# 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.

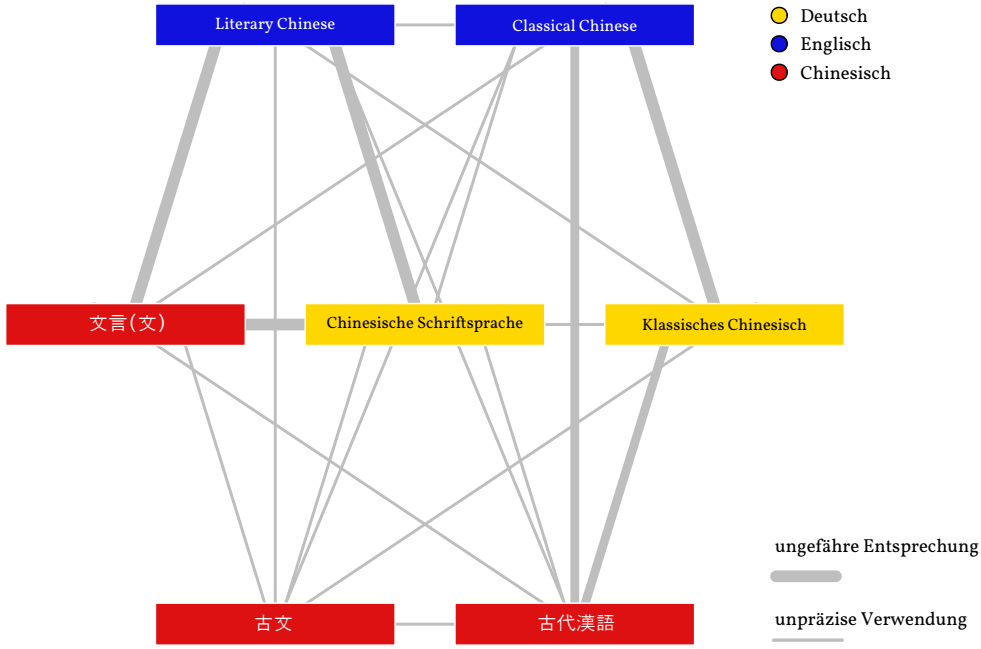


**Abbildung 1.1** Klassisches und schriftsprachliches Chinesisch: Übersetz- und Austauschbarkeit der Begriffe

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.