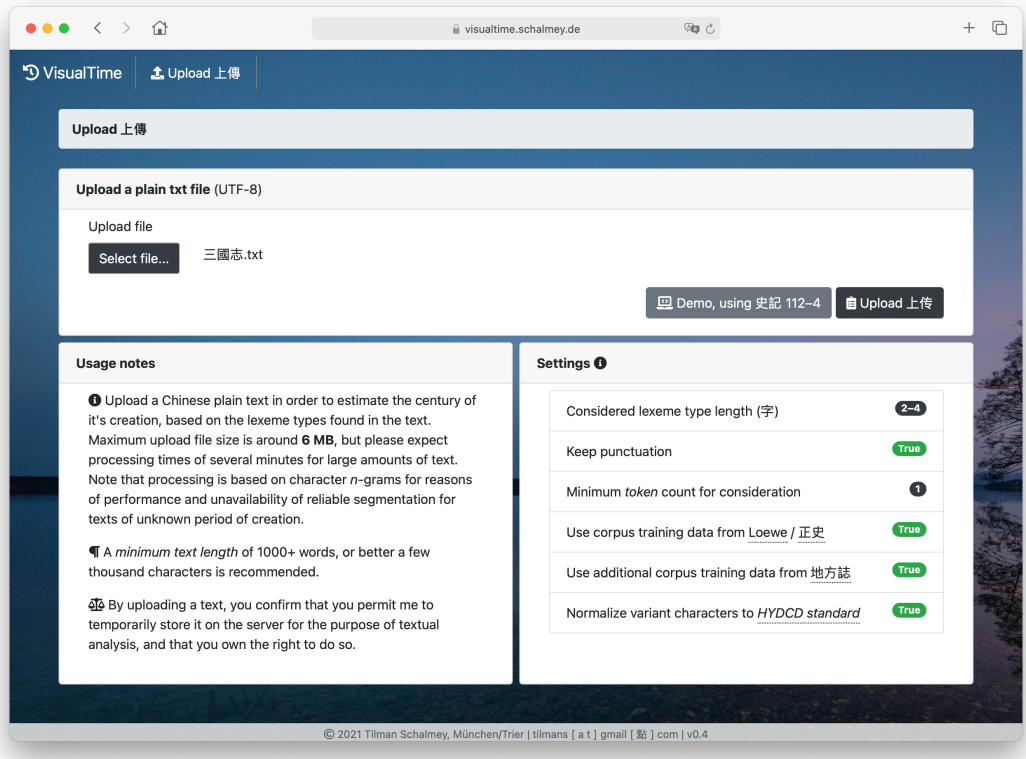


Abbildung 6.38 *VisualTime* Startseite – Datei auswählen und hochladen

Die für Kapitel 6.1, 6.2 und 6.3 entwickelten Komponenten und die dabei gewonnenen Erkenntnisse fließen dabei in eine Anwendung ein, der das *Python Framework Django*[284] zugrunde liegt. Die so erzeugte Benutzer:innenoberfläche ermöglicht den Upload einer *Plain Text* Datei, für deren Inhalt ein temporales Textprofil erzeugt wird. Zusätzlich erfolgt eine *SLM*-Klassifizierung sowie Berechnung des *AYL*. Das resultierende Neologismusprofil und die Ergebnisse der Datierung per Sprachmodell können im Detail erkundet und direkt miteinander verglichen werden.

VisualTime wird zunächst anhand eines Optimalbeispiels erläutert. Das *Sanguo zhi* 三國志 (*Chroniken der Drei Reiche*) ist in den *zhengshi* 正史 Trainingsdaten für die diachrone Lexemdatenbank und dem zur *AYL*-Projektion verwendeten Regressionsmodell enthalten. Zudem wird in 6.799 *DHYDCD*-Einträgen aus dem *Sanguo zhi* zitiert, so dass auch die *chronons* 200–300 und 250–350 des verwendeten temporalen Sprachmodells direkt von diesem Text geprägt sind. Durch diese „Schummelei“, die teilweise Identität von Trainings- und Testdaten, lässt sich die Ausgabe unter Optimalbedingungen veranschaulichen.

284 Django Software Foundation 2005–: *django – The web framework for perfectionists with deadlines*. URL: https://www.djangoproject.com/ (besucht am 12. 10. 2021), *Django* wird hier primär als *templating engine* zur Erzeugung der Seiten mit *HTML* & *CSS* verwendet, die Kernfunktionalität der Verwaltung von Datenmodellen wird hier (noch) nicht genutzt, stattdessen wird die in Kapitel 5.5 (ab S. 120) erzeugte Datenbank angebunden.

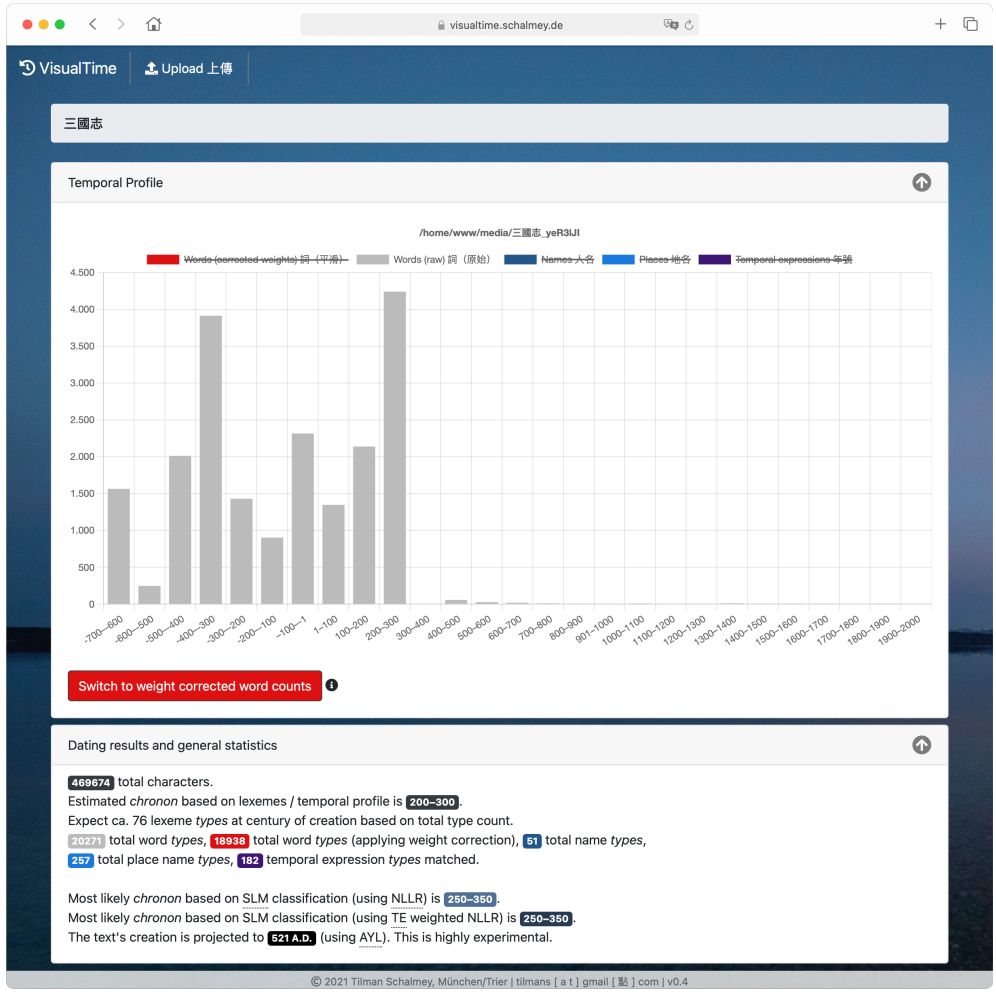


Abbildung 6.39 *VisualTime* Ergebnisseite

Allgemeine Statistiken und Temporales Textprofil

Auf der Ergebnisseite (Abb. 6.39) werden allgemeine Statistiken zum hochgeladenen Text und die Ergebnisse der unterschiedlichen Datierungsmethoden angezeigt. Durch die Verwendung von *Chart.js*,[285] einem *JavaScript Framework* zur Visualisierung von Daten im Browser bzw. innerhalb von Webseiten, kann ein interaktives temporales Profil des Textes mit an- und abklickbaren Lexemen, Namen, Ortsnamen und temporalen Ausdrücken dargestellt werden. Initial werden die im Text ausgemachten Lexeme jahrhundertweise in einem Balkendiagramm angezeigt, wahlweise kann zu einer Darstellung mit Gewichtungskorrektur gewechselt werden.[286] Da der betrachtete Text in den Trainingsdaten für die Lexemdatenbank enthalten ist, kann in beiden Darstellungen die Entstehung im 3. Jh. klar abgelesen werden – *types* aus späteren Jahrhunder-

285 Evert Timberg et al. 2013–: *Chart.js*. GitHub Repository. URL: `https://github.com/chartjs/Chart.js` (besucht am 12. 10. 2021).

286 In Abschnitt 6.2.1, ab S. 182, wird die Auswertung der dieser Darstellung zugrunde liegenden Daten ausführlich erläutert. Ab S. 185 wird auf das hier verwendete *Smoothing* eingegangen.

ten sind kaum enthalten. Entsprechend wählt auch der Datierungsalgorithmus mit 200–300 das korrekte *chronon*.[287]

Abbildung 6.40 *VisualTime* – Temporales Profil des *Sanguo zhi* mit flexibel aktivierbaren *types*

Durch Deaktivieren der Lexemanzeige (*word types*) bzw. Aktivierung der übrigen *types* können Namen oder temporale Ausdrücke separat betrachtet werden (Abb. 6.40). Die im *Sanguo zhi* erkannten *temporal expressions* (unten rechts) zeigen deutlich, dass die im Text beschriebene Zeit primär ins 2.–3. Jh. fällt, was ebenfalls den Tatsachen entspricht. Deutlich problematischer ist die zeitliche Einordnung der im Text erkannten Personen- und Ortsnamen. Es kommt hier zu unterschiedlichen Arten von *false positives*, da Zeichenfolgen fälschlich als Namen erkannt werden können, Personen gleichen Namens abweichende biographische Daten haben, bzw. die als Datenquelle verwendete *CBDB* nicht ausreicht, um eine klare Interpretation zu ermöglichen.[288]

Um die *NER*-Ergebnisse dennoch sinnvoll nutzbar zu machen, kann durch Klick auf entsprechende *Tabs* auf der Ergebnisseite die Liste der erkannten Personennamen ausgeklappt und chronologisch, oder alternativ nach Häufigkeit der Nennung sortiert werden (Abb. 6.41).

Abbildung 6.41 *VisualTime* – Erkannte Namen im *Sanguo zhi* (Ausschnitt)

287 Siehe dazu Abschnitt 6.2.5, ab S. 197.
288 Siehe dazu auch Kapitel 4.7, ab S. 97, sowie Abschnitt 6.2.2, ab S. 189.

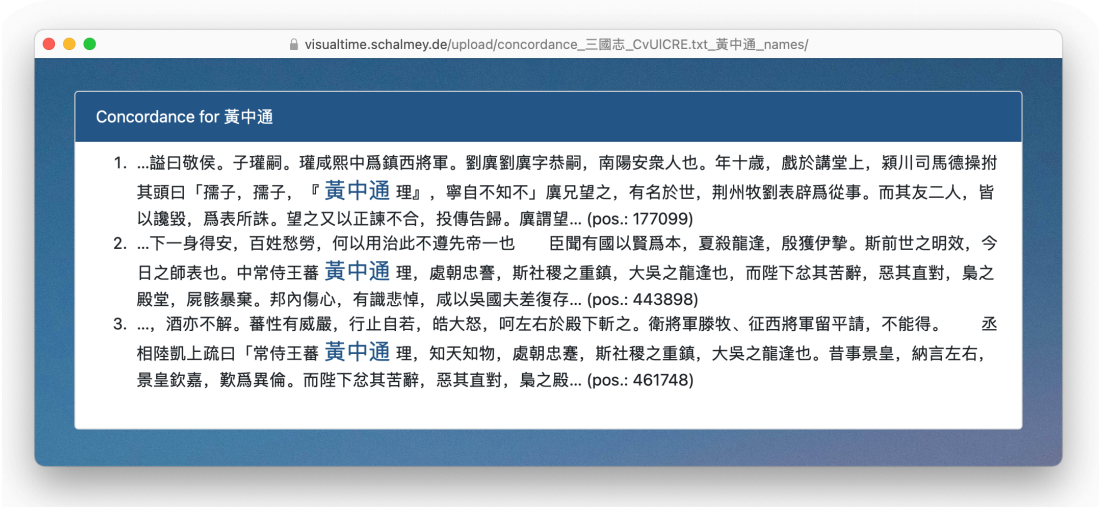


Abbildung 6.42 *VisualTime* – Anzeige aller Textstellen mit Huang Zhongtong 黃中通 im *Sanguo zhi*

Bei im Text häufigen Zeichenfolgen wie Zhuge Liang 諸葛亮 (181–234, 109 Nennungen) handelt sich mit höherer Wahrscheinlichkeit tatsächlich um die jeweilige Person. Zum Abgleich der Daten bzw. für die weitergehende Recherche ist zudem rechts die *CBDB* ID der angezeigten Einträge aufgeführt. Personen mit in der *CBDB* vollständigen biographischen Daten werden blau hervorgehoben.[289] Um zu verifizieren, ob im Text tatsächlich eine Person genannt wird und es sich nicht um Vorkommen einer zufällig mit dem Namen übereinstimmenden Zeichenkombination handelt, kann durch Klick auf den Namen eine Konkordanz der betroffenen Textstellen in einem neuen Fenster angezeigt werden (Abb 6.42).

Für Huang Zhongtong 黃中通 ist in der *CBDB* das sogenannte Indexjahr mit 1679 angegeben,[290] es muss sich hier also – trotz der drei Vorkommen – um *false positives* handeln. Alle drei Textstellen enthalten die Zeichenfolge *huang zhong tongli* 黃中通理, ein feststehender Ausdruck, der auf das *Yijing* 易經 (*Buch der Wandlungen*) zurückgeführt werden kann.[291]

Eine strukturierte Detailanzeige der im Text gefundenen *types* sowie von Vorkommen dieser *types* ist gleichermaßen auch für Lexeme und temporale Ausdrücke möglich.

289 Zu biographischen Angaben in der *CBDB* siehe auch Kapitel 4.7, S. 99.
290 Siehe *CBDB*, Zum Indexjahr in der *CBDB* siehe auch Kapitel 4.7, ab S. 99.
291 Die Farbe gelb in der Mitte steht für einen *junzi* 君子, der schnelle Auffassungsgabe und eine gute Urteilsfähigkeit aufweist. James Legge übersetzt „ 君子黃中通理“ aus dem Text zum Hexagramm *kun* 坤 mit „The superior man (emblemed here) by the yellow and correct (colour), is possessed of comprehension and discrimination.“ James Legge 1882: *The Yi King*. Übers. von James Legge. Sacred Books of the East XVI. Oxford: Clarendon Press, S. 421.

Datierung mit *SLM*

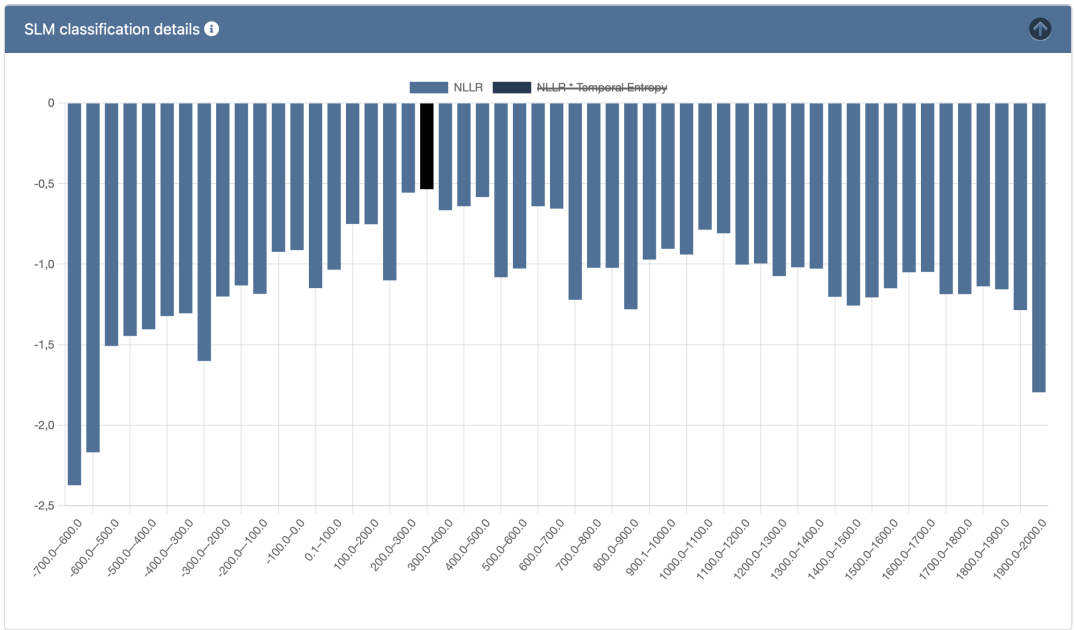


Abbildung 6.43 *VisualTime* – Anzeige *NLLR* des *Sanguo zhi* zu einzelnen *DHYDCD-chronons*

Wie in Abb. 6.39 zu sehen ist, wurde das *Sanguo zhi* mit einem temporalen *SLM* dem wahrscheinlichsten *chronon* 250–350 zugeordnet. Wie bereits angedeutet wird die Richtigkeit dieser Zuordnung stark dadurch begünstigt, dass Sätze aus dem *Sanguo zhi* 30 % der Trainingsdaten dieses *chronon*s ausmachen.[292] Um die Qualität bzw. Verlässlichkeit einer solchen Zuordnung zu analysieren, können die *NLLR*-Ähnlichkeiten des Texts zu allen *chronons* des Modells eingeblendet werden (Abb. 6.43).

Die Detailansicht zeigt eine tendenziell steigende *NLLR* von den frühesten *chronons* hin zur tatsächlichen Fertigstellung des Texts im Jahr 297. Die höchsten Werte erhalten dabei die beiden überlappenden *chronons* 200–300 und 250–350, letzteres in der Graphik farblich hervorgehoben. Für die späteren *chronons* ist wieder eine abnehmende Tendenz zu beobachten. Diese klar ablesbaren Tendenzen und die Nachbarschaft der ähnlichsten *chronons* sprechen für die Verlässlichkeit der Zuordnung.[293] In der Darstellung in Abb. 6.43 kann zu einem mit Temporaler Entropie gewichteten Modell umgeschaltet werden (ohne Abb.).[294]

Zur Veranschaulichung der Funktionsweise und Darstellungen sei zum Vergleich die Ausgabe in *VisualTime* für einen weiteren bekannten Text erläutert. Das *Sanguo zhi yanyi* 三國志演義 (*Die ausführliche und erläuterte Geschichte der Drei Reiche*) ist ein historischer Roman, der Luo Guanzhong

292 Die Trainingsdaten des *chronons* 250–350 setzen sich aus 22.452 Belegstellen aus dem *DHYDCD* zusammen, 6.799 davon sind dem *Sanguo zhi* entnommen. Siehe dazu auch Kapitel 5.6 (ab S. 137), sowie Abschnitt 6.1.3 (ab S. 171) zur Erzeugung und Verwendung des auf dem *DHYDCD* basierenden temporalen Sprachmodells.

293 Vgl. auch de Jong, Rode und Hiemstra 2005, S. 6: „A simple confidence measure for dating could be the relative distance between the score of the top-ranked time partition to the scores of the following ones. A more sophisticated measure could also take into account the level of timely scattering in the top-ranked partitions."

294 Siehe dazu Kapitel 3.3, S. 53.

羅貫中 (ca. 14. Jh.) zugeschrieben wird. Die älteste überlieferte Ausgabe stammt aus dem Jahr 1522.[295] Wie auch beim *Sanguo zhi* dienen die Geschehnisse zur Zeit der Drei Reiche als Grundlage, hier allerdings literarisch und mehr als 1.000 Jahre nach den historischen Ereignissen erzählt.

Anders als das *Sanguo zhi* ist das *Sanguo zhi yanyi* nicht Teil der Trainingsdaten. Zwar ist es im *DHYDCD* 1.644 mal als Beleg präsent, aber in der Datenbank fehlt eine Datierung des Texts, so dass weder das Sprachmodell noch die Lexemdatenbank davon beeinflusst sind.

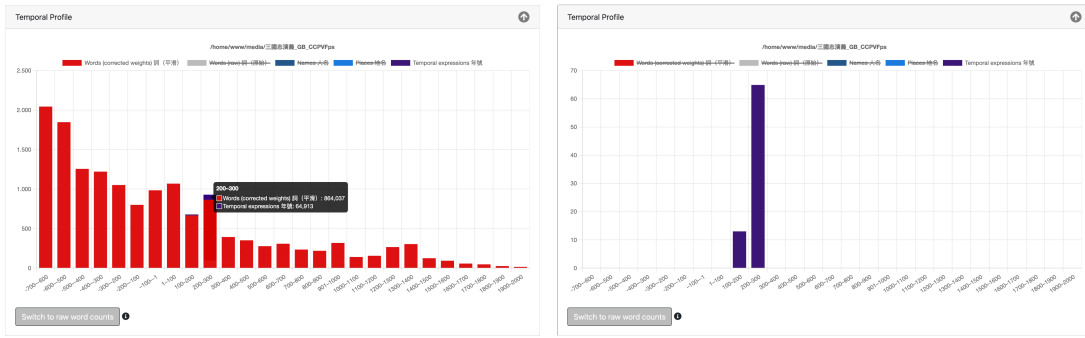


Abbildung 6.44 *VisualTime* – Profil des *Sanguo zhi yanyi*

Abb. 6.44 (links) zeigt das gewichtungskorrigierte Neologismusprofil des *Sanguo zhi yanyi*[296] mit eingeblendeten temporalen Ausdrücken. Eindeutig lässt sich nach Ausblenden der Lexem-*types* (Abb. 6.44 rechts) erkennen, welche Zeit in dem Text behandelt wird: das 2. und v. a. 3. Jh. Ganz anders als in der Darstellung des *Sanguo zhi* (Abb. 6.40) sind aber zahlreiche Lexeme enthalten, die erst nach dem 3. Jh. nachgewiesen sind. Ein deutliche Abnahme ist vom 14. hin zum 15. Jh. erkennbar, was für eine Entstehung des Texts im 14. oder Anfang des 15. Jhs. sprechen würde – zu Lebzeiten von Luo Guanzhong. Der Profil-Datierungsalgorithmus verortet den Roman in das *chronon* 1600–1700.[297]

Das *chronon* mit dem höchsten Wert für die *NLLR* ist 1550–1650, wobei hier ebenfalls die benachbarten *chronons*, v. a. 1600–1700, nahezu identische Werte erreichen (Abb. 6.45). Wie schon bei der Darstellung der *NLLR* für das deutlich ältere *Sanguo zhi* (Abb. 6.43) lässt sich ein zum Jahrhundert der Textentstehung steigender Wert erkennen, der anschließend (hier ab dem 18. Jh.) wieder abnimmt. Allerdings verläuft diese Zunahme ab dem 7. Jh. bereits sehr flach, was auf eine geringe Klarheit bzw. Verlässlichkeit der *SLM*-Datierung hindeutet.

295 Siehe Clemens Treter 2004: „Die Literatur der Ming- und Qing-Zeit". In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 225–287, S. 239.

296 Verwendete Version: Luo Guanzhong 羅貫中 2007 [?1522]: *Sanguo zhi yanyi* 三國志演義 *(Romance of the Three Kingdoms)*. Project Gutenberg eBook. URL: https://www.gutenberg.org/ebooks/23950 (besucht am 28. 05. 2021).

297 Ohne Abb. Die Datierung erfolgt hier aufgrund der Linearregression auf die für das Jahrhundert der Textentstehung erwartete Anzahl *types* (60) auf Basis der Gesamtzahl erkannter Lexem-*types* (16.021, mit korrigierter Gewichtung 15.127). Siehe dazu Abschnitt 6.2.5, ab S. 197.

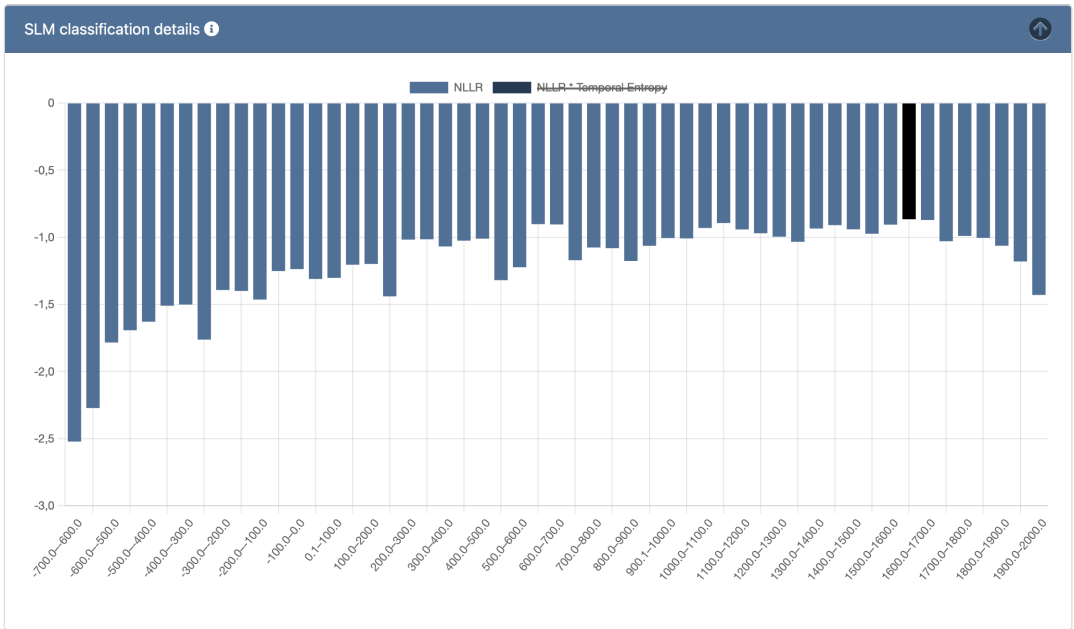


Abbildung 6.45 *VisualTime* – *NLLR*-Werte des *Sanguo zhi yanyi* für die einzelnen *chronons*

In diesem Fall lässt sich aus den genannten Interpretationen kein klares Bild ableiten. Alle drei genannten Datierungen können richtig sein, da weder klar ist, ob Luo der Autor war, wie stark die überlieferte Version von seiner Fassung abweicht und welche Version oder Ausgabe des *Sanguo zhi yanyi* hier genau vorliegt. Das Beispiel zeigt aber, dass sich sowohl mit dem *DHYDCD-SLM*, als auch mit der lexikographischen Datierungsmethode ähnliche Ergebnisse erzielen lassen, die in der Tendenz richtig sind.

Die Detaildarstellung der im Text erkannten *types* erlaubt es, weitere Analysen vorzunehmen. Die Betrachtung der temporalen Ausdrücke (Abb. 6.46) gibt Aufschluss über die Zeitangaben im Text – die späteste Angabe bezieht sich dabei auf das siebte Jahr der Ära Taikang (*Taikang qi nian* 太康七年, 286), während der Regierung von Kaiser Wu 武 der westlichen Jin 晉 (Sima Yan 司馬炎, reg. 266–290). Auffällig und erwartbar ist auch, dass der literarische Text trotz seinem etwas größereren Umfang (581.810 Zeichen) deutlich weniger unterschiedliche Zeitangaben – 78 – enthält als die historiographische Fassung mit 182.

Bei so umfangreichen Texten wenig aussichtsreich ist die Betrachtung der einzelnen Lexem-*types* (Abb. 6.47). Um z. B. Belege für eine Entstehung des Texts im 17. Jh. (oder später) zu suchen, müssen zumindest diejenigen *types* betrachtet werden, die dem 17. (52), 18. (35), 19. (6) und 20. Jh. (2) zugeordnet sind.

Abbildung 6.46 *VisualTime* – Temporale Ausdrücke im *Sanguo zhi yanyi* (Ausschnitt)

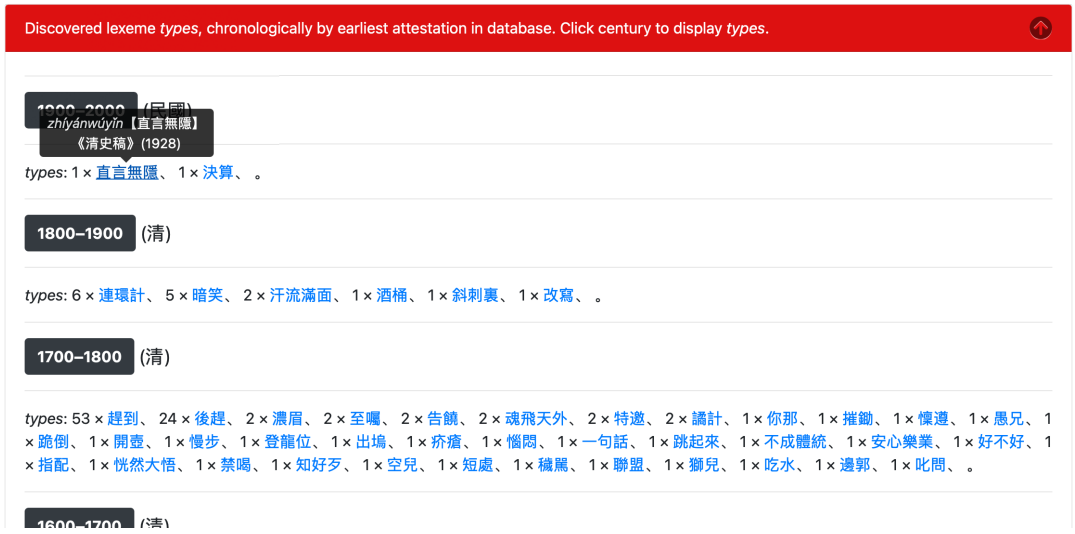


Abbildung 6.47 *VisualTime* – Lexeme im *Sanguo zhi yanyi* (Ausschnitt)

Der „neueste“ lexikalisierte Ausdruck mit 2–4 Zeichen im *Sanguo zhi yanyi* ist *zhi yan wu yin* 直言無隱 („offen seine Meinung sagen“). Wie im Screenshot zu sehen gibt das *HYDCD* als *attestation* das 1928 veröffentlichte *Qing shi gao* an.[298] Es kann davon ausgegangen werden, dass das *Sanguo zhi yanyi* ein deutliches *ante-dating* für diesen Ausdruck ermöglicht.

Das zweite ins 20. Jh. datierte Lexem, *juesuan* 決算 („Abschlussrechnung“), ist ein anschauliches Beispiel für *false positives*, da die beiden Zeichen im *Sanguo zhi yanyi* eher in ihren Einzelbedeutungen (etwa: „Pläne machen“ oder „Pläne ausführen“) zu verstehen sind.[299]

298 Siehe auch *HYDCD*, Bd 1., S. 858.

299 „後人有詩讚玄德曰運籌決算有神功，[…]“: „Die nachfolgenden Generationen machten ein Lied, in welchem sie Xuande 玄德 (i. e. Liu Bei 劉備) lobten: ‚Beim Entwerfen von Strategien und Machen von Plänen brachte er wunder-

Die Unvollständigkeit der diachronen Lexemdatenbank und die Problematik der fehlenden Tokenisierung oder sogar semantischen Analyse des zu datierenden Textes werden an den beiden Beispielen deutlich. Eine eingehende Untersuchung, wie sie z. B. in Abschnitt 6.2.4 anhand des *Zhongjing* 忠經 demonstriert wurde,[300] ist unter diesen Umständen mit zahlreichen *false positives* aufwändig – für kürzere Texte aber ein gangbarer Weg.

Datierung per *AYL*

Auf der Ergebnisseite wird zusätzlich die Datierung mittels *AYL* angezeigt.[301] Um ein breiteres Spektrum an schriftsprachlichen Texten einordnen zu können, wird zur Projektion ein Regressionsmodell mit den Texten der Loewe und *zhengshi*-Korpora trainiert.[302] Wie bereits in Abschnitt 6.3.1 wird zur Berechnung des *AYL* die Lexikalisierung der 90 % häufigsten 2–4-Zeichen Lexem-*types* verwendet. Als Grundlage dient die Datenbankabfrage für die Erzeugung des Neologismusprofils. Da hierbei die zusätzlichen Korpus-Belegstellen verwendet werden,[303] muss auch das Training für die Linearregression entsprechend berechnet werden. Zur Projektion dient auf dieser Basis die Funktion:

$$Y_{proj.} = 2,41 \times AYL_{0,9}^{2--4} + 904,03$$

Für das in den Trainingsdaten des Modells enthaltene *Sanguo zhi* ergibt sich daraus eine geschätzte Entstehung im Jahr 521 (+ 224 Jahre), für das umgangssprachlichere *Sanguo zhi yanyi* das Jahr 962 (- ?560 Jahre). Vor allem letztere Zuordnung ist unbrauchbar und deutet an, dass eine genre- oder stilübergreifende Datierung mit dieser Methodik nicht möglich ist.

Dass Ergebnisse aller untersuchten Methoden verglichen werden können, erleichtert die Einschätzung ihrer Verlässlichkeit. Übereinstimmende oder ähnliche Ergebnisse sprechen für eine korrekte Zuordnung, zwischen lexikographischer und statistischer Datierung stark abweichende Ergebnisse sollten Anlass zur Skepsis sein. Gerade in solchen Fällen ist eine qualitative Untersuchung erforderlich. Die diachrone Detailanzeige der *types* in *VisualTime* stellt dafür eine Hilfestellung dar, auf Basis derer zusätzliche Quellen konsultiert werden können.

bare Meisterleistungen hervor […]'". Luo Guanzhong 羅貫中 2007 [?1522].
300 Siehe S. 192.
301 Siehe dazu auch die Darstellung in Abb. 6.39 auf S. 234.
302 Siehe dazu Abschnitt 6.3.1, v. a. Abb. 6.35, S. 225.
303 Siehe dazu Kapitel 5.5.4, S. 134.

7 Ergebnisse und Ausblick

> *„Though the title suggests that they might actually date texts, in fact they argue that linguistic dating […] is a hopeless enterprise."*[1]
>
> Robert D. Holmstedt

In diesem letzten Kapitel werden die wichtigsten Resultate dieser Arbeit zusammengefasst und abschließend diskutiert. Es wird überdies ein Ausblick auf weiterführende Forschungsansätze gegeben, die Ergebnisse der vorliegenden Untersuchung aufgreifen.

Grundlage jeder linguistischen Datierung von Texten ist ein Verständnis diachroner sprachlicher Entwicklungen. In Kapitel 2 wurde daher zunächst der Sprachwandel im Hinblick auf für die Textdatierung relevante Aspekte beleuchtet. Davon ist der Wortschatzwandel für das Chinesische bislang aber am wenigsten systematisch erforscht. Es wurde auf das Piotrowski-Gesetz Bezug genommen und die bisher spärliche Forschung zu seiner Gültigkeit für das Chinesische betrachtet.

In einer Zusammenfassung des Forschungsstands zu Aspekten der historischen Entwicklung der chinesischen Sprache hat sich angedeutet, dass die Rigidität der chinesischen Schriftsprache bzw. der Konservativismus einiger Textgattungen und Sprachstile die Textdatierung erschweren und einige klassische Texte auch bei sorgfältiger Exegese nicht allein mit Methoden der historischen Sprachwissenschaft datiert werden können. Dies gilt umso mehr für Texte mit einer komplexen Überlieferungsgeschichte, wenn sie aus Fragmenten kompiliert wurden und/oder Inhalt und Form des Textes sich im Laufe der Jahrhunderte verändert hat, so dass teilweise eine Unschärfe in der Datierung von mehreren hundert Jahren akzeptiert werden muss.

Phonologische Veränderungen an der Sprache, die sich bei der Verwendung von Alphabetschriften oft in orthographischen Veränderungen widerspiegeln, werden mit chinesischen Zeichen in der Regel nicht verschriftlicht und stilistische, sowie syntaktische Veränderungen geschehen in der Schriftsprache sehr langsam. Dennoch konnten auch in dem stilistisch recht homogenen Korpus der offiziellen Dynastiegeschichten (*zhengshi* 正史) leichte Veränderungen in der Häufigkeit von Funktionswörtern beobachtet werden, die einen syntaktischen und stilistischen Wandel – unter Vorbehalt des engen Betrachtungsrahmens – nahelegen. Offensichtlichere Veränderungen finden aber im Wortschatz statt, was im selben Textkorpus gezeigt wurde. Als Beispiele für solche lexikalischen Innovationen und Veränderungen wurden die Verwendung einiger Amtstitel und mit dem Buddhismus verbundener Begriffe analysiert. Die Entstehung neuer Wörter ist dabei einfacher nachvollziehbar als ihr Aussterben. Neologismen können Indizien für die Textdatierung auch dann liefern, wenn Struktur und Stil eine hohe Kontinuität aufweisen.

1 Robert D. Holmstedt 2009: *Dating the Language of Ruth: A Study in Method*. Paper presented at the annual meeting of the Canadian Society of Biblical Studies (Ottawa, May 23, 2009). URL: http://individual.utoronto.ca/holmstedt/Holmstedt_DatingLangRuth_CSBSrevAug2009.pdf (besucht am 02. 08. 2021), S. 1. Holmstedt geht es in diesem Zitat um die Datierung althebräischer Texte in Young und Rezetko 2014.

In Kapitel 3 wurden sinologische Aspekte der linguistischen Datierung aufgezeigt und ein systematischer Überblick über die verfügbaren computerlinguistischen Methoden gegeben, die für die Datierung von indoeuropäischen Texten eingesetzt werden. Dabei kommen vor allem statistische Sprachmodelle zum Einsatz, die mit diachronen Textkorpora trainiert werden müssen und denen meist eine *Bag of Words* zugrunde liegt. Weitere Datierungsmethoden greifen auch auf Metadaten und explizite Zeitangaben in Texten zurück.

Für schriftsprachliches Chinesisch ergaben sich daraus zwei wesentliche Herausforderungen, die in Kapitel 4 diskutiert wurden: Diachrone Korpora, wie sie im Optimalfall für die Evaluation von Datierungsmethoden eingesetzt werden, stehen nicht in gewünschter Qualität und Umfang zur Verfügung. Eine diachrone Erprobung von Tokenizern hat zudem gezeigt, dass *state of the art* Segmentierung und *PoS*-Tagging für moderne Texte und neuerdings auch für klassisches Chinesisch in ausreichender Genauigkeit möglich ist. Für den langen, dazwischen liegenden Zeitraum bleiben die Ergebnisse aber unbefriedigend, da Wortgrenzen oft nicht korrekt erkannt werden. Mit einem einfachen *maximum matching*-Ansatz konnten teilweise bessere Erfolge erzielt werden, wenn individuell auf den Text angepasste, diachrone Wörterbücher verwendet wurden. Dies ist jedoch nur möglich, wenn die Entstehungszeit des zu segmentierenden Textes bekannt ist.

Es wurden zudem wichtige Aspekte von Namen und die Verwendung der *China Biographical Database* (*CBDB*) als Quelle für eine datenbankgestützte Erkennung von Namen diskutiert. Aus der Nutzung dieser biographischen Daten ergibt sich ein Potenzial, das für die Textdatierung über das einer herkömmlichen *Named Entity Recognition* hinausgeht, aber auch besondere Herausforderungen mit sich bringt. Es wurde festgestellt, dass dabei Ambiguitäten berücksichtigt werden müssen, die vor allem zweisilbige Namen betreffen. Als problematisch hat sich auch der hohe Anteil homonymer Personen herausgestellt, sowie vor allem *false positives* durch Namens-Zeichenkombinationen, die auch einfache Vorkommen lexikalisierter Bedeutungen dieser Zeichen sein können.

Inspiriert durch die für *MARKUS*[2] genutzte Implementierung der *Dharma Drum Buddhist College Time Authority Database*[3] konnte eine Möglichkeit der Erkennung und zeitlichen Einordnung von Regierungsdevisen erarbeitet werden, die für Zeitangaben in schriftsprachlichen Texten wesentlich sind. Anders als gregorianische Jahreszahlen beziehen sich diese in der Regel auf einen Zeitraum, der aus der Perspektive der Verfasser:in in der Vergangenheit, Gegenwart oder nahen Zukunft liegen muss.

Als Alternative zur Segmentierung wird die Zerlegung der Texte in *n*-Gramme vorgenommen. Hierfür wurde etabliert, dass für das schriftsprachliche Chinesisch die Betrachtung von maximal 1–4 bzw. 2–4 Grammen empfehlenswert ist. Um die Anzahl der *features* einer solchen Textrepräsentation zu begrenzen, wurde die Reduktion dieser *n*-Gramme auf lexikalisierte Wörter, Zeitausdrücke und Namen vorgeschlagen. Zur Erprobung von Datierungsmethoden konnten somit auch *n*-Gramm-Datensätze verwendet werden, wie sie seit 2019 von CrossAsia bereitgestellt werden.[4]

Mit dem Ziel, eine Datengrundlage für Textdatierungsmethoden zu schaffen, bei denen der Aspekt der Wortbildung im Vordergrund steht, wurde in Kapitel 5 aus einer digitalen Ausgabe

2 Ho und DeWeerdt. 2014–.
3 *DDBC*.
4 Siehe CrossAsia, Staatsbibliothek zu Berlin 2019–: *CrossAsia N-Gramm Service*. Website. URL: https://crossasia.org/service/crossasia-lab/crossasia-n-gram-service/ (besucht am 19. 04. 2019).

des *Hanyu da cidian* 漢語大詞典[5] (*DHYDCD*) erstmals eine diachrone Lexemdatenbank für das Chinesische erzeugt. Die Wahl von *Python* für die Implementierung der notwendigen Software hat sich – wie für alle weiteren Programmierungen im Rahmen dieser Arbeit – als adäquat erwiesen. Die Lexikalisierungsdaten aus dem *DHYDCD* konnten mithilfe biblio- und biographischer Daten aus der *CBDB* sowie durch Betrachtung datierter Primärtexte weiter verdichtet werden, so dass eine Datenbank mit 272.892 Lexemen zur Verfügung steht, die mit einer Genauigkeit von ca. 80 Jahren datiert sind. Hinzu kommen über 600.000 zugehörige Textbelegstellen, aus denen nach dem Vorbild von Hoffmann[6] ein diachrones Behelfskorpus generiert wurde. Daraus konnten temporale Sprachmodelle für einen Betrachtungszeitraum von über 2.000 Jahren erzeugt werden.

Die Datenbank hat zudem die Gewinnung neuer Erkenntnisse über Machart sowie Stärken und Schwächen des *HYDCD* ermöglicht. Eine Analyse der verwendeten Belegstellen konnte lexikographische Präferenzen für bestimmte Texte und Textgattungen aufdecken, in der eine tiefe Verwurzelung des Wörterbuchs in der schriftsprachlichen Texttradition zum Ausdruck kommt.

Mit einer chronologischen Analyse der Lexikalisierungsdaten wurde überdies neues Licht auf die Geschichte des chinesischen Wortschatzes geworfen. Was in der Forschungsliteratur als unscharf getrennte Phasen der Sprachentwicklung dargestellt wird, ließ sich als kontinuierliche, logistische Entwicklung darstellen. Diese Beobachtung legt nahe, dass das Piotrowski-Gesetz auch für die Modellierung des Wortschatzwachstums des Chinesischen geeignet ist. Ob Schwankungen in der Aufnahme neuen Vokabulars bestimmte (Entlehnungs-)Wellen oder Krisen widerspiegeln, ließ sich dabei nicht eindeutig klären. Dass Krisen einen immensen sprachlichen Innovationsschub bewirken können, zeigt sich aktuell in einer Vielzahl von Wortbildungen wie „Flockdown“ oder „Lollitest“.[7] In der retrospektiven Lexikographie werden solche aber nur erfasst, wenn sie sich längerfristig im Sprachgebrauch durchsetzen können.

Es konnte verdeutlicht werden, dass bereits ab dem 4. Jh. v. u. Z. ein hoher Anteil nicht nur zwei, sondern auch drei- und viersilbiger Wortbildungen belegt sind. Während dabei anfangs noch eine – erwartbare – Präferenz für tetra- gegenüber trisyllabischen Lexemen besteht, ist der Anteil an drei- und viersilbigen Wortbildungen ab dem 4./5. Jh. weitestgehend identisch. Mehr als die Hälfte der Wörter, sogar in modernen Texten, bleiben einsilbig,[8] aber über 80 % der Einträge im *DHYDCD* sind disyllabische Wortbildungen. Entgegen dem Eindruck, der durch von Jahrhundert zu Jahrhundert wachsende Zeichenwörterbücher geweckt werden kann, sind fast alle auch heute gebräuchlichen Schriftzeichen – von vereinfachten Formen abgesehen – bereits in der Han-Zeit vorhanden, während sich neue Zeichen – im Gegensatz zu mehrsilbigen Wortbildungen – nur schwer durchsetzen können.

In Kapitel 6.1 wurden für westliche Sprachen erfolgreiche Methoden der Textdatierung auf Basis temporaler Sprachmodelle implementiert und auf ihre Eignung für schriftsprachliche chinesische Texte getestet. Solche Ansätze, bei denen die Datierung als Kategorisierungsproblem aufgefasst wird, haben sich als auf beliebige Entwicklungsstufen des Chinesischen übertragbar erwiesen. Mit dem *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌, *DFZ*) konnten z. B. mehr als 60 % der Testdaten aus demselben Korpus einem korrekten

5 *DHYDCD*.
6 Siehe Hoffmann 2004.
7 Eine umfassende Sammlung von Neologismen, die mit der Coronapandemie in Verbindung stehen, findet sich in: Leibniz-Institut für Deutsche Sprache (IDS) 2020.
8 Siehe auch Breiter 1994, 224ff.

chronon, hier einem Zeitraum von 50 Jahren, bei einem Betrachtungsfenster von 450 Jahren, korrekt zugeordnet werden. Die durchschnittliche Genauigkeit der Datierung lag bei etwa 42 Jahren.[9] Als Ähnlichkeitsmaße haben sich dabei die *Normalized-Log-Likelihood-Ratio* bzw. die Kullback-Leibler-Divergenz am stärksten bewährt. Durch Gewichtung der betrachteten *types* mittels *Temporaler Entropie* konnte nur teilweise eine leichte Verbesserung der *Accuracy* erzielt werden. Bei einer Analyse von Glättungsmethoden hat sich die naive Annahme einer Häufigkeit von *unseen events*, die niedriger als die geringste im Korpus beobachtete Häufigkeit ist, als am effektivsten herausgestellt.

Um die zeitliche Einordnung von Texten auch über Genregrenzen hinweg und für einen größeren Zeitraum zu ermöglichen, wurden zudem temporale Sprachmodelle aus den Belegstellen im *DHYDCD* erzeugt. Die Klassifizierung von Texten in *chronons* von 100 Jahren bei einem Zeithorizont von über 2.000 Jahren kann damit grundsätzlich als vielversprechend angesehen werden. Wie Tests mit dem *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*) gezeigt haben, ist die Datierung schriftsprachlicher Texte mit solchen Sprachmodellen gleichzeitig aber zutiefst problematisch. Dies liegt primär am Modellcharakter einiger antiker Texte, deren Stil und damit Grammatik über Jahrtausende als Vorbild gedient haben. Dass für westliche Sprachen erfolgreiche Datierungsmethoden an solchen Texten scheitern, verdeutlicht die beeindruckende Kontinuität der Tradition bestimmter schriftsprachlicher Textgattungen. Das rigide Schriftsystem, das von kurzlebigen orthographischen Anpassungen aufgrund von phonologischem Wandel weitestgehend unberührt bleibt, verstärkt diesen Effekt.

In Kapitel 6.2 wurde mit temporalen Textprofilen eine neue lexikographische Methode eingeführt, die eine temporale Einordnung von Texten auf Basis der frühesten Belege der enthaltenen Lexeme, sowie den vorkommenden Namen und Zeitangaben ermöglicht. Die gewählte Darstellung, die die Anzahl der pro Jahrhundert zugeordneten *types* als Balkendiagramm visualisiert, kann als Hilfestellung für qualitative Analysen genutzt werden. Anhand der Profile lässt sich ablesen, wieviele *types* durch Textbelege bzw. biographische Daten und Zeitangaben mit einem Jahrhundert verbunden sind, was Rückschlüsse sowohl über Inhalt als auch Genese des Textes zulässt – unabhängig von Genre oder Alter des Eingabetextes.

Mithilfe von Trainingsdaten können diese temporalen Profile auch für die automatische Datierung von Texten eingesetzt werden. Bei Experimenten mit dem *DFZ*-Datensatz wurden dabei Ergebnisse erzielt, die mit denen bei Verwendung temporaler Sprachmodelle vergleichbar sind. Etwa 50 % der untersuchten Texte konnten auf etwa 100 Jahre genau eingeordnet werden, wobei der Betrachtungszeitraum über 2.000 Jahre betrug. Bei zusätzlicher Betrachtung von Personennamen konnte der Anteil der korrekt datierten Texte auf über 60 % erhöht werden. Bei einer einfachen Betrachtung von *temporal expressions* konnten in diesem Korpus sogar 88 % der Texte korrekt zugeordnet werden. Die Besonderheit, dass Zeitangaben in Form von Regierungsdevisen erfolgen, hat sich dabei als nützlich herausgestellt, da – im Gegensatz zu westlichsprachigen Texten – damit keine Referenz auf die ferne Zukunft erfolgen kann.

Experimente mit dem *XXSKQS*-Datensatz konnten zudem zeigen, dass eine automatisierte lexikographische Datierung auch unabhängig von Trainingsdaten erfolgreich sein kann.

Neben der ungefähren zeitlichen Einordnung von Dokumenten kann das Konzept der temporalen Textprofile auch zur Unterstützung bei der Analyse von vermuteten Fälschungen, bzw. Texten von fragwürdiger Provenienz oder ungeklärter Autorschaft herangezogen werden. Da die zur Verfügung stehenden Lexikalisierungsdaten große Ungenauigkeiten aufweisen können

9 Siehe Kapitel 6.1.1, ab S. 158.

und die Verwendung von *n*-Grammen zur Segmentierung der Texte einen hohen Anteil an *false positives* mit sich bringt, erfordert dieser Anwendungsbereich ein hohes Maß zusätzlicher, sorgfältiger Recherche.[10]

In Kapitel 6.3 wurde zuletzt ein experimenteller Ansatz betrachtet, der die bereits in 6.2 verwendeten Lexikalisierungsdaten aus dem *DHYDCD* auf einen einzigen Messwert abstrahiert: das Jahr der durchschnittlichen Lexikalisierung der enthaltenen Lexeme. Mit den *DHYDCD*-nahen *zhengshi* 正史 konnte dabei ein verblüffender Zusammenhang zwischen diesem *Average Year of Lexicalization* (*AYL*) und der ursprünglichen Entstehungszeit beobachtet werden, der im betrachteten Zeitraum linear zu sein scheint. Es wurde unter anderem festgestellt, dass der statistische Zusammenhang zwischen Textentstehung und *AYL* besser ist, wenn Unigramme unberücksichtigt bleiben und stattdessen nur ein geringer Anteil der häufigsten 2–4-Gramme betrachtet wird. Weitere Experimente legen aber nahe, dass diese Art der Modellierung nur innerhalb der engen Grenzen eines homogenen Korpus gut funktioniert.

Zuletzt wurden in Kapitel 6.4 die drei für das schriftsprachliche Chinesische untersuchten Datierungsmethoden gegenübergestellt und ein *user interface* vorgestellt, das für einen hochgeladenen *Plain Text* einen Vergleich der Ergebnisse im Browser ermöglicht.

Unabhängig von der gewählten Methodik wurden die Limitationen, schriftsprachliche chinesische Texte (computer)linguistisch zu datieren, immer wieder sichtbar. Bei der automatisierten Textdatierung kann der korrekte Zeitstempel für zu datierende Texte um viele hundert Jahre, in Einzelfällen sogar mehr als ein Jahrtausend, verfehlt werden. Davon betroffene Texte weisen tendenziell ein hohes Maß an Intertextualität auf, bilden ein „Mosaik von Zitaten“[11] älterer Schriften oder kommentieren diese.

Dass solche Texte fälschlich etwa der Zeit zugeordnet werden, deren Syntax und Stil darin imitiert oder aufgegriffen wird, kann in gewisser Hinsicht auch als Bestätigung dafür gewertet werden, dass die chronologische Einordnung grundsätzlich funktioniert. Manche Autor:innen kommen gänzlich ohne zeitgenössisches Vokabular aus, was entsprechende Texte – auch bei Recherche der einzelnen *types* aus temporalen Textprofilen – gegen eine rein linguistische Analyse vollkommen resistent macht. Wegen des inhaltlichen und sprachlichen Vergangenheitsbezugs könnten solche Texte in gewisser Hinsicht auch als „transtemporal“ bezeichnet werden.[12] Für die Klärung der Authentizität oder Entstehungszeit solcher Dokumente verbleibt nur der Verweis auf die qualitative, inhaltliche und kontextuelle Analyse, sowie die zusätzlichen Möglichkeiten, die gegebenenfalls für physisch vorliegende Exemplare zur Verfügung stehen: Bestimmung des Alters des verwendeten Papiers bzw. Trägers, der verwendeten Farbe, Analyse von Kalligraphie-, Zeichen- und Druckstilen,[13] sowie Tabuzeichen.[14] Derartige Untersuchungen erlauben Rückschlüsse aber stets nur auf die vorliegende Kopie, auf eine Ausgabe, die Manifestation eines Textes, nicht über seine ursprüngliche Genese.

10 Siehe zur Genauigkeit der Lexemdatierungsdaten und zur Polysemie der Lexeme auch Kapitel 5.7.1, ab S. 139. Viele der erkannten Zeichenkombinationen sind zudem im geg. Kontext keine „Wörter“, vgl. auch Kapitel 2.3 und 5.5.4.

11 Julia Kristeva 1972 [1965]: „Bachtin, das Wort, der Dialog und der Roman“. In: *Zur linguistischen Basis der Literaturwissenschaft II*. Hrsg. von Jens Ihwe. Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven Bd. 3. Frankfurt am Main: Athenäum, S. 345–375, S. 348. Kristevas Formulierung scheint hier sehr zutreffend, obwohl sie ihrem literaturwissenschaftlichen Kontext entrissen ist.

12 Der Begriff der Transtemporalität im Kontext von Anspielungen auf die Vergangenheit ist entlehnt aus Christian Soffel 2020: „Transcultural Aspects in Chang Yi-Jen's 張以仁 Poetry“. In: *Interface – Journal of European Languages and Literatures* 12.2, S. 113–137. doi: 10.6667/interface.12.2020.111, S. 133.

13 Vgl. aber Galambos 2006.

14 Siehe dazu Kapitel 4.3, sowie v. a. Adamek 2012.

Wichtige Texte der schriftsprachlichen bzw. klassischen Texttradition, wie das *Shijing* 詩經 (*Buch der Lieder*) oder das *Shangshu* 尚書 (*Buch der Urkunden*), deren Datierung besonders intensiv diskutiert wird, wurden vollkommen ungeachtet unserer heutigen, eurozentrischen Auffassung von Autorschaft über einen Zeitraum von mehreren hundert Jahren bis zum Erreichen ihrer heutigen Form immer wieder verändert oder ergänzt. Das Streben nach einer präzisen Datierung solcher Textgewebe wirkt in diesem Kontext ebenso unpassend, wie es die Zuschreibung zu einer einzigen Autor:in wäre.[15] Zumindest müsste die ohnehin komplexe Frage nach der Datierung durch zusätzliche Aspekte verkompliziert werden.[16] Der Versuch einer genauen, linguistischen Datierung solcher Texte gerät zu einer Art Henne-Ei-Problem, wenn Trainingsdaten bzw. Primärquellen selbst nur sehr vage datiert sind, denn „eine robuste lexikalische Chronologie setzt zunächst die Existenz einer großen Reihe von sicher datierten Texten voraus."[17] Gesicherte Belege lassen sich teilweise aus archäologischen Funden gewinnen, deren Alter naturwissenschaftlich bestimmt, oder aus dem Kontext des Fundorts ermittelt werden kann.[18] Tatsächlich kann durch solche Funde aber nur – wenn überhaupt – das *Mindest*alter eines Texts korrigiert werden, da sie keinen Aufschluss über die ursprüngliche Textgenese erlauben.

7.1 Ausblick

Naheliegend für eine Weiterentwicklung der in Kapitel 6.2 vorgestellten Methodik zur zeitlichen Einordnung von Texten ist eine kontinuierliche Verbesserung der Datengrundlage. Dies kann durch eine „mitlernende" Datenbank geschehen, in die laufend zusätzliche Lexeme und Namen sowie Belegstellen – auch zu bereits vorhandenen Einträgen – aus sicher datierten Texten aufgenommen werden. Die Verwendung der Profile für die Analyse moderner chinesischer Texte könnte dann ebenfalls erwogen werden. Auch ein *crowd sourcing* mit Ergänzung von Worteinträgen und Belegen durch Anwender:innen ist denkbar. Die Zuverlässigkeit der Textprofile könnte so immer weiter verbessert und letzlich auch eine feinere Granularität erreicht werden als der momentan noch unscharfe Zeitraum von 100 Jahren. Dazu müsste allerdings die Toleranzschwelle, wie viele „zu neue" Lexeme ein Text enthalten darf, ebenfalls kontinuierlich trainiert werden. Unlösbar bleibt die Tatsache, dass viele schriftsprachliche Texte bewusst in einem antiken Stil verfasst sind und wenig bis gar keine zeitgenössischen Lexeme enthalten.

Neben der möglichst genauen Sammlung von *Loci classici* sollte auch die systematische Erfassung des Lebenszyklus von Wörtern angestrebt werden. Die im Rahmen von Kapitel 5.5 gesammelten Belege können dafür einen Anfang bilden und perspektivisch auch helfen, Fragen über das Verschwinden von Wörtern zu beantworten. Überdies ist auch das Potenzial des *DHYDCD* als Datenquelle noch nicht ausgeschöpft, da neben den hier verwendeten Informationen zum lexikalischen Sprachwandel auch Belege zu unterschiedlichen Bedeutungen gegeben

15 Vgl. z. B. Edward Shaughnessy 1993b: „*Shang shu* 尚書 (*Shu ching* 書經)". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 376–389, v. a. S. 376–380; vgl. z. B. Boltz 2007, S. 70–71; für eine Diskussion des Autoschaftsbegriffs für Asien siehe auch Christian Schwermann und Raji C. Steineck 2014: „Introduction". In: *That Wonderful Composite Called Author*. Hrsg. von Christian Schwermann und Raji C. Steineck. East Asian Comparative Literature and Culture 4. Leiden: Brill, S. 1–29, v. a. S. 20–26.

16 Siehe dazu auch Kapitel 3, S. 36, sowie Harbsmeier 2019, S. 189.

17 Toner und Han Xiwu 2019, S. 36, übersetzt durch den Verfasser.

18 Ein Beispiel sind umgangssprachliche Funde aus Dunhuang 敦煌, siehe z. B. Jing-Schmidt und Hsieh 2019, S. 516; Als weiteres Beispiel seien die 1993 in Guodian 郭店 gefundenen Bambusstreifen genannt, die auf die Mitte des 4.–3. Jh. v. u. Z. datiert werden können. Siehe z. B. Shirley Chan 2019: „Introduction: The Excavated Guodian 郭店 Bamboo Manuscripts". In: *Dao Companion to the Excavated Guodian Bamboo Manuscripts*. Hrsg. von Shirley Chan. Dao Companions to Chinese Philosophy 10. Cham: Springer Nature, S. 1–20. DOI: 10.1007/978-3-030-04633-0_1, S. 4–7.

werden. Aus diesen ließen sich auch umfangreiche Daten zum semantischen Wandel gewinnen. Um sie für die Datierung nutzbar zu machen, wäre selbstverständlich auch eine semantische Analyse der zu datierenden Texte erforderlich.[19]

Dass „es für fast alle europäischen Sprachen historische Wörterbücher gibt, die – oft genaue – Angaben über die erste schriftliche Erwähnung eines jeden Wortes liefern",[20] macht es interessant, Neologismusprofile auch für europäische Sprachen einzusetzen. Mit der digitalen Fassung des *Oxford English Dictionary*[21] oder der Neubearbeitung des *Deutschen Wörterbuchs* von Jacob und Wilhelm Grimm[22] stehen Quellen zur Verfügung, in denen – im Gegensatz zu den oft vagen Angaben im *HYDCD* – die Belege mit Jahreszahlen angegeben werden. Entsprechende Experimente ließen also sogar eine genauere Darstellung zu. Dass die chinesische Sprache stark isolierend und ihre Schrift eine fast zeitlose, von orthographischen Veränderungen verschonte Schreibung ermöglicht, war hier von Vorteil. Ein Abgleich historischer, westlicher Texte mit diachronen Wörterbüchern setzt hingegen eine intensivere Beschäftigung mit der Normalisierung von Schreibweisen und mit Lemmatisierung voraus.

Entwicklungen sowohl in Bereichen des *machine learning*, sowie wachsende digitale Sammlungen schriftsprachlicher, darunter auch mittelchinesischer und spätkaiserzeitlicher Texte lassen in naher Zukunft auch eine befriedigende Tokenisierung und vor allem ein *PoS-Tagging* für solche Texte möglich erscheinen. Die Erfassung von Wortarten hätte dabei ein großes Potenzial in der Erforschung von Sprachwandel, da anstatt bloßer Vorkommen von Zeichen und Zeichenkombinationen Wörter und – zumindest in Teilen – auch ihre unterschiedlichen Bedeutungen betrachtet werden könnten. Eine Analyse der Häufigkeit von *zhi* 之 in Kapitel 2.3 konnte dies nur erahnen lassen.[23] Dass durch *PoS-Tagging*, sowie auch durch die Unterscheidung von Bedeutungen und Berücksichtigung gängiger Wortverbindungen (*collocations*) auch die Aussagekraft temporaler Sprachmodelle verbessert werden kann, wurde von Kanhabua und Nørvåg bereits gezeigt.[24] Auf Datensätze mit *n*-Gramm Häufigkeiten können diese Techniken allerdings nicht angewandt werden.

Ähnlich den in Kapitel 6.1 verwendeten temporalen Sprachmodellen könnten Methoden aus dem Bereich des *machine learning* auch für die Einordnung (schriftsprachlicher) chinesischer Texte getestet werden. Als Einstiegsmöglichkeit bieten sich *support vector machines* an, die bereits im Bereich der Textdatierung eingesetzt wurden.[25]

Spannend wäre dabei auch zu sehen, ob Textdatierungen für modernes (*xiandai* 現代) Chinesisch innerhalb des 20. Jahrhunderts spürbar genauer möglich sind, als dies bei den in relativ große *chronons* eingeordneten schriftsprachlichen Texten der Fall war. Der stark angestiegene sprachliche Wandel, wie ihn die Beobachtungen aus Kapitel 6.1.4 für diesen Zeitraum andeu-

19 Bei Verfügbarkeit passender Trainingsdaten, stehen *machine learning* Technologien zur Verfügung, mit denen sich entsprechende Modelle trainieren ließen. Dazu zählen neben sogenannten *word embeddings* auch die bereits in Kapitel 4.5 (ab S. 77) erwähnten *Transformers*.

20 Alinei 2004, S. 213, übersetzt durch den Verfasser.

21 John A. Simpson, Hrsg. 2014 [2002]: *Oxford English Dictionary, 2nd Edition [Second edition on CD-ROM Version 3.0]*. Oxford University Press. URL: http://njw.me.uk/oed (besucht am 10. 04. 2014).

22 *Deutsches Wörterbuch von Jacob* Grimm *und Wilhelm* Grimm*, Neubearbeitung (A–F), Version 01/21* 2021. Digitale Fassung im Wörterbuchnetz des Trier Center for Digital Humanities. URL: https://www.woerterbuchnetz.de/DWB2 (besucht am 30. 05. 2021).

23 Siehe Kapitel 2.3, ab S. 24.

24 Siehe Kanhabua und Nørvåg 2008, S. 361, S. 367–369, siehe auch Kapitel 3.3, ab S. 45.

25 Siehe Garcia-Fernandez et al. 2011, S. 8–10, siehe auch Kapitel 3.3, ab S. 45. Die Verfügbarkeit von *Python*-Bibliotheken wie `scikit-learn` ermöglicht überdies Experimente mit zahlreichen weiteren Methoden, die dem *machine learning* zugerechnet werden können. Vgl. auch Fabian Pedregosa et al. 2011: „Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, S. 2825–2830; Vanderplas 2018.

ten,[26] sollte dies begünstigen. Für solche Analysen bieten sich aufgrund der genauen Datierung diachrone Sammlungen von Zeitungsartikeln, wie etwa aus der *Renmin ribao* 人民日報 (*RMRB*), an. Eine temporale Veränderung der darin behandelten *Themen* dürfte dabei allerdings einfacher zu beobachten sein als rein sprachliche Veränderungen.[27]

Unter dem Eindruck von Diskussionen um den Missbrauch von *Open Source* Software wie dem Apache Webserver im Einsatz bei Ölfirmen oder *NationBuilder* als Plattform für die Brexit-Kampagne,[28] scheint es zuletzt auch angezeigt, mögliche *misuse cases*[29] zu betrachten.[30] Besteht ein solches Dilemma auch für Software zur Datierung von Texten?

Sie ließe sich sicherlich einsetzen, um die Produktion glaubhafterer Fälschungen historischer Texte zu erleichtern, indem entlarvende Anachronismen vorab erkannt werden. Abgesehen vom zweifelhaften Nutzen eines solchen Unterfangens bleibt es der kritischen Leser:in unbenommen, eine Fälschung an anderen, inhaltlichen Kriterien zu erkennen, selbst wenn die Fälscher:in z. B. einen perfekten Song-zeitlichen Wortschatz verwendet. Die Datierungssoftware selbst wäre durch diesen Schwindel allerdings leicht hinters Licht zu führen. Der hier beschriebene *misuse case* kann zugleich ein valider Anwendungsfall sein, etwa wenn authentische Dialoge für fiktionale historische Werke wie Fernsehserien verfasst werden sollen. Näherliegend, gerade für das schriftsprachliche Chinesische, ist aber die Verwendung für die Vergabe von Zeitstempeln innerhalb von *NLP*-Workflows. Die zeitliche Einordnung des Textes kann so zur Verbesserung von *NER* oder der Tokenisierung von Texten beitragen und somit für eine breite Vielfalt computerlinguistischer Anwendungen genauso nützlich sein, wie für eine erleichterte Textlektüre.

26 Siehe ab S. 177.

27 Siehe aber Tang Xuri, Qu Weiguang und Chen Xiaohe 2015, Die Autoren untersuchen anhand von 59 Jahrgängen der *RMRB* den semantischen Wandel einzelner Lexeme wie *touming* 透明. vgl. auch de Jong, Rode und Hiemstra 2005, S. 4.

28 Siehe dazu z. B. Chris Jensen 2018: *Why we need an Open Source License that considers the misuse of our code*. URL: https://hackernoon.com/why-we-need-an-open-source-licence-that-considers-the-misuse-of-our-code-8d19b65d425 (besucht am 27. 09. 2018).

29 „A misuse case is a special kind of use case, describing behavior that the system/entity owner does not want to occur." Guttorm Sindre und Andreas L. Opdahl 2000: „Eliciting Security Requirements by Misuse Cases". In: *Proceedings of TOOLS Pacific*, S. 120–131, S. 122; Zu industrierelevanten *misuse cases* siehe z. B. auch Ian Alexander 2003: „Misuse Cases: Use Cases with Hostile Intent". In: *IEEE Software*, S. 58–66. DOI: 10.1109/MS.2003.1159030.

30 Eine ausführtliche Diskussion zu *Dual Use* und weiteren ethischen Fragen in den *DH* findet sich in Malte Rehbein und Christian Thies 2017: „Ethik". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 353–357, *passim*.

Literaturverzeichnis

„Si vous avez lu tout cela,
je vous plains.“[1]
Georges Heuyer

Anmerkung: Alle Quellen sind zugunsten der Auffindbarkeit alphabetisch sortiert, ohne Unterteilung in Primär- und Sekundärliteratur bzw. in gedruckte und elektronische Quellen.

Sigel

CBDB Michael A. Fuller 2017: *China Biographical Database Project (CBDB)*. URL: https://projects.iq.harvard.edu/cbdb (besucht am 24. 04. 2017).

DDBC Buddhist Studies Authority Database Project 佛學規範資料庫, Hrsg. 2010–2020: *Dharma Drum Buddhist College Time Authority Database*. GitHub Repository. New Taipei City 新台北市. URL: https://github.com/DILA-edu/Authority-Databases/tree/master/authority_time (besucht am 17. 10. 2020).

DFZ CrossAsia, Staatsbibliothek zu Berlin 2019a: *N-gram dataset of Chinese local gazetteers (Zhongguo Difangzhi* 中國地方誌). Datenset, Version 0.0.2-20190408. DOI: 10.5281/zenodo.2594596.

DHYDCD Luo Zhufeng 羅竹風, Hrsg. 2005: *Hanyu da cidian* 漢語大詞典 *UTF-8 (Großes Wörterbuch der chinesischen Sprache, Unicode-Version)*. Shanghai 上海. URL: http://bbs.gxsd.com.cn/forum.php?mod=viewthread&tid=498015 (besucht am 13. 01. 2013).

HHS Fan Ye 范曄 1965 [445]: *Hou Han shu* 後漢書. *12 Bde*. Beijing 北京: Zhonghua shuju 中華書局.

HYDCD Luo Zhufeng 羅竹風, Hrsg. 1986–1994: *Hanyu da cidian* 漢語大詞典 *(Großes Wörterbuch der chinesischen Sprache)*. Bd. 1-13. Shanghai 上海: Cishu chubanshe 辭書出版社.

HYDZD Xu Zhongshu 徐中舒, Hrsg. 1986–1990: *Hanyu da zidian* 漢語大字典 *(Großes Lexikon chinesischer Schriftzeichen)*. 3 Bde. Wuhan 武漢: Sichuan cishu chubanshe 四川辭書出版社, Hubei cishu chubanshe 湖北辭書出版社.

MXBT Shen Kuo 沈括 2008 [1088]: *Meng xi bi tan* 夢溪筆談 *(Pinselunterhaltungen am Traumbach)*. Project Gutenberg eBook. URL: http://www.gutenberg.net (besucht am 10. 09. 2018).

OED James A. H. Murray et al., Hrsg. 1913–1933: *Oxford English Dictionary*. Bd. 1–13. London: Oxford University Press.

XXSKQS CrossAsia, Staatsbibliothek zu Berlin 2019b: *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書. Datenset. Version 0.0.1-20190307. DOI: 10.5281/zenodo.2586940.

1 Georges Heuyer, zitiert in Peter Schalmey 1977: *Die Bewährung psychoanalytischer Hypothesen*. Wissenschaftstheorie und Grundlagenforschung 7. Kronberg: Scriptor, S. 217.

A

Academia Sinica 中央研究院 1984–: *Han ji dianzi wenxian ziliaoku* 漢籍電子文獻資料庫 *(Scripta Sinica database)*. Website. URL: http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm (besucht am 07. 12. 2021).

Academia Sinica, Center for Digital Cultures 中央研究院數位文化中心 2018: *Academia Sinica Digital Humanities Research Platform (Zhongyang yanjiuyuan shuwei renwen yanjiu pingtai* 中央研究院數位人文研究平台*)*. URL: https://dh.ascdc.sinica.edu.tw/ (besucht am 24. 01. 2021).

——— *Liang qian nian zhong-xi li zhuanhuan* 兩千年中西曆轉換 *(Umwandlung zwischen chinesischem und westlichem Kalender für 2000 Jahre)*. Website. URL: http://sinocal.sinica.edu.tw/ (besucht am 08. 09. 2019).

Adamek, Piotr 2012: „Good Son is Sad If He Hears the Name of His Father: The Tabooing of Names in China as a Way of Implementing Social Values". Diss. Leiden: Leiden University.

——— 2015 [2012]: *Good Son is Sad If He Hears the Name of His Father: The Tabooing of Names in China as a Way of Implementing Social Values*. Monumenta Serica Monograph Series 66. St. Augustin & London [Leiden, Diss.]: Monumenta Serica & Routledge.

Aisin-Gioro Xuanye 愛新覺羅 · 玄燁 (als Shengzu ren huangdi 聖祖仁皇帝), Hrsg. 1922 [1716]: *Yuding Kangxi zidian* 御定康熙字典 *(„Kaiserliches Kangxi-Zeichenwörterbuch")*. Shanghai 上海: Tongwen shuju 同文書局.

Aitchison, Jean 2001 [1991]: *Language Change – Progress or Decay*. 3. Aufl. Cambridge: Cambridge University Press.

Akaike Hirotsugu 赤池弘次 1992 [1973]: „Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Hrsg. von Samuel Kotz und Norman L. Johnson. Bd. I: Foundations and Basic Theory. New York: Springer, S. 610–624.

Alexander, Ian 2003: „Misuse Cases: Use Cases with Hostile Intent". In: *IEEE Software*, S. 58–66. DOI: 10.1109/MS.2003.1159030.

Alford, William P. 1995: *To Steal a Book Is an Elegant Offense*. Stanford: Stanford University Press.

Alinei, Mario 2004: „The Problem of Dating in Linguistics". In: *Quaderni di semantica* 25.2, S. 211–232.

Allan, Kathryn 2012: „Using OED data as evidence". In: *Current Methods in Historical Semantics*. Hrsg. von Kathryn Allan und Justyna A. Robinson. Topics in English Linguistics. Berlin & Boston: Walter de Gruyter, S. 17–39.

Allison, Sarah et al. 2011: „Quantitative Formalism: an Experiment". In: *Pamphlets of the Stanford Literary Lab* 1, S. 1–24.

Altmann, Gabriel 1983: „Das Piotrowski-Gesetz und seine Verallgemeinerungen". In: *Exakte Sprachwandelforschung*. Hrsg. von Karl-Heinz Best und Jörg Kohlhaase. Göttingen: Herodot, S. 59–90.

Altmann, Gabriel et al. 1983: „A law of change in language". In: *Historical Linguistics*. Hrsg. von Barron Brainerd. Quantitative Linguistics 18. Bochum: Dr. N. Brockmeyer, S. 104–115.

Anderl, Christoph 2020: „Some Reflections on the Database of Medieval Chinese Texts as a Multi-Purpose Tool for Research, Teaching, and International Collaboration". In: *Corpus-Based Research on Chinese Language and Linguistics*. Hrsg. von Bianca Basciano, Franco Gatti und

Anna Morbiato. Sinica venetiana 6. Venezia: Edizioni Ca'Foscarini, S. 339–358. doi: 10.30687/978-88-6969-406-6/011.

Anderl, Christoph et al. 2015–: *A Database of Mediaval Chinese Texts*. url: https://www.database-of-medieval-chinese-texts.be/ (besucht am 30. 04. 2021).

Andrist, Eleni 2015: „Internet Language". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Apache Software Foundation 2011–2016: *Lucene*. url: https://lucene.apache.org/core/ (besucht am 01. 05. 2018).

Arapov, Mikhail V. 1983: „Word Replacement Rates for Standard Russian (A.D. 1100–1850)". In: *Historical Linguistics*. Hrsg. von Barron Brainerd. Quantitative Linguistics 18. Bochum: Dr. N. Brockmeyer, S. 50–61.

Arapov, Mikhail V. und Maja M. Cherc 1983 [1974]: *Mathematische Methoden in der historischen Linguistik [Matematičeskiye metody v istoričeskoy lingvistike, Математические методы в исторической лингвистике]*. Übers. von Reinhard Köhler und Peter Schmidt. Quantitative Linguistics 17. Bochum [Moskau]: Dr. N. Brockmeyer [Nauka].

Association for Computational Linguistics: *Association for Computational Linguistics*. Website. url: https://www.aclweb.org/ (besucht am 29. 09. 2018).

B

Bamman, David et al. 2017: „Estimating the Date of First Publication in a Large-Scale Digital Library". In: *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, Toronto, Canada, June 2017 (JCDL '17)*, S. 1–10. doi: 10.475/1234.

Baxter, William H. und Laurent Sagart 2014: *Old Chinese: A New Reconstruction*. Oxford & New York: Oxford University Press.

Behr, Wolfgang 2011: *The Analects: A Western Han Text?* Conference Presentation: The Lunyu as a Han Text, Princeton University. doi: https://doi.org/10.5281/zenodo.1405049.

——— 2018: „»Monosyllabism« and Some Other Perennial Clichés". In: *Asia and Europe – Interconnected: Agents, Concepts, and Things*. Hrsg. von Angelika Malinar und Simone Müller. Wiesbaden: Harrassowitz, S. 155–209.

——— 2019: „‚Urheimat' der Chinesen. Die Sprachwissenschaft und die Suche nach ‚Wurzeln'". In: *Geschichte der Gegenwart*. url: https://geschichtedergegenwart.ch/urheimat-der-chinesen-die-sprachwissenschaft-und-die-suche-nach-wurzeln/ (besucht am 24. 04. 2021).

Beijing yuyan xueyuan jiaoxue yanjiusuo 北京語言學院教學研究所, Hrsg. 1986: *Xiandai Hanyu pinlü cidian 現代漢語頻率詞典 (Häufigkeitswörterbuch der modernen chinesischen Sprache)*. Beijing 北京: Yuwen chubanshe 語文出版社.

Bentham, Jeremy 1825: *A Treatise on Judicial Evidence*. Hrsg. von Étienne Dumont. London: Baldwin, Cradock und Joy.

Benz, Wolfgang 2019 [2007]: *Die Protokolle der Weisen von Zion: Die Legende der jüdischen Weltverschwörung*. 4. Aufl. München: C. H. Beck.

Best, Karl-Heinz 2003: „Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes". In: *Glottometrics* 6, S. 9–34.

Best, Karl-Heinz und Zhu Jinyang 2006: „Sprachwandel im Chinesischen". In: *Archív Orientální* 74.2, S. 203–214.

Bibiko, Hans-Jörg 2006–: *Japanisch-Deutsches Kanji-Lexikon*. URL: https://mpi-lingweb.shh.mpg.de/kanji/ (besucht am 16. 05. 2021).

Bielenstein, Hans 1954: „The Restoration of the Han Dynasty. With Prolegomena on the Historiography of the Hou Han Shu". In: *BMFEA [Bulletin of the Museum of Far Eastern Antiquities]* 26, S. 1–209.

Bingenheimer, Marcus und Zhang Boyong 張伯雍 2017: *XML Data for „Four Early Chan Texts from Dunhuang - A TEI-based Edition"*. XML-Datensatz. Version Dec 2017. DOI: 10.5281/zenodo.1133490. (Besucht am 30. 04. 2021).

Bingenheimer, Marcus et al. 2016: „Modelling East Asian Calendars in an Open Source Authority Database". In: *International Journal of Humanities and Arts Computing* 10.2, S. 127–144. DOI: 10.3366/ijhac.2016.0164.

Binongo, Jose Nilo G. und M. W. A. Smith 1999: „The Application of Principal Component Analysis to Stylometry". In: *Literary and Linguistic Computing* 14.4, S. 445–465.

Bird, Steven, Ewan Klein und Edward Loper 2009: *Natural Language Processing with Python*. 1. Aufl. Sebastopol: O'Reilly.

——— 2014: *Natural Language Processing with Python*. 2. Aufl. URL: http://nltk.org/ (besucht am 12. 09. 2018).

Bochkarev, Vladimir, Valery Solovyev und Søren Wichmann 2014: „Universals versus historical contingencies in lexical evolution". In: *Journal of The Royal Society: Interface* 11.101. DOI: 10.1098/rsif.2014.0841.

Bojcov, Michail A. 2015: „Die Konstantinische Schenkung und ähnliche Gaben – im Westen und im Osten Europas". In: *Jahrbücher für Geschichte Osteuropas* 63.1, S. 23–46. URL: http://www.jstor.org/stable/43819721.

Bol, Peter K. 2020: „Introduction to the Utilities". In: *Journal of Chinese History* 4.2, S. 483–486. DOI: 10.1017/jch.2020.10.

Boltz, William G. 1993a: „Hsiao ching 孝經". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 141–153.

——— 1993b: „Shuo wen chieh tzu 說文解字". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 429–443.

——— 2007: „The Composite Nature of Early Chinese Texts". In: *Text and Ritual in Early China*. Hrsg. von Martin Kern. Seattle und London: Washington University Press, S. 50–78.

——— 2015: „Etymology". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Brainerd, Barron 1980: „The Chronology of Shakespeare's Plays: A Statistical Study". In: *Computers and the Humanities* 14, S. 221–230.

Breiter, Maria 1994: „Length of Chinese words in relation to their other systemic features". In: *Journal of quantitative linguistics* 1.3, S. 224–231.

Brezina, Vaclav 2018: *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge & New York: Cambridge University Press.

Brochlos, Astrid 2004: *Kanbun* 漢文の基礎 – *Grundlagen der klassischen sino-japanischen Schriftsprache*. Wiesbaden: Harrassowitz.

Buck, Christian und Philipp Koehn 2016: „Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance". In: *Proceedings of the First Conference on Machine Translation, Berlin, Germany, August 11-12, 2016*. Bd. 2. Berlin: Association for Computational Linguistics, S. 672–678.

Buddhist Studies Authority Database Project 佛學規範資料庫, Hrsg. 2010–2020: *Dharma Drum Buddhist College Time Authority Database*. GitHub Repository. New Taipei City 新台北市. URL: https://github.com/DILA-edu/Authority-Databases/tree/master/authority_time (besucht am 17. 10. 2020).

Bybee, Joan 2015: *Language Change*. Cambridge: Cambridge University Press.

C

Cao Liang, Wu Weiming und Gu Yonghao 2011: „The Research of Performance of Lucene's Chinese Tokenizer". In: *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*. IEEE, S. 7398–7401. DOI: 10.1109@AIMSEC.2011.6011478.

Chai Hongmai 柴红梅 2005: „To Remedy Some Flaws of Hanyu da cidian (汉语大词典) – On the Basis of Entry C in Xiandai Hanyu cidian (现代汉语词典)《汉语大词典》瑕疵补正 —- 以《现代汉语词典》C 字条为例". In: *Research In Ancient Chinese Language* 古汉语研究 3.

Chambers, Nathanael 2012: „Labeling Documents with Timestamps: Learning from their Time Expressions: Learning from their Time Expressions". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju* 濟州*, Republic of Korea, 8–14 July 2012*, S. 98–106.

Chan, Shirley 2019: „Introduction: The Excavated Guodian 郭店 Bamboo Manuscripts". In: *Dao Companion to the Excavated Guodian Bamboo Manuscripts*. Hrsg. von Shirley Chan. Dao Companions to Chinese Philosophy 10. Cham: Springer Nature, S. 1–20. DOI: 10.1007/978-3-030-04633-0_1.

Chan Wing-Tsit 陳榮捷 1963: *A Source Book in Chinese Philosophy*. Princeton, New Jersey: Princeton University Press.

Chen, Stanley und Joshua Goodman 1998: „An Empirical Study of Smoothing Techniques for Language Modeling". In: *Harvard Computer Science Group Technical Report* 10.

Chen Dingyi 2013: *libUnihan*. URL: https://sourceforge.net/projects/libunihan/ (besucht am 30. 11. 2016).

Chen Shih-Pei 陳詩沛 et al. 2017: *LoGaRT: Local Gazetteers Research Tools*. Software. Berlin: Max-Planck-Institut für Wissenschaftsgeschichte. URL: https://www.mpiwg-berlin.mpg.de/research/projects/logart-local-gazetteers-research-tools (besucht am 22. 10. 2021).

Chen Shih-Pei 陳詩沛 et al. 2020: „Local Gazetteers Research Tools: Overview and Research Application". In: *Journal of Chinese History* 4.2, S. 544–558. DOI: doi:10.1017/jch.2020.26.

Chen Shou 陳壽 1971 [297]: *Sanguo zhi* 三國志. *5 Bde*. Beijing 北京: Zhonghua shuju 中華書局.

Chen Xiaohe et al. 2017: *Ancient Chinese Corpus*. URL: https://catalog.ldc.upenn.edu/LDC2017T14 (besucht am 18. 10. 2017).

Chen Yanglu 陳錫輅 und Cha Qichang 查岐昌, Hrsg. 2016 [1754]: *[Qianlong] Guide fu zhi 36 juan [乾隆] 歸德府志 36 卷 ([Qianlong] Chronik der Präfektur Guide, 36 juan)*. Online-Datenbank Diaolong 雕龍 / *Zhongguo Difang zhi* 中國地方誌, via CrossAsia. Nagoya 名古屋 & Taipeh 台北: Kaixi MS 日本凱希多媒體 & tts 大鐸資訊.

Cheng Chung-ying 成中英 2019 [1971]: *Tai Chen's Inquiry into Goodness: A Translation of the Yuan Shan, With an Introductory Essay*. Honolulu: University of Hawai'i Press.

Cheng Yizhong 程毅中 et al., Hrsg. 2020–: *Shuzi renwen* 数字人文 *Digital Humanities*. Beijing 北京: Zhonghua shuju 中華書局.

Cheung Kwan-hin 張韋顯 und Robert S. Bauer 2002: *The Representation of Cantonese with Chinese Characters*. Bd. 18. Journal of Chinese Linguistics Monograph Series. Hong Kong: Chinese University Press.

China Biographical Database Project 2013: *Rules for Index Years*. URL: https://projects.iq.harvard.edu/cbdb/supporting-documents (besucht am 30. 11. 2017).

Chiru, Costin-Gabriel und Traian Rebedea 2014: „Archaisms and Neologisms Identification in Texts". In: *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference*. Chişinău: IEEE. DOI: 10.1109/RoEduNet-RENAM.2014.6955312.

Choi, Charles Q. 2020: *World's First Classical Chinese Programming Language*. URL: https://spectrum.ieee.org/tech-talk/computing/software/classical-chinese (besucht am 12. 09. 2020).

Chou Ya-Min 周亞民 und Huang Chu-ren 黃居仁 2010: „Hantology: conceptual system discovery based on orthographic convention". In: *Ontology and the Lexicon*. Hrsg. von Huang Chu-ren 黃居仁 et al. Studies in Natural Language Processing. Cambridge & New York: Cambridge University Press, S. 122–143.

CKIP Lab 2020: *CKIP Lab*. Website. URL: https://ckip.iis.sinica.edu.tw/ (besucht am 21. 05. 2021).

Coblin, South W. 1993: „Erh ya 爾雅". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 94–99.

Crespigny, Rafe de 2007: *A Biographical Dictionary of Later Han to the Three Kingdoms (23-220 AD)*. Handbook of Oriental Studies, Section Four: China 19. Leiden & Boston: Brill.

CrossAsia, Staatsbibliothek zu Berlin 2019a: *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌). Datenset, Version 0.0.2-20190408. DOI: 10.5281/zenodo.2594596.

——— 2019b: *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書. Datenset. Version 0.0.1-20190307. DOI: 10.5281/zenodo.2586940.

——— 2019–: *CrossAsia N-Gramm Service*. Website. URL: https://crossasia.org/service/crossasia-lab/crossasia-n-gram-service/ (besucht am 19. 04. 2019).

D

Dauses, August 1990: *Theorien des Sprachwandels – Eine kritische Übersicht*. Stuttgart: Franz Steiner Verlag.

Davis, Richard L. 2004: *Historical Records of the Five Dynasties*. New York: Columbia University Press.

de Jong, Franciska M. G., Henning Rode und Djoerd Hiemstra 2005: „Temporal Language Models for the Disclosure of Historical Text". In: *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*. Amsterdam: Koninklijke Nederlandse Academie van Wetenschappen, S. 161–168.

Debon, Günther 1989: *Chinesische Dichtung: Geschichte, Struktur, Theorie*. Handbook of Oriental Studies, Section Four: China 2. Leiden: Brill.

DeFrancis, John 1984: *The Chinese Language – Fact and Fantasy*. Honolulu: University of Hawaií Press.

deLeeuw, Jan 1992: „Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Hrsg. von Samuel Kotz und Norman L. Johnson. Bd. I: Foundations and Basic Theory. New York: Springer, S. 599–609.

Dennis, Joseph 2015: *Writing, Publishing, and Reading Local Gazetteers in Imperial China, 1100–1700*. Harvard East Asian Monographs 379. Cambridge, MA & London: Harvard University Asia Center, Harvard University Press.

Dershowitz, Nachum und Edward M. Reingold 2008: *Calendrical Calculations*. 3. Aufl. Cambridge & New York: Cambridge University Press.

Deutsches Wörterbuch von Jacob Grimm *und Wilhelm* Grimm*, Neubearbeitung (A–F), Version 01/21* 2021. Digitale Fassung im Wörterbuchnetz des Trier Center for Digital Humanities. url: https://www.woerterbuchnetz.de/DWB2 (besucht am 30. 05. 2021).

Devlin, Jacob et al. 2019: „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *ArXiv* 1810.04805. doi: 10.18653/v1/N19-1423.

deWeerdt, Hilde 2006: „Changing Minds through Examinations: Examination Critics in Late Imperial China". In: *Journal of the American Oriental Society* 126.3, S. 367–377. doi: 10.2307/20064514.

Ding, Chris und He Xiaofeng 2004: „*K*-means Clustering via Principal Component Analysis". In: *Proceedings Of International Conference of Machine Learning (ICML 2004)*. Hrsg. von Russ Greiner und Dale Schuurmans. New York: ACM Press, S. 225–232.

Django Software Foundation 2005–: *django – The web framework for perfectionists with deadlines*. url: https://www.djangoproject.com/ (besucht am 12. 10. 2021).

Dong Yubing 2012: *CSCI 562 Final Project, Building a machine translation system that translates Modern Chinese into Classical Chinese*. GitHub Repository. url: https://github.com/tomtung/nlp-class/blob/master/final/report.pdf.

Driscoll, Matthew J. 2007: *Electronic Textual Editing: Levels of transcription*. url: http://www.tei-c.org/Vault/ETE/Preview/driscoll.html (besucht am 25. 09. 2018).

E

Eder, Maciej 2018: „Words that Have Made History, or Modeling the Dynamics of Linguistic Changes“. In: *Digital Humanities 2018 Puentes-Bridges: Book of Abstracts*. Hrsg. von Jonathan Girón Palau und Isabel Galina Russell. Mexico City: El Colegio de México, S. 362–365.

Ellegård, Alvar 1953: *The auxiliary do: the establishment and regulation of its use in English*. Gothenburg studies in English. Göteborg: Almquist & Wiksell. URL: https://books.google.de/books?id=VcRZAAAAMAAJ.

Elman, Benjamin 2013: „The Civil Examination System in Late Imperial China, 1400-1900“. In: *Frontiers of History in China* 8.1, S. 32–50. DOI: 10.3868/s020-002-013-0003-9.

Emmerich, Reinhard 2004: „östliche Han bis Tang“. In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 88–186.

Ethan-yt 2020: *GuwenBERT Guwen yu xunlian moxing* 古文预训练模型. GitHub Repository. URL: https://github.com/Ethan-yt/guwenbert (besucht am 25. 05. 2021).

F

Fahrmeir, Ludwig et al. 2013: *Regression – Models, Methods and Applications*. Berlin & Heidelberg: Springer.

Fan Ye 范曄 1965 [445]: *Hou Han shu* 後漢書. *12 Bde*. Beijing 北京: Zhonghua shuju 中華書局.

Farmer, Edward L. et al. 1994: *Ming History: An Introductory Guide to Research*. Ming Studies Research Series 3. Minneapolis: Center for Early Modern History, University of Minnesota.

Feldman, Ronen und James Sanger 2006: *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Findeisen, Raoul David 2004: „Literatur im 20. Jahrhundert“. In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 288–395.

Forster, Malcolm und Elliot Sober 1994: „How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions“. In: *The British Journal for the Philosophy of Science* 45.1, S. 1–35.

Forsyth, Richard 1999: „Stylochronometry with substrings, or: A poet young and old“. In: *Literary and Linguistic Computing* 14. DOI: 10.1093/llc/14.4.467.

Fuller, Michael A. 2017: *China Biographical Database Project (CBDB)*. URL: https://projects.iq.harvard.edu/cbdb (besucht am 24. 04. 2017).

Furnivall, Frederick J. 1874: „Inaugural address to the New Shakspere Society“. In: *The New Shakspere Society's Transactions* 1.1–2, S. v–vi.

G

Gabelentz, Hans Georg Conon von der 1881: *Chinesische Grammatik: mit Ausschluss des niederen Stiles und der heutigen Umgangssprache*. Leipzig: T. O. Weigel.

——— 1901 [1891]: *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Hrsg. von Albrecht Graf von der Schulenburg. 2. vermehrte und verbesserte Auflage. Leipzig: Tauchnitz.

Galambos, Imre 2006: *Orthography of Early Chinese Writing: Evidence from Newly Excavated Manuscripts*. Budapest: Department of East Asian Studies, Eötvs Loránd University.

——— 2015: „Variant Characters". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Gamillscheg, Ernst 1928: *Die Sprachgeographie und ihre Ergebnisse für die allgemeine Sprachwissenschaft*. Bielefeld & Leipzig: Velhagen & Klasing.

Garcia-Fernandez, Anne et al. 2011: „When Was It Written? Automatically Determining Publication Dates". In: *String Processing and Information Retrieval*. Hrsg. von Roberto Grossi, Fabrizio Sebastiani und Fabrizio Silvestri. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 221–236.

Ge Hong 葛洪 2020 [Anfang 4. Jh.] *Baopuzi* 抱朴子. Digitalisierte Version der *Sibu congkan*-Ausgabe von *Baopuzi nei wai pian* 《四部叢刊初編》本《抱朴子內外篇》. URL: https://ctext.org/baopuzi (besucht am 20. 09. 2020).

Genette, Gérard 1982: *Palimpsestes: La littérature au second degré*. Paris: Éditions du Seuil.

Ginzburg, Carlo 1989: *Clues, Myths, and the Historical Method*. Baltimore & London: Johns Hopkins University Press.

Glahn, Richard von 1996: *Fountain of Fortune: Money and Monetary Policy in China, 1000–1700*. Berkeley: University of California Press.

Goethe, Johann Wolfgang von 1871 [1808]: *Faust: Eine Tragödie*. Berlin: G. Grote'sche Verlagsbuchhandlung.

Graliński, Filip et al. 2017: „The RetroC Challenge: How to Guess the Publication Year of a Text?" In: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. DATeCH2017. Göttingen: ACM, S. 29–34. DOI: 10.1145/3078081.3078095.

Greg, Walter W. 1950/51: „The Rationale of Copy-Text". In: *Studies in Bibliography* 3, S. 19–36. URL: https://www.jstor.org/stable/40381874.

Grimm, Jacob und Wilhelm 1854: *Deutsches Wörterbuch*. Bd. I. A–Biermolke. Leipzig: S. Hirzel.

Guan Xi 冠西 1997: „难忘罗老风范 (Schwer, das Gebaren des alten Luo [Zhufeng 羅竹風] zu vergessen)". In: *Renmin ribao* 人民日报 11.10.

Gumpen, Kristoffer Berg und Øyvind Vik Nygard 2017: „Automatic Document Timestamping". Masterarbeit. Trondheim: Norwegian University od Science und Technology (NTNU).

Guo Siyuan et al. 2015: „Toward Enhanced Metadata Quality of Large-Scale Digital Libraries: Estimating Volume Time Range". In: *iConference 2015 Proceedings*. Urbana: iSchools / University of Illinois. URL: http://hdl.handle.net/2142/73656.

H

Hallet, Wolfgang 2006: „Intertextualität als methodisches Konzept einer kulturwissenschaftlichen Literaturwissenschaft". In: *Kulturelles Wissen und Intertextualität. Theoriekonzeptionen und*

Fallstudien zur Kontextualisierung von Literatur. Hrsg. von Marion Gymnich, Birgit Neumann und Ansgar Nünning. Trier: Wissenschaftlicher Verlag Trier, S. 53–70.

Harbsmeier, Christoph 1998: *Language and Logic*. Hrsg. von Kenneth Robinson. Science and Civilization in China Volume 7, Part 1. Cambridge: Cambridge University Press.

——— 2019: „The Authenticity and Nature of the *Analects* of Confucius". In: *Journal of Chinese Studies* 68, S. 171–233.

Hargett, James M. 1990: „Review: Hanyu dacidian 漢語大詞典 by Luo Zhufeng 羅稷風". In: *Chinese Literature: Essays, Articles, Reviews (CLEAR)* 12, S. 138–143. doi: 10.2307/495232.

He Jiazheng 何加正 und Li Hongbing 李泓冰 1994: „中华民族五千年文化的结晶中国辞书出版史上的壮举《汉语大词典》大功告成首都隆重举行庆功会江泽民李鹏等到会祝贺全书 13 卷，收词语 37.5 万余条，约 5000 万字，是千余专家学者 18 年艰苦努力的结果 (Die Quintessenz der 5.000-jährigen Kultur des chinesischen Volkes, die Höchstleistung der chinesischen Geschichte der Herausgabe von Wörterbüchern, das *HYDCD*, wurde endlich abgeschlossen und zu diesem Anlass in der Hauptstadt eine große Feier ausgerichtet. An der Veranstaltung nahmen Jiang Zemin, Li Peng und andere teil. Das Werk hat insgesamt 13 Bände, 375.000 Wörter wurden aufgenommen, etwa 50 Mio. Zeichen, das Ergebnis der harten 18 Jahre dauernden Arbeit von über 1.000 Spezialisten und Gelehrten)". In: *Renmin ribao* 人民日报 05.11.

He Ying und Mehmet Kayaalp 2006: *A Comparison of 13 Tokenizers on MEDLINE*. Technical Report. doi: 10.1.1.216.2433.

Henningsen, Lena 2010: *Copyright Matters: Imitation, Creativity and Authenticity in Contemporary Chinese Literature*. Berlin: BWV.

Ho, Hou-Ieong Brent und Hilde DeWeerdt. 2014–: *MARKUS. Text Analysis and Reading Platform*. url: https://dh.chinese-empires.eu/beta/.

Hoffmann, Sebastian 2004: „Using the OED quotations database as a corpus – a linguistic appraisal". In: *ICAME* 28, S. 17–30.

Holman, Eric W. et al. 2011: „Automated Dating of the World's Language Families Based on Lexical Similarity". In: *Current Anthropology* 52.6, S. 1–35. doi: 10.1086/662127.

Holmstedt, Robert D. 2009: *Dating the Language of Ruth: A Study in Method*. Paper presented at the annual meeting of the Canadian Society of Biblical Studies (Ottawa, May 23, 2009). url: http://individual.utoronto.ca/holmstedt/Holmstedt_DatingLangRuth_CSBSrevAug2009.pdf (besucht am 02. 08. 2021).

Hong Ye 洪業 (William Hung), Hrsg. 1966 [1949]: *Combined indices to Hou Han shu and the notes of Liu Chao and Li Hsien (後漢書及注釋綜合引得)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 41. Taipei 台北 [Beijing 北京]: Harvard-Yenching [Yanjing xue she 燕京學社].

——— Hrsg. 1955 [1947]: *Combined indices to Shih chi and the notes of P'ei Yin, Ssu-ma Cheng, Chang Shou-chieh, and Takigawa Kametaro (Shi ji ji zhu shi zong he yin de 史記及注釋綜合引得)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 40. Cambridge, MA [Beijing 北京]: Harvard University Press [Yenching University Press].

Hong Ye 洪業 (William Hung) et al., Hrsg. 1934: *A concordance to Shih ching (Mao shi yin de 毛詩引得)*. Harvard-Yenching Institute Sinological Index Series Supplement (Hafo Yanjing xue she

yin de te 哈佛燕京大學圖書館引得特刊) 9. Beijing 北京: Harvard-Yenching (Hafo Yanjing xueshe 哈佛燕京學社).

Hong Ye 洪業 (William Hung) et al., Hrsg. 1938: *Combined Indices to San Kuo Chih and the Notes of P'ei Sung-chih (三國志及裴注綜合引得)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 33. Beijing 北京: Harvard-Yenching (Hafo Yanjing xueshe 哈佛燕京學社).

——— Hrsg. 1966 [1940]: *Combined indices to Han Shu and the notes of Yen Shih-ku and Wang Hsien-ch'ien (Hanshu ji buzhu zonghe yinde 漢書及補註綜合引得)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京學社引得) 36. Taipei 台北 [Beijing 北京]: Harvard Yenching Institute [Yanjing da xue tu shu guan 燕京大學圖書館].

Hsia Chih-tsing 夏志清 1968: *The Classic Chinese Novel: A Critical Introduction*. New York und London: Columbia University Press.

Hsieh Feng-fan 謝豐帆 2015: „Transcribing Foreign Names". In: *Encyclopedia of Chinese Language and Linguistics Online*. Leiden: Brill.

Hu Shaowen 胡绍文 2002: „The Shortages of Hanyu Da Cidian (汉语大词典) From the View of Yi Jian Zhi (夷坚志) – 从《夷坚志》看《汉语大词典》的若干阙失". In: *Research In Ancient Chinese Language* 古汉语研究 4, S. 87–89.

Hu Xianfeng, Wang Yang und Wu Qiang 2014: „Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber". In: *Advances in Adaptive Data Analysis* 6. DOI: 10.1142/S1793536914500125.

Hu Xiaoling, Nigel Williamson und Jamie McLaughlin 2004: *Sheffield Corpus of Chinese*. DOI: 10.1093/llc/fqi034. URL: http://purl.ox.ac.uk/ota/2481 (besucht am 09. 02. 2019).

——— 2005: „Sheffield Corpus of Chinese for Diachronic Linguistic Study". In: *Literary and Linguistic Computing* 20.3, S. 281–293. DOI: 10.1093/llc/fqi034.

Huang Chu-ren 黃居仁, Tokunaga Takenobu 徳永健伸 und Sophia Yat Mei Lee 2006: „Asian language processing: current state-of-the-art". In: *Language Resources & Evaluation* 30, S. 203–218.

Huang Chu-ren 黃居仁 und Xue Nianwen 2019: „Digital Language Resources and NLP tools". In: *The Routledge Handbook of Chinese Applied Linguistics*. Hrsg. von Huang Chu-ren 黃居仁, Zhuo Jing-Schmidt und Barbara Meisterernst. Abingdon, Oxon & New York: Routledge, S. 461–482.

Huang Chu-ren 黃居仁 et. al. 1990: *Academia Sinica Ancient Chinese Corpus* 中央研究院古漢語標記語料庫. URL: http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh (besucht am 10. 02. 2019).

——— 2001: *Academia Sinica Tagged Corpus of Early Mandarin Chinese* 中央研究院近代漢語語料庫. URL: http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/pkiwi/kiwi.sh (besucht am 17. 02. 2019).

Huang Liang et al. 2002a: „PCFG Parsing for Restricted Classical Chinese Texts". In: *COLING-02: The First SIGHAN Workshop on Chinese Language Processing*. ACL Anthology. DOI: 10.3115/1118824.1118830.

——— 2002b: „Statistical Part-of-Speech Tagging for Classical Chinese“. In: *Text, Speech and Dialogue: 5th International Conference, TSD 2002, Brno, Czech Republic September 9-12, 2002*. Hrsg. von Petr Sojka, Ivan Kopecek und Karel Pala. Berlin & Heidelberg: Springer, S. 115–122.

Huang Shigong 黃石公 ca. 100–9 v. Chr. *San lüe* 三略. Hrsg. von Donald Sturgeon. ctext.org.

Hucker, Charles O. 1987 [1985]: *A Dictionary of Official Titles in Imperial China*. Taipei 台北 [Stanford]: Nantian shuju 南天書局 [Stanford University Press].

Hugging Face 🤗 2020–: *Hugging Face Models*. Website. URL: https://huggingface.co/models (besucht am 15. 07. 2021).

Hulsewé, Anthony François Paulus 1993a: „Han shu 漢書“. In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 129–136.

——— 1993b: „Shih chi 史記“. In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS.

Hunter, Michael und Martin Kern, Hrsg. 2018: *Confucius and the Analects Revisited: New Perspectives on Composition, Dating, and Authorship*. Leiden & Boston: Brill.

Hurvitz, Avi 1973: „Linguistic Criteria for Dating Problematic Biblical Texts“. In: *Hebrew Abstracts* 14, S. 74–79.

J

Jaccard, Paul 1902: „Lois de distribution florale dans la zone alpine“. In: *Bulletin de la Sociéé Vaudoise des Sciences Naturelles* 38.144, S. 69–130. DOI: 10.5169/seals-266762.

Jannidis, Fotis 2017a: „Grundbegriffe des Programmierens“. In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 68–95.

——— 2017b: „Zahlen und Zeichen“. In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 59–67.

Jannidis, Fotis, Hubertus Kohle und Malte Rehbein, Hrsg. 2017: *Digital Humanities – Eine Einführung*. Stuttgart: Metzler.

Jensen, Chris 2018: *Why we need an Open Source License that considers the misuse of our code*. URL: https://hackernoon.com/why-we-need-an-open-source-licence-that-considers-the-misuse-of-our-code-8d19b65d425 (besucht am 27. 09. 2018).

Jespersen, Otto 1912 [1905]: *Growth and Structure of the English Language*. Leipzig: B. G. Teubner.

Ji Meng 2010: „A corpus-based study of lexical periodization in historical Chinese“. In: *Literary and Linguistic Computing* 25.2, S. 199–213. DOI: 10.1093/llc/fqq002.

Jiang Shaoyu 蒋绍愚 2015: *Hanyu lishi cihui xue gaiyao* 汉语历史词汇学概要 *(Outline of the History of Chinese Lexicology)*. Beijing 北京: Shangwu yinshuguan 商务印书馆 (The Commercial Press).

Jiaoyubu 教育部 (Bildungsministerium [der Republik China]) 2017: *Yitizi biao* 異體字表 *(Variantenzeichentabelle)*. Yitizi zidian 異體字字典 (Variantenzeichenwörterbuch). URL: https://dict.variants.moe.edu.tw/variants/rbt/variant_modified_record_tiles.rbt (besucht am 14. 07. 2021).

Jing-Schmidt, Zhuo und Shu-Kai Hsieh 2019: „Chinese neologisms". In: *The Routledge Handbook of Chinese Applied Linguistics*. Hrsg. von Huang Chu-ren 黃居仁, Zhuo Jing-Schmidt und Barbara Meisterernst. Abingdon, Oxon & New York: Routledge, S. 514–534.

Jockers, Matthew L. 2013: *Macroanalysis: Digital Methods and Literary History*. Topics in The Digital Humanities. Urbana, Chicago und Springfield: University of Illinois Press.

Jüngling, Ralf und Gabriel Altmann 2003: „Python for linguistics?" In: *Glottometrics* 6, S. 70–82.

K

Kanhabua, Nattiya und Kjetil Nørvåg 2008: „Improving Temporal Language Models for Determining Time of Non-timestamped Documents". In: *Research and Advanced Technology for Digital Libraries: 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Proceedings*. Hrsg. von Birte Christensen-Dalsgaard et al. Berlin & Heidelberg: Springer, S. 358–370. doi: 10.1007/978-3-540-87599-4_37.

Kaske, Elisabeth 2007: *The Politics of Language in Chinese Education, 1895–1919*. Leiden: Brill.

Kauermann, Göran und Helmut Küchenhoff 2010: *Stichproben – Methoden und praktische Umsetzung mit R*. Berlin & Heidelberg: Springer. doi: 10.1007/978-3-642-12318-4.

Keller, Rudi 2003: *Sprachwandel*. 3. Aufl. Tübingen & Basel: A. Francke.

Kennedy, George A. 1951: „The Monosyllabic Myth". In: *Journal of the American Oriental Society* 71.3, S. 161–166. doi: 10.2307/595185 .

——— 1964 [1955]: „The Butterfly Case, Part I". In: *Selected Works of George A. Kennedy*. Hrsg. von Li Tien-yi. New Haven: Far Eastern Publications, S. 274–322.

Kern, Martin 2004: „Die Anfänge der chinesischen Literatur". In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler.

——— 2018: „Kongzi as Author in the Han". In: *Confucius and the Analects Revisited: New Perspectives on Composition, Dating, and Authorship*. Hrsg. von Michael Hunter und Martin Kern. Leiden & Boston: Brill, S. 268–307.

Kilgarriff, Adam et al. 2004: „The Sketch Engine". In: *Proceedings of the 11th EURALEX International Congress*. Hrsg. von Geoffrey Williams und Sandra Vessier. Lorient, France: Université des lettres et des sciences humaines, S. 105–115.

Klaussner, Carmen und Carl Vogel 2015: „Stylochronometry: Timeline Prediction in Stylometric Analysis. Proceedings of AI-2015, The Thirty-Fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence". In: *Research and Development in Intelligent Systems XXXII*. Hrsg. von Max Bramer und Miltos Petridis. Cham & Heidelberg: Springer, S. 91–106. doi: 10.1007/978-3-319-25032-8_6.

——— 2018: „A Diachronic Corpus for Literary Style Analysis". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7–12, 2018*. European Language Resources Association (ELRA), S. 3496–3503.

Klinke, Harald 2017: „Information Retrieval". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 268–278.

Klöter, Henning 2013: „Chinese lexicography". In: *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Hrsg. von Rufus H. Gouws et al. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: DeGruyter Mouton, S. 884–893.

Knipf-Komlósi, Elisabeth, Roberta V. Rada und Csilla Bernáth 2006: *Aspekte des Deutschen Wortschatzes*. Budapest: Bölcsész Konzorcium.

Köhler, Reinhard 1986: *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Quantitative linguistics 31. Bochum: Dr. N. Brockmeyer.

Kotsakos, Dimitrios et al. 2014: „A Burstiness-aware Approach for Document Dating". In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14. Gold Coast, Queensland, Australia: ACM, S. 1003–1006. doi: 10.1145/2600428.2609495.

Kraaij, Wessel 2004: *Variations on Language Modeling for Information Retrieval (Diss.)* Enschede: Neslia Paniculata.

Kristeva, Julia 1972 [1965]: „Bachtin, das Wort, der Dialog und der Roman". In: *Zur linguistischen Basis der Literaturwissenschaft II*. Hrsg. von Jens Ihwe. Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven Bd. 3. Frankfurt am Main: Athenäum, S. 345–375.

Kroch, Anthony S. 1989: „Reflexes of grammar in patterns of language change". In: *Language Variation and Change* 1, S. 199–244. doi: 10.1017/s0954394500000168.

Kulkarni, Vivek et al. 2018: „Simple Neologism Based Domain Independent Models to Predict Year of Authorship". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, S. 202–212. url: https://www.aclweb.org/anthology/C18-1017.

Kullback, Solomon und Richard A. Leibler 1951: „On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, S. 79–86. doi: 10.1214/aoms/1177729694.

Kumar, Abhimanu 2013: „Supervised Language Models for Temporal Resolution of Text in Absence of Explicit Temporal Cues". Diss. Austin: University of Texas.

Kumar, Abhimanu et al. 2012: „Dating Texts without Explicit Temporal Cues". In: *arXiv [cs. CL]* 1211.2290, S. 1–12.

Kupfer, Peter 2009: „Language". In: *Brill's Encyclopedia of China*. Hrsg. von Daniel Leese. Leiden & Boston: Brill, S. 544–549.

L

Lackner, Michael, Iwo Amelung und Joachim Kurtz 2001: „Introduction". In: *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*. Hrsg. von Michael Lackner, Iwo Amelung und Joachim Kurtz. Leiden: Brill.

Langacker, Ronald W. 1977: „Syntactic reanalysis". In: *Mechanisms of Syntactic change*. Hrsg. von Charles N. Li. Austin: University of Texas Press, S. 57–139.

Laozi 老子 2009: *Lau-zi dao de jing* 老子《道德經》. eBook. url: http://www.gutenberg.org/ebooks/7337 (besucht am 19. 05. 2019).

Lappas, Theodoros et al. 2009: „On Burstiness-Aware Search for Document Sequences“. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. Paris, France: Association for Computing Machinery, S. 477–486. doi: 10.1145/1557019.1557075.

Lass, Roger 1980: *On Explaining Language Change*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

——— 1984: *Language and Time: A Historian's View*. Inaugural lectures, University of Cape Town 90. Cape Town: University of Cape Town.

——— 1997: *Historical Linguistics and Language Change*. Cambridge Studies in Linguistics. Cambridge & New York: Cambridge University Press.

——— 2014: „Lineage and the Constructive Imagination: The Birth of Historical Linguistics“. In: *The Routledge Handbook of Historical Linguistics*. Hrsg. von Claire Bowern und Bethwyn Evans. London & New York: Routledge, S. 45–63.

Lee, Thomas H. C. 2000: *Education in Traditional China: A History*. Hrsg. von Erik Zürcher, Stephen F. Teiser und Martin Kern. Handbook of Oriental Studies, Section Four: China 13. Leiden: Brill.

Lees, Robert B. 1953: „The Basis of Glottochronology“. In: *Language* 29.2, S. 113–127. doi: 10.2307/410164.

Legge, James 1882: *The Yî King*. Übers. von James Legge. Sacred Books of the East XVI. Oxford: Clarendon Press.

Lehtinen, Jyri 2009: „Language change as an evolutionary process“. Masterarbeit. Helsinki: University of Helsinki.

Leibniz-Institut für Deutsche Sprache (IDS), Hrsg. 2020: *Neuer Wortschatz rund um die Coronapandemie*. Online-Neologismenwörterbuch OWID. Mannheim. url: https://www.owid.de/docs/neo/listen/corona.jsp (besucht am 09. 11. 2022).

Leopold, Edda 2005: „Das Piotrowski-Gesetz“. In: *Quantitative Linguistik – Quantitative Linguistics – An International Handbook / Ein internationales Handbuch*. Hrsg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. Berlin & New York: Walter de Gruyter, S. 627–633.

Li Hongbing 李泓冰 1994: „龙飞在天 —— 《汉语大词典》编纂前前后后 (Der Drache fliegt – Die ganze Geschichte hinter der Kompilation des HYDCD)“. In: *Renmin ribao* 人民日报 05.11.

Li Peng-Hsuan 李朋軒 und Ma Wei-Yun 馬偉雲 2019–: *CKIP Tagger*. GitHub Repository. url: https://github.com/ckiplab/ckiptagger (besucht am 30. 05. 2021).

Li Rui 李鋭 und Li Yingnan 黎应南 2000 [1823]: *Kaifang shuo* 開方說. Online-Datenbank Diaolong 雕龍 / *Xuxiu Siku quan shu* 續修四庫全書, via CrossAsia. Nagoya 名古屋 & Taipeh 台北: Kaixi MS 日本凱希多媒體 & tts 大鐸資訊.

Li Shen 李申 und Wang Benling 王本靈 2015: *Hanyu da cidian yanjiu* 《汉语大词典》研究 *(A study on Hanyu da cidian)*. Beijing 北京: Shangwu yinshuguan 商務印書館 (The Commercial Press).

Li Wai-Yee 李惠儀 2017: „Concepts of Authorship“. In: *Oxford Handbook of Classical Chinese Literature (1000 BCE–900CE)*. Hrsg. von Li Wai-Yee 李惠仪 Wiebke Denecke und Tian Xiaofei 田曉菲. New York: Oxford University Press, S. 360–376.

Li Xiaonan et al. 2020: „FLAT: Chinese NER Using Flat-Lattice Transformer“. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, S. 6836–6842.

Li Yuanpeng et al. 2015: „Publication Date Estimation for Printed Historical Documents Using Convolutional Neural Networks“. In: *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*. HIP '15. Gammarth, Tunisia: ACM, S. 99–106. DOI: 10.1145/2809544.2809550.

Lin, Trey 2013: *Paoding Analysis*. GitHub Repository. URL: https://github.com/cslinmiso/paoding-analysis (besucht am 26. 02. 2019).

Lin Liangyi 2012: *ik-analyzer IK-Analyzer java* 开源中文分词器. URL: https://code.google.com/p/ik-analyzer/ (besucht am 13. 01. 2016).

List, Johann-Mattis 2013: „Theoretische und praktische Aspekte der quantitativen historischen Linguistik“. Seminarskript, Universität Marburg.

——— 2018: *SinoPy: Python Library for quantitative tasks in Chinese historical linguistics*. Jena. URL: https://pypi.org/project/sinopy/ (besucht am 26. 04. 2020).

Lit, Cornelis van et al., Hrsg. 2015–: *Digital Orientalist, The*. URL: https://digitalorientalist.com/ (besucht am 29. 05. 2021).

Liu Bing 劉冰 2009: „《汉语大词典》书证迟后例补——以《先秦漢魏晉南北朝詩(梁詩)》为例 (Ergänzungen für späte Belegstellen im *HYDCD* - anhand von Gedichten der Prä-Qin, Han, Wei, Jin und Nanbei-Zeit [Liang Gedichte])“. In: *Yuwen Xuekan* 语文学刊 *(Journal of language and literature studies)* 19, S. 72–73.

Liu Xiaogan 劉笑敢 2005: *Laozi niandai xin kao yu sixiang xin quan* 老子年代新考與思想新詮 *[Neue Untersuchungen über Laozis Zeit und neue Interpretationen seiner Philosophie]*. Taipei 台北: Dongda tushu 東大圖書.

Lochbaum, Karen E. und Lynn A. Streeter 1989: „Comparing and Combining the Effectiveness of Latent Semantic Indexing and the Ordinary Vector Space Model for Information Retrieval“. In: *Information Processing & Management* 25.6, S. 665–676.

Loewe, Michael, Hrsg. 1993: *Early Chinese Texts: A Bibliographical Guide*. Berkeley: The Society for the Study of Early China; The Institute of East Asian Studies.

Lu Qin 2019: „Computers and Chinese writing systems“. In: *The Routledge Handbook of Chinese Applied Linguistics*. Hrsg. von Huang Chu-ren 黃居仁, Zhuo Jing-Schmidt und Barbara Meisterernst. Abingdon, Oxon & New York: Routledge, S. 461–482.

Lü Shuxiang 呂叔湘 1963: „Xiandai Hanyu danshuang yinjie wenti chutan 现代汉语单双音节问题初探 (Vorläufige Studie zum Problem von Mono- und Disyllabizität im modernen Chinesischen)“. In: *Zhongguo yuwen* 中国语文 1, S. 10–22.

Lüdi Kong, Eva 2018: „“随文入观”: 古文的阅读、理解与翻译 (Hinein in den Text: Lesen, Verstehen und übersetzen klassischer Chinesischer Texte)“. Konferenzbeitrag vom 15. Dezember 2018 im Rahmen des *International Symposium on the Teaching of Classical Chinese* in Bonn.

Luo Guanzhong 羅貫中 2007 [?1522]: *Sanguo zhi yanyi* 三國志演義 *(Romance of the Three Kingdoms)*. Project Gutenberg eBook. URL: https://www.gutenberg.org/ebooks/23950 (besucht am 28. 05. 2021).

Luo Zhufeng 羅竹風, Hrsg. 2005: *Hanyu da cidian* 漢語大詞典 *UTF-8 (Großes Wörterbuch der chinesischen Sprache, Unicode-Version)*. Shanghai 上海. URL: http://bbs.gxsd.com.cn/forum.php?mod=viewthread&tid=498015 (besucht am 13. 01. 2013).

——— Hrsg. 1986–1994: *Hanyu da cidian* 漢語大詞典 *(Großes Wörterbuch der chinesischen Sprache)*. Bd. 1-13. Shanghai 上海: Cishu chubanshe 辭書出版社.

M

Ma Wei-Yun 馬偉雲 und Chen Keh-Jiann 陳克健 2003: „Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff". In: *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, S. 168–171.

Mair, Victor H., Hrsg. 2003: *An Alphabetical Index to the Hanyu da cidian*. Honolulu: University of Hawai'i Press.

Mangrum, Benjamin 2018: „Aggregation, Public Criticism, and the History of Reading Big Data". In: *PMLA* 133.5, S. 1207–1224. DOI: 10.1632/pmla.2018.133.5.1207.

Mani, Inderjeet und George Wilson 2000: „Robust Temporal Processing of News". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. ACL '00. Hong Kong: Association for Computational Linguistics, S. 69–76. DOI: 10.3115/1075218.1075228. URL: https://doi.org/10.3115/1075218.1075228.

Manning, Christopher D, Prabhakar Raghavan und Hinrich Schütze 2008: *Introduction to Information Retrieval*. Cambridge & New York: Cambridge University Press.

Masini, Federico 2015: „Modern Lexicon, Formation". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Maspero, Henri 1981 [1950, 1971]: *Taoism and Chinese Religion [Le Taoïsme et les religions chinoises]*. Übers. von Frank A. Kierman jr. Amherst [Paris]: University of Massachusetts Press [Gallimard].

McKinney, Wes 2010: „Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Hrsg. von Stéfan van der Walt und Jarrod Millman, S. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

Meng Yuxian et al. 2019: „Is Word Segmentation Necessary for Deep Learning of Chinese Representations?" In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Hrsg. von Anna Korhonen, David R. Traum und Lluís Màrquez. Association for Computational Linguistics, S. 3242–3252. DOI: 10.18653/v1/p19-1314.

Mengzi 孟子 1990: „Mengzi 孟子". In: Academia Sinica 中央研究院. Kap. 1. URL: http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh (besucht am 10. 02. 2019).

Michel, Jean-Baptiste et al. 2011: „Quantitative Analysis of Culture Using Millions of Digitized Books". In: *Science* 331.6014, S. 176–182. DOI: 10.1126/science.1199644.

Morohashi Tetsuji 諸橋轍次, Hrsg. 1955–1960: *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)*. Bd. 1–12. Tokyo 東京: Taishukan shoten 大修館書店.

Morohashi Tetsuji 諸橋轍次 (Komp.), Kamata Tadashi 鎌田正 und Yoneyama Torataro 米山寅太郎 (Rev.), Hrsg. 2021 [1955, 1990, 2000]: *Dai Kan-Wa jiten* 大漢和辞典 *Web* 版 *(Großes*

Chinesisch-Japanisches Wörterbuch, Online-Ausgabe). Bd. 1–13; 15. Tokyo 東京: Taishukan shoten 大修館書店, JapanKnowledge.

Moss, Stephen 2011: *Francis Fukuyama: 'Americans are not very good at nation-building'*. URL: https://www.theguardian.com/books/2011/may/23/francis-fukuyama-americans-not-good-nation-building (besucht am 15. 12. 2021).

Mosteller, Frederick und David L. Wallace 1984 [1964]: *Applied Bayesian and Classical Inference – The Case of The Federalist Papers*. 2. Aufl. New York: Springer.

Mu Yang 慕楊 2020: *CKIP BERT Base Chinese*. BERT Modell. URL: https://huggingface.co/ckiplab/bert-base-chinese (besucht am 13. 10. 2021).

Mu Yang 慕楊 und Ma Wei-Yun 馬偉雲 2020–: *CKIP Transformers*. GitHub Repository. URL: https://github.com/ckiplab/ckip-transformers (besucht am 30. 05. 2021).

Murphy, Andrew 2003: *Shakespeare in Print: A History and Chronology of Shakespeare Publishing*. Cambridge: Cambridge University Press.

Murray, James A. H. et al., Hrsg. 1913–1933: *Oxford English Dictionary*. Bd. 1–13. London: Oxford University Press.

Murray, Katherine M. Elisabeth 1977: *Caught in the Web of Words: James Murray and the Oxford English Dictionary*. New Haven & London: Yale University Press.

N

Neidorf, Leonard 2014: „Lexical Evidence for the Relative Chronology of Old English Poetry". In: *SELIM* 20, S. 7–48.

Nespeca-Moser, Roberto 2005: „Auf dem Weg zu einem Lexikon, Tagger und Parser für das Antikchinesische". Lizenziatsarbeit. Zürich: Universität Zürich.

Nichols, Ryan et al. 2018: „Modeling the Contested Relationship between *Analects*, *Mencius*, and *Xunzi*: Preliminary Evidence from a Machine-Learning Approach". In: *Journal of Asian Studies* 77.1, S. 19–57.

Nielbo, Kristoffer, Ryan Nichols und Edward Slingerland 2018: „Mining the Past – Data-Intensive Knowledge Discovery in the Study of Historical Textual Traditions". In: *Journal of Cognitive Historiography* 3.1–2, S. 93–118. DOI: 10.1558/jch.31662.

Norman, Jerry 1988: *Chinese*. Cambridge: Cambridge University Press.

Nunberg, Geoffrey 2009: „Google's Book Search: A Disaster for Scholars". In: *The Chronicle*. URL: https://www.chronicle.com/article/googles-book-search-a-disaster-for-scholars/ (besucht am 31. 08. 2009).

Nylan, Michael 2001: *The Five 'Confucian' Classics*. New Haven: Yale University Press.

O

OCLC 2019: *oclc.org – Worldcat Identities*. Website. URL: https://www.worldcat.org/identities (besucht am 19. 05. 2019).

Osgood, Charles E. und Thomas A. Sebeok 1954: *Psycholinguistics: A Survey of Theory and Research Problems*. Baltimore: Waverly Press.

Osman, Nabil 1988 [1971]: *Kleines Lexikon untergegangener Wörter – Wortuntergang seit dem Ende des 18. Jahrhunderts*. 5. Aufl. München: C. H. Beck.

Ouyang Xun 歐陽詢 et. al. 0624: *Yiwen leiju* 藝文類聚. Hrsg. von Donald Sturgeon. ctext.org.

P

Palmer, Martha et al. 2007: *Chinese Treebank 6.0*. URL: https://catalog.ldc.upenn.edu/LDC2007T36.

Pan Yunzhong 潘云中 1989: *Hanyu cihui shi gaiyao* 汉语词汇史概要 (*Zusammenfassende Geschichte des Wortschatzes des Chinesischen*). Shanghai 上海: Guji chubanshe 古籍出版社.

Pedregosa, Fabian et al. 2011: „Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, S. 2825–2830.

Peyraube, Alain 2004: „Ancient Chinese". In: *The Cambridge Encyclopedia of the World's Ancient Languages*. Hrsg. von Roger D. Woodard. Cambridge University Press, S. 988–1014.

——— 2015: „Periodization". In: *Encyclopedia of Chinese Language and Linguistics Online*. Leiden: Brill.

Project Gutenberg 1971–. URL: https://www.gutenberg.org/ (besucht am 07. 12. 2021).

Python Software Foundation 2017: *Python 2.7.14 documentation*. URL: https://docs.python.org/2/ (besucht am 26. 09. 2018).

Q

Qiu Xigui 裘錫圭 2000: *Chinese Writing. übersetzt von Gilbert L. Mattos und Jerry Norman*. Berkeley: The Society for the Study of Early China; The Institute of East Asian Studies.

R

Rama, Taraka 2014: *Vocabulary lists in computational historical linguistics*. Data linguistica 25. Göteborg: Språkbanken, Department of Swedish.

——— 2015: *Studies in computational historical linguistics*. Hrsg. von Lars Borin. Data linguistica 27. Göteborg: Språkbanken, Department of Swedish.

Rehbein, Malte und Christian Thies 2017: „Ethik". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 353–357.

Richardson, Charles 1836: *A New Dictionary of the English Language*. London: W. Pickering.

——— 1851: *A New Dictionary of the English Language. 2 vols.* Philadelphia: E. H. Butler & Co.

Richardson, Leonard 1996–2020: *BeautifulSoup*. Python module. URL: https://www.crummy.com/software/BeautifulSoup/ (besucht am 02. 02. 2020).

Rimal, Raju 2014: *Diagnostic Plots using ggplot2*. Website. URL: https://rpubs.com/therimalaya/43190 (besucht am 07. 11. 2018).

Ritz, Christian 2016: *drc Analysis of Dose-Response Curves, Version 3.0-1. R* package. URL: rdocumentation.org/packages/drc/versions/3.0-1 (besucht am 10. 02. 2021).

Rogers, Henry 2005: *Writing systems: a linguistic approach*. Blackwell textbooks in linguistics 18. Malden, MA & Oxford: Blackwell.

Rong Xinjiang 榮新江 2004: „Land Route or Sea Route? Commentary on the Study of the Paths of Transmission and Areas in which Buddhism was Disseminated during the Han Period". In: *Sino-Platonic Papers* 144.

Roten, Thomas 2017: *zhon - Constants used in Chinese text processing*. URL: https://github.com/tsroten/zhon (besucht am 10. 11. 2017).

Rusk, Bruce 2006: „Not Written in Stone: Ming Readers of the *Great Learning* and the Impact of Forgery". In: *Harvard Journal of Asiatic Studies* 66.1, S. 189–231. DOI: 10.2307/25066803.

S

Sagart, Laurent et al. 2019: „Dated language phylogenies shed light on the ancestry of Sino-Tibetan". In: *Proceedings of the National Academy of Sciences* 116.21, S. 10317–10322. DOI: 10.1073/pnas.1817972116.

Sasaki Yutaka 佐々木裕 2007: „The Truth of the F-measure". In: *Toyota Technological Institute (Toyota Kōgyō Daigaku 豊田工業大学)*. URL: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf (besucht am 04. 11. 2022).

Schaaf, Herman 2017: *Mafan - Toolkit for working with Chinese in Python*. Python module. URL: https://pypi.org/project/mafan/ (besucht am 26. 04. 2020).

Schalmey, Peter 1977: *Die Bewährung psychoanalytischer Hypothesen*. Wissenschaftstheorie und Grundlagenforschung 7. Kronberg: Scriptor.

Schalmey, Tilman 2009: „Überlegungen zur Konzeption eines neuen Lehrbuchs für das Klassische Chinesische". Magisterarbeit. München: Ludwig-Maximilians-Universität.

——— 2020: „Das *Hanyu Da Cidian* 漢語大詞典 als Sprachgedächtnis". In: *Erinnern und Erinnerung, Gedächtnis und Gedenken*. Hrsg. von Maria Khayutina und Sebastian Eicher. Jahrbuch der Deutschen Vereinigung für Chinastudien. Wiesbaden: Harrassowitz, S. 73–90.

——— 2021: „Thoughts on »Reliable« Learner's Vocabularies for Classical and Literary Chinese". In: *Teaching Classical Chinese / Zum Unterricht des Klassischen Chinesischen / Wenyan wen jiaoxue* 文言文教学 *(Proceedings of the International Symposium on the Teaching of Classical Chinese, December 14–16, 2018)*. Hrsg. von Li Wen 李文 und Ralph Kauz. Gossenberg: Ostasien Verlag, S. 251–261.

Schicker, Edwin 2017: *Datenbanken und SQL*. 5. Aufl. Wiesbaden: Springer Vieweg.

Schindelin, Cornelia 2005a: „Die quantitative Erforschung der chinesischen Sprache und Schrift". In: *Quantitative Linguistik – Quantitative Linguistics – An International Handbook / Ein internationales Handbuch*. Hrsg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. Berlin & New York: Walter de Gruyter, S. 947–970.

——— 2005b: „Zur Geschichte quantitativ-linguistischer Forschungen in China". In: *Quantitative Linguistik – Quantitative Linguistics – An International Handbook / Ein internationales Handbuch*. Hrsg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. Berlin & New York: Walter de Gruyter, S. 96–115.

Schneider, Florian, Hrsg. 2014–: *Asiascape: Digital Asia*. URL: https://brill.com/view/journals/dias/dias-overview.xml (besucht am 29. 05. 2021).

Schöch, Christof 2012: „Author or genre? Assessing the quality of cluster analysis graphs in two-dimensional classification problems". In: *The Dragonfly's Gaze: Computational analysis of literary texts*. URL: https://dragonfly.hypotheses.org/148 (besucht am 30. 12. 2018).

Schöch, Christof et al. 2020: „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2020_006.

Schuessler, Axel 2015: „Old Chinese Morphology". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Schwermann, Christian und Raji C. Steineck 2014: „Introduction". In: *That Wonderful Composite Called Author*. Hrsg. von Christian Schwermann und Raji C. Steineck. East Asian Comparative Literature and Culture 4. Leiden: Brill, S. 1–29.

Shalmanese 2008: „Quick easy way to migrate SQLite3 to MySQL?" In: *Stack Overflow*. URL: http://stackoverflow.com/questions/18671/quick-easy-way-to-migrate-sqlite3-to-mysql (besucht am 10. 07. 2016).

Shaughnessy, Edward 1993a: „*I ching* 易經 (*Chou I* 周易)". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS.

——— 1993b: „*Shang shu* 尚書 (*Shu ching* 書經)". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 376–389.

Shen Kuo 沈括 2008 [1088]: *Meng xi bi tan* 夢溪筆談 *(Pinselunterhaltungen am Traumbach)*. Project Gutenberg eBook. URL: http://www.gutenberg.net (besucht am 10. 09. 2018).

——— 1997 [1088]: *Pinselunterhaltungen am Traumbach. übs. von Konrad Herrmann*. München: Diederichs.

Shih Hsiang-lin 施祥林 und David R. Knechtges 2014: „Song Yu 宋玉". In: *Ancient and Early Medieval Chinese Literature. A Reference Guide. Part Two*. Hrsg. von David R. Knechtges und Chang Taiping 張泰平. Handbook of Oriental Studies. Leiden: Brill, S. 1007–1022.

SIGHAN 2005–: *SIGHAN Home Page*. URL: http://sighan.cs.uchicago.edu/ (besucht am 29. 09. 2018).

Sima Qian 司馬遷 2008 [91 v. u. Z.] *Shiji* 史記 *(Records of the Grand Historian)*. Project Gutenberg eBook. URL: https://www.gutenberg.org/ebooks/24226 (besucht am 16. 05. 2021).

——— 1959 [91 v. u. Z.] *Shiji* 史記 *(Records of the Grand Historian)*. Beijing 北京: Zhonghua shuju 中華書局.

Simpson, John A., Hrsg. 2014 [2002]: *Oxford English Dictionary, 2nd Edition [Second edition on CD-ROM Version 3.0]*. Oxford University Press. URL: http://njw.me.uk/oed (besucht am 10. 04. 2014).

Simpson, John A. und Edmund Weiner, Hrsg. 1989: *Oxford English Dictionary*. 2. Aufl. Oxford: Clarendon Press.

Sinclair, Stéfan und Geoffrey Rockwell 2016: *Voyant Tools*. URL: https://voyant-tools.org/ (besucht am 13. 04. 2023).

Sindre, Guttorm und Andreas L. Opdahl 2000: „Eliciting Security Requirements by Misuse Cases“. In: *Proceedings of TOOLS Pacific*, S. 120–131.

Soffel, Christian 2004: *Ein Universalgelehrter verarbeitet das Ende seiner Dynastie – Eine Analyse des Kunxue jiwen von Wang Yinglin*. Wiesbaden: Harrassowitz.

——— 2020: „Transcultural Aspects in Chang Yi-Jen's 張以仁 Poetry“. In: *Interface – Journal of European Languages and Literatures* 12.2, S. 113–137. DOI: 10.6667/interface.12.2020.111.

Soothill, William E. und Lewis Hodous 2003 [1937]: *A Dictionary of Chinese Buddhist Terms*. Online Version. URL: http://mahajana.net/texts/soothill-hodous.html (besucht am 28. 11. 2017).

Sproat, Richard W. und Thomas Emerson 2003: „The first international Chinese word segmentation Bakeoff“. In: *Proceedings of the second SIGHAN workshop on Chinese language processing* 17, S. 133–143. DOI: 10.3115/1119250.1119269.

Sproat, Richard W. et al. 1996: „A Stochastic Finite-State Word-Segmentation Algorithm for Chinese“. In: *Computational Linguistics* 22.3, S. 377–404.

Sproat, Richard W. et al. 2001: „Normalization of non-standard words“. In: *Computer Speech & Language* 15.3, S. 287–333. DOI: 10.1006/csla.2001.0169.

Stachowski, Kamil 2020: „Piotrowski-Altmann law: State of the art“. In: *Glottotheory* 11.1, S. 1–12. DOI: 10.1515/glot-2020-2002.

Stamou, Constantina 2007: „Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating“. In: *Literary and Linguistic Computing* 23.2, S. 181–199. DOI: 10.1093/llc/fqm029.

Stanford Natural Language Processing Group 2015: *Stanford Word Segmenter*. URL: http://nlp.stanford.edu/software/segmenter.shtml (besucht am 14. 01. 2016).

——— 2019: *Stanford CoreNLP*. GitHub Repository. URL: https://github.com/stanfordnlp/CoreNLP (besucht am 23. 03. 2019).

Starostin, George 2013: „Lexicostatistics as a basis for language classification: increasing the pros, reducing the cons“. In: *Classification and Evolution in Biology, Linguistics and the History of Science*. Hrsg. von Heiner Fangerau et al. Stuttgart: Franz Steiner Verlag, S. 125–146.

Stefanowitsch, Anatol 2020: *Corpus Linguistics: A Guide to the Methodology*. Textbooks in Language Sciences 7. Berlin: Language Science Press. DOI: 10.5281/zenodo.3735822.

Strötgen, Jannik 2015: „Domain-sensitive Temporal Tagging for Event-centric Information Retrieval“. Diss. Heidelberg: Universität Heidelberg.

Strötgen, Jannik und Michael Gertz 2010: „HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions“. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, S. 321–324. URL: http://www.aclweb.org/anthology/S10-1071.

——— 2016: *Domain-Sensitive Temporal Tagging*. Hrsg. von Graeme Hirst. Synthesis Lectures on Human Language Technologies 36. San Rafael: Morgan & Claypool.

Sturgeon, Donald, Hrsg. 2011: *Chinese Text Project*. URL: https://ctext.org (besucht am 24. 04. 2021).

——— 2019: „Chinese Text Project: a dynamic digital library of premodern Chinese“. In: *Digital Scholarship in the Humanities* 0.0, S. 1–12. DOI: 10.1093/llc/fqz046.

Su, Eugene 2017: *IK Analyzer Solr 5*. URL: https://github.com/EugenePig/ik-analyzer-solr5 (besucht am 07. 01. 2018).

Sun Junyi 2018: *Jieba zhongwen fenci* 结巴中文分词 *(Jieba Chinesisch-Tokenizer)*. GitHub Repository. URL: https://github.com/fxsjy/jieba (besucht am 26. 02. 2019).

Sun Xiaoxuan 孫曉玄 2011: „基于《漢語大詞典》語料庫的宋代新詞研究 (On the New Vocabulary of Song Dynasty Based on Corpus According to <The Great Chinese Dictionary>)“. Diss. Shandong daxue 山東大學.

Suwald, Judith 2008: „Zhong 忠 und das Zhongjing 忠經“. Diss. München: LMU München.

Swadesh, Morris 1955: „Towards Greater Accuracy in Lexicostatistic Dating“. In: *International Journal of American Linguistics* 21.2, S. 121–137. URL: http://www.jstor.org/stable/1263939.

Swan, Russell und David Jensen 2000: „TimeMines: Constructing Timelines with Statistical Models of Word Usage“. In: *Proceedings of KDD-2000 Workshop on Text Mining*.

T

Tahmasebi, Nina, Lars Borin und Adam Jatowt 2019: „Survey of Computational Approaches to Lexical Semantic Change Detection“. In: *arXiv [cs. CL]* arXiv:1811.06278v2, S. 1–55.

Tai, James H.-Y. und Marjorie K. M. Chan 1999: „Some Reflections on the Periodization of the Chinese Language“. In: *Studies in Chinese Historical Syntax and Morphology: Linguistic Essays in Honor of Mei Tsu-lin*. Hrsg. von Alain Peyraube und Sun Chaofen. Paris: École des Hautes Études en Sciences Sociales, S. 223–239.

Tan Pang-Ning 陳封能, Michael Steinbach und Vipin Kumar 2013 [2005]: *Introduction to Data Mining*. Essex: Pearson.

Tang Feng 唐鳳 2009: *Lingua::Sinica::PerlYuYan – Perl in Classical Chinese in Perl – Zhongshuyu* 中書玲. URL: https://metacpan.org/pod/release/AUDREYT/Lingua-Sinica-PerlYuYan-1257340475/lib/Lingua/Sinica/PerlYuYan.pm (besucht am 12. 09. 2020).

Tang Xuri, Qu Weiguang und Chen Xiaohe 2015: „Semantic Change Compuation: A Successive Approach“. In: *World Wide Web* 19.3, S. 375–415. DOI: 10.1007/s11280-014-0316-y.

Tao Hongyin 2015: „Author Identification and Dating of Texts“. In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Teahan, William J. et al. 2000: „A Compression-based Algorithm for Chinese Word Segmentation“. In: *Computational Linguistic* 26.3, S. 375–393. DOI: 10.1162/089120100561746.

TEI Consortium 2019: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Website. URL: http://www.tei-c.org/Guidelines/P5/ (besucht am 07. 05. 2019).

Thaller, Manfred 2017: „Geschichte der Digital Humanities“. In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 3–12.

The pandas development team 2020: *pandas 1.5.0*. DOI: 10.5281/zenodo.3509134.

Thurneysen, Rudolf 1921: *Die irische Helden- und Königssage bis zum siebzehnten Jahrhundert*. 2 Bde. Halle: Max Niemeyer.

Timberg, Evert et al. 2013–: *Chart.js*. GitHub Repository. URL: https://github.com/chartjs/Chart.js (besucht am 12. 10. 2021).

Toner, Gregory und Han Xiwu 2019: *Language and Chronology – Text Dating by Machine Learning*. Language and Computers, Vol. 84. Leiden & Boston: Brill.

Treter, Clemens 2004: „Die Literatur der Ming- und Qing-Zeit". In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 225–287.

Triglia, Scott 2013: *Elegant n-gram generation in Python*. Blog entry. URL: http://locallyoptimal.com/blog/2013/01/20/elegant-n-gram-generation-in-python/ (besucht am 27. 07. 2016).

Tsai Yu-Fang und Chen Keh-Jiann 陳克健 2004: „Reliable and Cost-Effective Pos-Tagging". In: *International Journal of Computational Linguistics & Chinese Language Processing* 9.1, S. 83–96.

Tuldava, Juhan 1998: *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Übers. von Gabriel Altmann und Reinhard Köhler. Trier: Wissenschaftlicher Verlag Trier.

Tung Tso-Pin 董作賓, Hrsg. 1960: *Chronological Tables of Chinese History*. Hong Kong 香港: Hong Kong University Press.

U

Underwood, Ted E. 2016: „The Life Cycles of Genres". In: *Journal of Cultural Analytics* 1.1. DOI: 10.22148/16.005.

Unger, Ulrich 1985a: *Einführung in das Klassische Chinesisch, Teil I: Allgemeines, Chinesische Texte, Indices*. Wiesbaden: Harrassowitz.

——— 1985b: *Einführung in das Klassische Chinesisch, Teil II: Erläuterungen*. Wiesbaden: Harrassowitz.

Unicode, Inc. 2016: *Glossary of Unicode Terms*. URL: http://www.unicode.org/glossary/ (besucht am 30. 11. 2016).

Universität Trier, Computerlinguistik und Digital Humanities: *Concluded: DGfS-CL Fall School 2015*. URL: https://www.uni-trier.de/index.php?id=55983 (besucht am 14. 09. 2021).

Unschuld, Paul Ulrich 1986: *Medicine in China: A History of Pharmaceutics*. Comparative Studies of Health. Berkeley & Los Angeles: University of California Press.

V

Valla, Lorenzo 2007 [1440]: *On the Donation of Constantine*. Übers. von Glen W. Bowersock. The I Tatti Renaissance Library. Cambridge & London: Harvard University Press.

van Hulle, Dries und Mike Kestemont 2016: „Periodizing Samuel Beckett's Works: A Stylochronometric Approach". In: *Style* 50.2, S. 172–202.

Vanderplas, Jake 2018: *Data Science mit Python – Das Handbuch für den Einsatz von IPython, Jupyter, NumPy, Pandas, Matplotlib, Scikit-Learn*. Übers. von Knut Lorenzen. 1. Aufl.

Vierthaler, Paul 2016a: „Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature". In: *Journal of Cultural Analytics*, S. 1–32. DOI: 10.22148/16.003.

——— 2016b: *Late Imperial Chinese Texts: The Corpus for Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature*. DOI: 10.7910/DVN/GDYFAG. URL: https://doi.org/10.7910/DVN/GDYFAG.

——— 2020: „Digital humanities and East Asian studies in 2020". In: *History Compass* e12628. DOI: 10.1111/hic3.12628.

Vogelsang, Kai 2012: *Geschichte Chinas*. Stuttgart: Reclam.

——— 2021: *Introduction to Classical Chinese*. Oxford: Oxford University Press.

W

Wang Fu 王符 ca. 102–167 v. Chr. *Qian Fu Lun* 潛夫論. Hrsg. von Donald Sturgeon. ctext.org.

Wang Li 王力 2011 [1958]: *Hanyu shi gao* 漢語史稿 (*Entwurf einer Geschichte des Chinesischen*). Beijing 北京: Zhonghua shuju 中華書局.

Wang Yankun 王彦坤, Hrsg. 1997: *Lidai bihuizi huidian* 歷代避諱字匯典 *(Geschichtliches Lexikon von Tabuzeichen)*. Zhongzhou guji chubanshe 中州古籍出版社.

Watson, William 1973: „On Some Categories of Archaism in Chinese Bronze". In: *Ars Orientalis* 9, S. 1–13.

Wei Shou 魏收 1974 [554]: *Wei shu* 魏書. *8 Bde*. Beijing 北京: Zhonghua shuju 中華書局.

Wei Zheng 魏徵 1973 [636]: *Sui shu* 隋書. *6 Bde*. Beijing 北京: Zhonghua shuju 中華書局.

Weiji wenku 維基文庫 *(Wikisource)* 2003–. Website. URL: https://www.gutenberg.org/.

Weinreich, Uriel, William Labov und Marvin I. Herzog 1968: „Empirical Foundations for a Theory of Language Change". In: *Directions for Historical Linguistics*. Hrsg. von Winfried P. Lehmann und Yakov Malkiel. Austin: University of Texas Press, S. 95–195.

Wenlin Institute, Inc. 2015: *Wénlín* 文林 *Software for Learning Chinese, Version 4.2.0. macOS* App.

Wenxue 100 文學 *100* 2015–. Website. URL: http://www.wenxue100.com/ (besucht am 07. 12. 2021).

Wickham, Hadley 2016: *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. URL: https://ggplot2.tidyverse.org.

Wickmann, Dieter 1989: „Computergestützte Philologie: Bestimmung der Echtheit und Datierung von Texten / Computer-Aided Philology: Authorship and Chronological Determination". In: *Computational Linguistics, An International Handbook of Computer Oriented Language Research and Applications*. Hrsg. von István S. Bátori, Winfried Lenders und Wolfgang Putschke. Handbücher zur Sprach- und Kommunikationswissenschaft, Band 4. Berlin & New York: De Gruyter, S. 528–534.

Wilhelm, Richard 1982: *Mong Dsï: Die Lehrgespräche des Meisters Meng K'o*. Köln: Eugen Diederichs.

Wilkinson, Endymion 2000: *Chinese History. A Manual*. Revised and Enlarged. Cambridge, MA & London: Harvard University Asia Center, Harvard University Press.

Williams, Hugh E. und David Lane 2004: *Web Database Applications with PHP and MySQL*. 2. Aufl. Sebastopol: O'Reilly.

Willinsky, John 1994: *Empire of Words: The Reign of the OED*. Princeton, New Jersey: Princeton University Press.

Wilson, Thomas 1994: „Confucian Sectarianism and the Compilation of the Ming History". In: *Late Imperial China* 15, S. 53–84. DOI: 10.1353/late.1994.0002.

Winchester, Simon 1998: *The Professor and the Madman: A Tale of Murder, Insanity, and the Making of The Oxford English Dictionary*. New York: HarperCollins.

Winter, Marc 2015: „Kāngxī zìdiǎn 康熙字典". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Wittgenstein, Ludwig 2003 [1922]: *Tractatus logico-philosophicus*. Frankfurt am Main: Suhrkamp.

Wolf, Thomas et al. 2020: „Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, S. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

Wong Kam-Fai 黃錦輝 et al. 2010: *Introduction to Chinese Natural Language Processing*. Hrsg. von Graeme Hirst. Synthesis Lectures on Human Language Technologies. San Rafael: Morgan & Claypool.

X

Xiao Hang 2008: „On the Applicability of Zipf's Law in Chinese Word Frequency Distribution". In: *Journal of Chinese Language and Computing* 18.1, S. 33–46.

Xu Shen 許慎 1985 [121]: *Shuo wen jie zi* 說文解字. *2 Bde.* Beijing 北京: Zhonghua shuju 中華書局.

Xu Wenming 徐文明 1999: „《*Niepan wu ming lun*》*zhen wei bian* 《涅盤無名論》真偽辨 (Die Authentizität des *Niepan wu ming lun*)". In: *Yuanguang foxue xuebao* 圓光佛學學報 *(Yuanguang Buddhist Journal)* 7.

Xu Zhongshu 徐中舒, Hrsg. 1986–1990: *Hanyu da zidian* 漢語大字典 *(Großes Lexikon chinesischer Schriftzeichen)*. 3 Bde. Wuhan 武漢: Sichuan cishu chubanshe 四川辭書出版社, Hubei cishu chubanshe 湖北辭書出版社.

Xue Nianwen et al. 2005: „The Penn Chinese TreeBank: Phrase structure annotation of a large corpus". In: *Natural Language Engineering* 11.2, S. 207–238.

Xue Nianwen et al. 2016: *Chinese Treebank 9.0*. URL: https://catalog.ldc.upenn.edu/LDC2016T13 (besucht am 15. 06. 2016).

Y

Yamada Takahito 山田崇仁 2004: „N-gram moderu o riyōshite senshin bunken no seisho jiki o saguru: ‚Sonshi' jūsanhen o jirei toshite N-gram モデルを利用して先秦文献の成書時期を探る —『孫子』十三篇を事例として —, *n*-Gramm Ansatz zur Datierung von chinesischen-Qin Texten am Beispiel der 13 *Sunzi*-Kapitel)". In: *Tōkyōdaigaku tōyō bunka kenkyūjo fuzoku tōyō-gaku kenkyū jōhō sentā, Ajia kenkyū jōhō* 東京大学東洋文化研究所附属東洋学研究情報センター, アジア研究情報 *(Tokyo University Research & Information Center for Asian Studies, Gateway to Asian Studies in Japan)*.

Yang Lin 杨琳 2011: *"Hanyu da cidian"guangpan ban yu zhizhiban de qubie* 《汉语大词典》光盘版与纸质版的区别 *(Unterschiede zwischen der CD-Rom und der Papierausgabe des Hanyu da cidian)*. URL: http://www.guoxue.com/?p=4453 (besucht am 22. 07. 2018).

Yang Zhiyi 楊治宜 2015: „The Modernity of the Ancient-Style Verse". In: *Frontiers of Literary Studies in China* 9.4, S. 551–580.

Yardi, Madhukar R. 1946: „A Statistical Approach to the Problem of Chronology of Shakespeare's Plays". In: *Sankhyā: The Indian Journal of Statistics* 7.3, S. 263–268.

Yasuoka Kōichi 安岡孝一 2019: „Universal Dependencies Treebank of the Four Books in Classical Chinese". In: *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, S. 20–28.

——— 2019–: *UD-Kanbun*. GitHub Repository. URL: https://github.com/KoichiYasuoka/UD-Kanbun (besucht am 25. 05. 2021).

Ye Guoliang 葉國良 2008: *„Xian Qin lishu zhong baocun de guyu ji qi yiyi* 先秦礼书中保存的古语及其意义 (Archaismen aus vor-Qinzeitlichen Ritenbüchern und ihre Bedeutungen)". In: *Journal of Northwest University (Philosophy and Social Sciences Edition)* 西北大学学报 *(*哲学社会科学版*)* 38.1, S. 86–90.

Yin Wen 尹文: *Yin Wen Zi* 尹文子. Hrsg. von Donald Sturgeon. URL: https://ctext.org/yin-wen-zi (besucht am 17. 02. 2020).

Yong Heming und Peng Jing 2008: *Chinese Lexicography: A History from 1046 BC to AD 1911*. Oxford: Oxford University Press.

Young, Ian und Robert Rezetko 2014: *Linguistic Dating of Biblical Texts*. Bd. 1. Routledge.

Yu Xuejin und Wei Huangfu 2019: „A Machine Learning Model for the Dating of Ancient Chinese Texts". In: *International Conference on Asian Language Processing, IALP 2019, Shanghai, China, November 15-17, 2019*. Hrsg. von Lan Man et al. IEEE, S. 115–120. DOI: 10.1109/IALP48816.2019.9037653.

Yu Zhangrui 余章瑞 1988: „为伊消得人憔悴——记《汉语大词典》的编纂及为其辛勤工作的人们 (Zum Gedenken an Yi Xiao, Erinnerung an die Herausgabe und die Menschen, die hart am *HYDCD* gearbeitet haben)". In: *Renmin ribao* 人民日报 06.23.

Z

Zádrapa, Lukáš 2015: „Word and Wordhood in Classical Chinese". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

Zampieri, Marcos, Shervin Malmasi und Mark Dras 2016: „Modeling Language Change in Historical Corpora: The Case of Portuguese". In: *ArXiv* abs/1610.00030.

Zhai Chengxiang 2008: „Statistical Language Models for Information Retrieval: A Critical Review". In: *Foundations and Trends in Information Retrieval* 2.3, S. 137–213. DOI: 10.1561/1500000008.

Zhang Huaping 張華平 2018: *NLPIR-ICTCLAS* 汉语分词系统 *(NLPIR-ICTCLAS Chinese lexical analysis system)*. Website. URL: http://ictclas.nlpir.org/index.html (besucht am 18. 03. 2019).

Zhang Huaping 張華平 et al. 2003: „HHMM-based Chinese Lexical Analyzer ICTCLAS". In: *Proceedings of the Second Workshop on Chinese Language Processing, SIGHAN 2003, Sapporo, Japan, July 11-12, 2003*. URL: https://aclanthology.info/papers/W03-1730/w03-1730.

Zhang Menghan et al. 2019: „Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic". In: *Nature* 569.7754, S. 112–115. DOI: 10.1038/s41586-019-1153-z.

Zhang Qiyun 張其昀 et. al., Hrsg. 1973–1979: *Zhongwen da cidian* 中文大辭典 *(Großes Wörterbuch des Chinesischen)*. Bd. 1–10. Yangmingshan 陽明山: Zhongguo wenhua xueyuan 中國文化學院.

Zhang Weizhong 張衛中 2016: „新词语与清末民初作家的科幻想象 (Neologism and the Imagination of Science Fiction Writers in the Late Qing Dynasty and the Early Republic of China)". In: *Journal of Central China Normal University (Humanities and Social Sciences)* 华中师范大学学报 *(*人文社会科学版*)* 55.6, S. 103–109.

Zhang Yue und Yang Jie 2018: „Chinese NER Using Lattice LSTM". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne: Association for Computational Linguistics, S. 1554–1564. DOI: 10.18653/v1/P18-1144.

Zhang Zhiyi 张志毅 2010: „, 辞书强国' 究竟有多远 (Wie weit es noch bis zum ‚starken Wörterbuchland' ist)". In: *Renmin ribao* 人民日报 10.12.

Zhao Jindan 趙金丹 2007: „《朱子語類》新詞新語初探 (Erste Untersuchungen über neue Wörter und Phrasen im «Zhuzi yulei»)". Masterarbeit. Shǎnxi 陝西: Shǎnxi shifan daxue 陝西師範大學.

Zhao Lanying 趙蘭英 1986: „《Hanyu da cidian》bianzuan wancheng《汉语大词典》编纂完成 (Die Kompilation des *HYDCD* ist abgeschlossen)". In: *Renmin ribao* 人民日报 01.11.

Zhao Shouhui und Zhang Dongbo 2008: „The Totality of Chinese Characters–A Digital Perspective". In: *Journal of Chinese Language and Computing* 17.2, S. 107–125.

Zheng Xianzhang 郑贤章 2000: „《Hanyu da cidian》shuzheng chushi li shi bu《汉语大词典》书证初始例试补 (Supplementing some Earlier Citations to Hanyu da cidian)". In: *Gu Hanyu yanjiu* 古汉语研究 *(Research in Ancient Chinese Language)* 2, S. 94–96.

Zhonghua renmin gongheguo wenhuabu 中华人民共和国文化部 und Zhongguo wenzi gaige weiyuanhui 中国文字改革委员会, Hrsg. 1988 [1955]: *Diyipi yitizi zhenglibiao* 第一批异体字整理表 *(Erste Tabelle mit Standardformen für Zeichen mit Varianten)*. Beijing 北京: Zhonghua renmin gongheguo wenhuabu 中华人民共和国文化部 und Zhongguo wenzi gaige weiyuanhui 中国文字改革委员会.

Zhongjing 忠經. Unkommentierte Ausgabe auf *WikiSource*. URL: https://zh.wikisource.org/zh-hant/%E5%BF%A0%E7%B6%93 (besucht am 30. 03. 2019).

Zhu Jinxiong 朱錦雄 2013: „*Lun* Ji *Kang ‚Shengxian gaoshi zhuanzan' zhong de ‚gao shi' fanxing* 論嵇康《聖賢高士傳贊》中的「高士」範型 (‚Diskussion des Begriffs *gaoshi* in Ji Kangs *Biographien heiliger gaoshi (etwa: untadeliger Menschen)*')". In: 國立臺北教育大學語文集刊 *Guo li Taibei daxue yuwen jikan (Journal of Language and Literature Studies)* 24, S. 233–260.

Zhu Jinyang und Karl-Heinz Best 1992: „Zum Wort im modernen Chinesisch". In: *Oriens extremus* 35, S. 45–60.

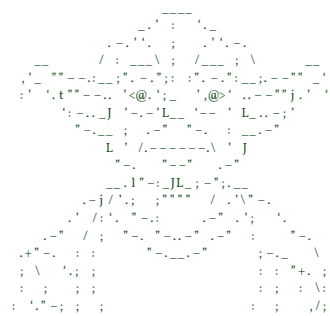
——— 1998: „Wortlääufigkeiten in chinesischen Kurzgeschichten". In: *Asian and African Studies* 7, S. 45–51.

Zhuang Zhou 莊周 o. J. [ca. 3. Jh. v. u. Z.] *Zhuangzi* 莊子. *Digitale Ausgabe. Guoxue jingdian shuku* 國學經典書庫. Dongyang ligong daxue tushuguan 東洋理工大學圖書館.

Zipf, George Kingsley 1947: „Prehistoric 'Cultural Strata' in the Evolution of Germanic: The Case of Gothic". In: *Modern Language Notes* 62.8, S. 522–530.

Zorkina, Mariana 2021: „Defining word boundaries for Modern and Classical Chinese". In: *The Digital Orientalist*. URL: https://digitalorientalist.com/ (besucht am 16. 05. 2021).

Zürcher, Erik 2007 [1959]: *The Buddhist Conquest of China*. 3. Aufl. Sinica Leidensia XI. Leiden: Brill.



Stichwortverzeichnis

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Epilog

„Die Grenzen meiner Sprache
bedeuten die Grenzen meiner Welt."[1]
Ludwig Wittgenstein

Während international in den vergangenen Jahren in Verbindung mit unterschiedlichen Bereichen der Ostasienwissenschaften ein regelrechter *Digital Humanities*-Boom stattgefunden hat, sind im deutschsprachigen Raum interdisziplinäre sinologisch-computerlinguistische Arbeiten selten. Diese Arbeit soll andere ermutigen, sich zur Beantwortung sinologischer oder anderer geisteswissenschaftlicher Fragestellungen über die Limitationen der inzwischen zahlreich verfügbaren Tools hinaus auch selbst Tools und Methoden (weiter) zu entwickeln – und quelloffen zur Verfügung zu stellen.[2] International bilden sich vermehrt Teams, in denen sich z. B. Philolog:innen, Historiker:innen und Informatiker:innen oder (Computer-)linguist:innen zusammenschließen, um gemeinsam an einer geisteswissenschaftlichen Fragestellung zu arbeiten. Diese Form der Kooperation erweist sich als sehr fruchtbar. Die klassische Form der Dissertation, die in den Geisteswisenschaften üblicherweise immer noch als Monographie eine:r einzigen Verfasser:in vorgelegt wird, ist in einem Feld, das sich so rasant weiterentwickelt wie die *Digital Humanities*, aus meiner Sicht leider nur bedingt geeignet.[3]

Als 2015 die Ideen zu den in Kapitel 6.2 und 6.3 vorgestellten Methoden entstanden, hatte ich noch kaum eine Zeile *Python* geschrieben – ich war guter Dinge, mit meinen „alten Bekannten" *PHP*[4] und *mySQL* „durchzukommen", die mich im Bereich der Webentwicklung privat und später auch beruflich treu begleitet hatten. Nach wenigen Monaten musste ich einsehen, dass man zwar auch einem Text vom Umfang des *HYDCD*, in der gedruckten Fassung zwölf große, schwere Bände, 21,7 Kilogramm Papier,[5] insgesamt über 18.000 Seiten „Kleingedrucktes", oder eben fast 140 *MegaByte Plain Text*, mit *PHP* begegnen kann – aber meine Geduld dafür nicht ausreichte. Laufzeiten von mehreren Tagen und im Nachgang oft die ernüchternde Entdeckung von Fehlern, die eine Wiederholung nötig machten, taten ihr Übriges.

1 Ludwig Wittgenstein 2003 [1922]: *Tractatus logico-philosophicus*. Frankfurt am Main: Suhrkamp, S. 86, These 5.6.
2 Dabei muss selbst das Erlernen einer (weiteren oder ersten) Programmiersprache kein Hindernis darstellen: dieser Prozess kann – wie das Erlernen einer neuen, natürlichen Sprache, eine große Bereicherung sein, zu neuen Denkmustern anregen und aus meiner Sicht helfen, Berührungsängste von Philolog:innen mit der Informatik und Mathematik abzubauen.
3 Das zeigt sich unter anderem darin, dass mehr als ein Viertel der verwendeten Literatur nach Beginn dieser Arbeit 2015 veröffentlicht wurde.
4 *PHP* (*Hypertext Preprocessor*) hat sich als Skriptsprache in der Kombination mit *mySQL* als Datenbank-Management zur Entwicklung datenbankgestützter Web-Anwendungen wie Foren und Webshops als ein kostenloser und einfach zu erlernender Standard etabliert. Siehe z. B. Hugh E. Williams und David Lane 2004: *Web Database Applications with PHP and MySQL*. 2. Aufl. Sebastopol: O'Reilly, S. 1.
5 Eigene Messung.

Veröffentlicht in: Tilman Schalmey: *Computerlinguistische Datierung schriftsprachlicher chinesischer Texte*.
Heidelberg: HASP, 2023. DOI: https://doi.org/10.11588/hasp.1153

Inspirationen während der *Fall School* für Computerlinguistik 2015[6] und die Veranstaltungen der von Hilde de Weerth initiierten „Summer School in Chinese Digital Humanities“ in Leiden 2016, insbesondere ein Workshop und die Arbeit von Paul Vierthaler[7] haben mich dazu inspiriert, *Python* zu erlernen und zu verwenden. Aus guten Gründen „gehört Python zu den beliebtesten Programmiersprachen weltweit“[8] und „hat sich in den letzten Jahrzehnten zu einem erstklassigen Tool für wissenschaftliche Berechnungen entwickelt, insbesondere auch für die Analyse und Visualisierung großer Datensätze“,[9] nicht zuletzt für die quantitative Linguistik.[10]

Die Performanceunterschiede bei der Erledigung computerlinguistischer Standardaufgaben im Vergleich zu *PHP* sind groß, so dass sich der Aufwand schnell ausgezahlt hat und das vorliegende Ergebnis sicher weit über das hinausgeht, was mit *PHP* möglich gewesen wäre.

Danksagung

Ein großes Dankeschön sei – *last but not least* – ausgesprochen an Christian Soffel, der als sinologischer Universalgelehrter diese Arbeit über die Grenzen der Sinologie hinweg zu betreuen wusste. An Christof Schöch, der sich recht kurzfristig bereit erklärt hat, die Zweitbegutachtung dieser Arbeit zu übernehmen und wertvolle Hinweise für den Feinschliff gegeben hat.

Roger Arbogast, Maura Dykstra, Irena Georgieva, Sven Naumann, Marc Nürnberger, Lisa Pramme und Grete Schönebeck dafür, dass sie sich viel Zeit genommen haben, meine Arbeit oder Aspekte daraus mit mir zu diskutieren. Allen meinen Kolleg:innen in der Trierer Sinologie, besonders Lydia Wolf, Heribert Lang, Liu Huiru 劉慧儒, Jan Goldenstein, Wu Jing 吴靜 und Chien Juo-ping 簡若玶, die über Jahre hinweg eine entspannte und inspirierende tägliche Routine mit mir pflegten und den vielen weiteren Wegbegleiter:innen, die hier nicht genannt sind. Sonja Scheungrab, die mich in den Jahren des Schaffensprozesses begleitet und unterstützt hat. Meinen Eltern, die immer wieder geholfen haben, Steine vom Weg zu nehmen.

6 Eine zweiwöchige Veranstaltung der Deutschen Gesellschaft für Sprachwissenschaft an der Universität Trier, mit Kursen von Thomas Hanneforth, Gero Kunter, Melanie Siegel und Caroline Sporleder, siehe Universität Trier, Computerlinguistik und Digital Humanities: *Concluded: DGfS-CL Fall School 2015*. URL: https://www.uni-trier.de/index.php?id=55983 (besucht am 14. 09. 2021).

7 Siehe dazu auch Kapitel 4.1, S. 60.

8 Jannidis 2017a, S. 68.

9 Vanderplas 2018, S. 14.

10 Vgl. Jüngling und Altmann 2003, *passim*.

Abstract

The chronological classification of texts can be crucial for clarifying authenticity and interpretation. For several Western languages, statistical language models (*SLMs*), amongst other methods, have been proven useful for automatically assigning timespan (*chronon*) labels to texts. This is made possible by changes in style, grammar, vocabulary, and phonology. When it comes to Classical and Literary Chinese sources, dating can be complicated not only by the existence of forgeries, complex textual lineages, and obscure authorship, but also the fact that many genres attempted to remain faithful to ancient rhetorical and linguistic patterns. Additionally, major phonological changes are not often reflected in Chinese script. The present study assesses both new and established computational dating methods and examines related issues in computational processing of Literary Chinese.

The history of the Chinese written language is the starting point for this study. On the basis of official dynastic histories (*zhengshi* 正史), it is shown that both grammatical and lexical changes can give clues to the time of production, even within a corpus of stylistically homogenous texts. Tied to the appearance of new concepts, lexical innovation is found to be a key indicator of the time of a text's creation.

Readers are then introduced to the current state of research on computational and philological methods of textual dating, emphasizing *SLMs*. The problems of Chinese word segmentation, the lack of diachronic Chinese corpora, as well as recognition of named entities and temporal expressions are discussed. A diachronic Chinese lexeme database for making lexical changes useful for dating is constructed from the earliest word use attestations (*loci classici*) sourced from a digital version of the *Hanyu da cidian* 漢語大詞典.

The main part of the study is dedicated to the development, testing, and comparison of dating methods for Literary Chinese texts. *SLMs* are adapted to be used with Chinese n-gram and plain text corpora. Text neologism profiles, generated from the lexeme database, are introduced as an innovative approach to emphasize lexicalization in automated dating. Accuracy of dating is increased by the usage of dated proper names and temporal expressions. In an attempt to treat time as a continuous variable, the average year of lexicalization (*AYL*) of words in a given text is also tested as a dating indicator.

It is found that *SLMs* can be successfully employed for assigning chronological categories to Literary Chinese texts. However, neologism profiles prove more robust against the rigidness of the written language, require less specific training, and can easily be combined with and aid the work of a philologist. Nevertheless, some Classical texts remain resistant to a linguistic analysis. All three evaluated methods can be tested through a ready-to-use online tool, *VisualTime*, developed by the author as part of this study (https://visualtime.schalmey.de).

論文提要

梳理文本的時間順序對真實性研究和詮釋至關重要。對一些西方語言來說，統計語言模型（*SLMs*）已被證明有助於將文本劃歸到一個歷史節點，即一個時段。這是通過語法、語音和詞匯以及歷時文體風格之變化進行的。中國書面資料因偽造、部分流傳史複雜和作者身份不明等因素，再加上某些文本體裁在語言上頗多取法古代範文，殊難確定產生年代。此外，漢文幾乎反映不出實實在在發生過的語音變化。本論文試圖通過研究比較計算機輔助文本測定的新老方法的應用來探討這一問題。

論文是以古漢語史為出發點的。官方正史研究的結果表明，文本即使具有同質的文體風格，其語法和詞匯的變化也能為確定文本產生的時間提供線索。新術語和新概念的出現為新構詞匯創造了條件，而新構詞匯被證明是文本產生時間的一個關鍵性指標。

本論文側重統計語言模型，對計算機輔助和文字學方法的文本斷代研究現狀作了綜述。接下來討論了一些基本性的挑戰：諸如古漢語文本的分節問題、缺乏歷時漢語語料庫，以及識別命名實體（*NER*）和時間概念的困難。為了使詞匯變化能更適用於斷定時間，還新建了一個歷時漢語詞匯庫。這是以《漢語大詞典》數字版中的詞條例證（*loci classici*）為基礎的。

論文的主要部分致力於開發、測試和比較古漢語文本的年代測定方法，使得統計語言模型適用於中文 ***n*** 元語法和純文本語料庫。作為一種創新探索，我們介紹了用詞庫生成文本的新詞條圖。把有具體時間的命名實體和時間概念納入考量，可以進一步提高確定時間的準確性。為了將時間作為連續變量，文本中單詞詞匯化的平均年份也被作為一個實驗性的斷定年代的指標進行測試。結果發現，統計語言模型可以成功地確定古漢語文本的產生時間。事實證明，新詞條圖更抗拒書面語言的風格僵化，也較少需要專門訓練，此外還可用於協助文字學工作。所評估的三種方法都可以通過在本研究框架下開發的在線工具 *VisualTime*（https://visualtime.schalmey.de）進行測試。

Die chronologische Einordnung von Texten kann für Authentizitätsforschung und Exegese entscheidend sein. Die Datierung schriftsprachlicher chinesischer Quellen kann unter anderem durch Imitation antiker Vorbilder und unklare Urheberschaft erschwert werden. Dieses Buch untersucht erstmals die Entwicklung und Anwendung computergestützter Methoden für die Datierung chinesischsprachiger Quellen. Dabei ermöglicht eine lexembasierte Methode, der stilistischen Rigidität der Schriftsprache zu begegnen und unterstützt damit die philologische Arbeit. Zudem werden der Sprachwandel, die Eignung digitaler Methoden für die Untersuchung klassischer Texte und das *Hanyu da cidian* 漢語大詞典 als wichtige Datenquelle für lexikographische Datierung untersucht.