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

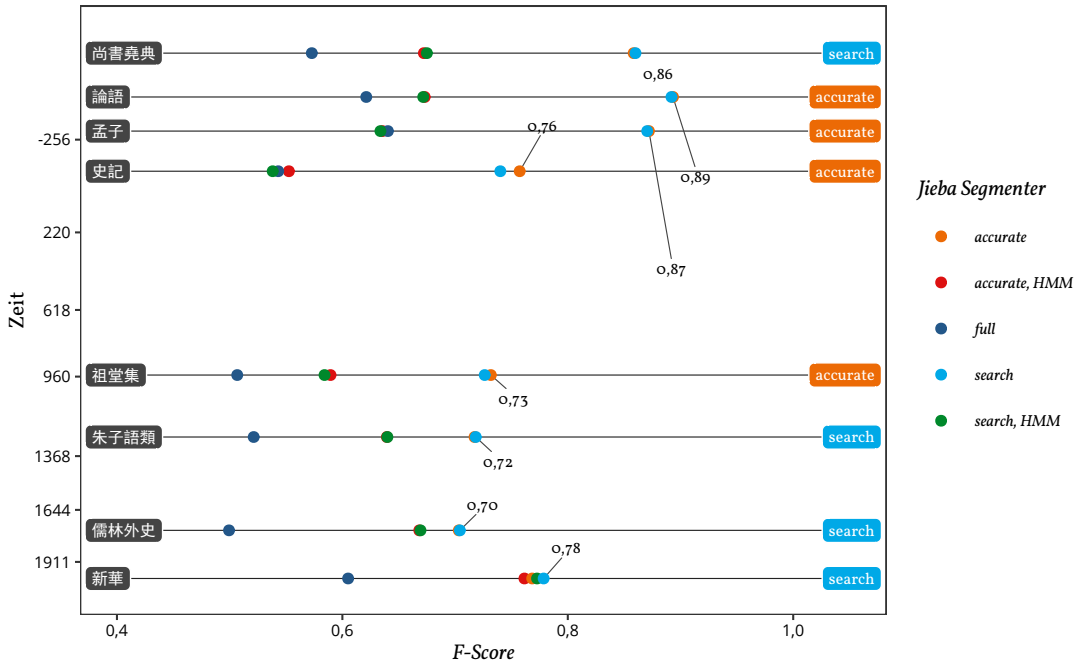

Abbildung 4.1 *F-Scores* der *Jieba-Segmenter*: links ist der segmentierte Goldstandardtext angegeben, rechts der Modus mit der jeweils besten Performance.

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.

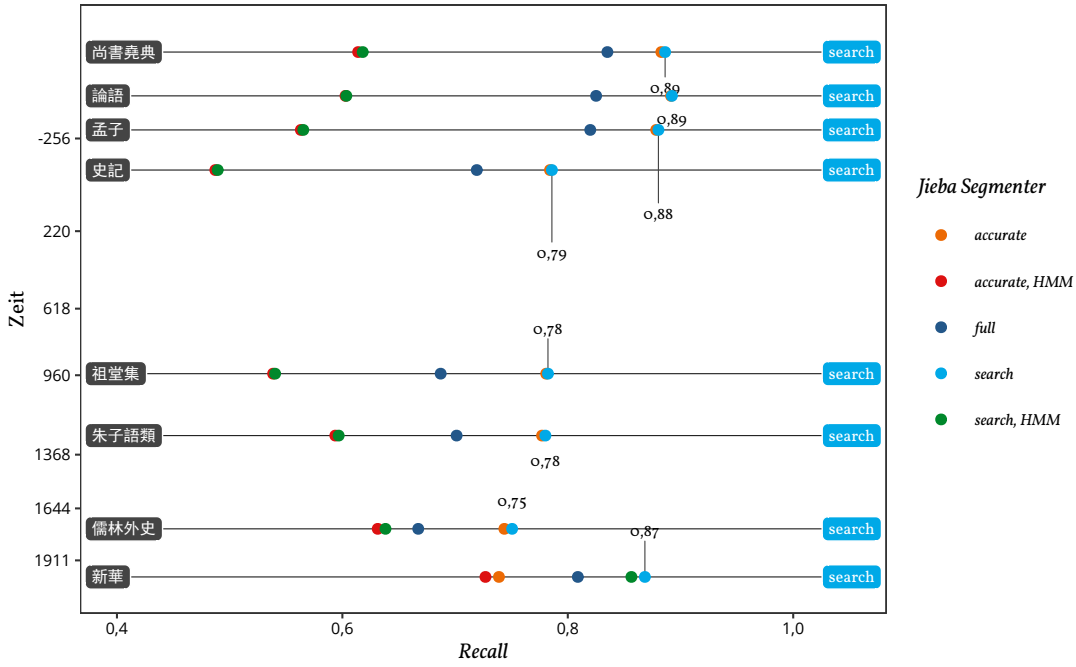

Abbildung 4.2 *Recall* der *Jieba* Modi, diachrone Goldstandards

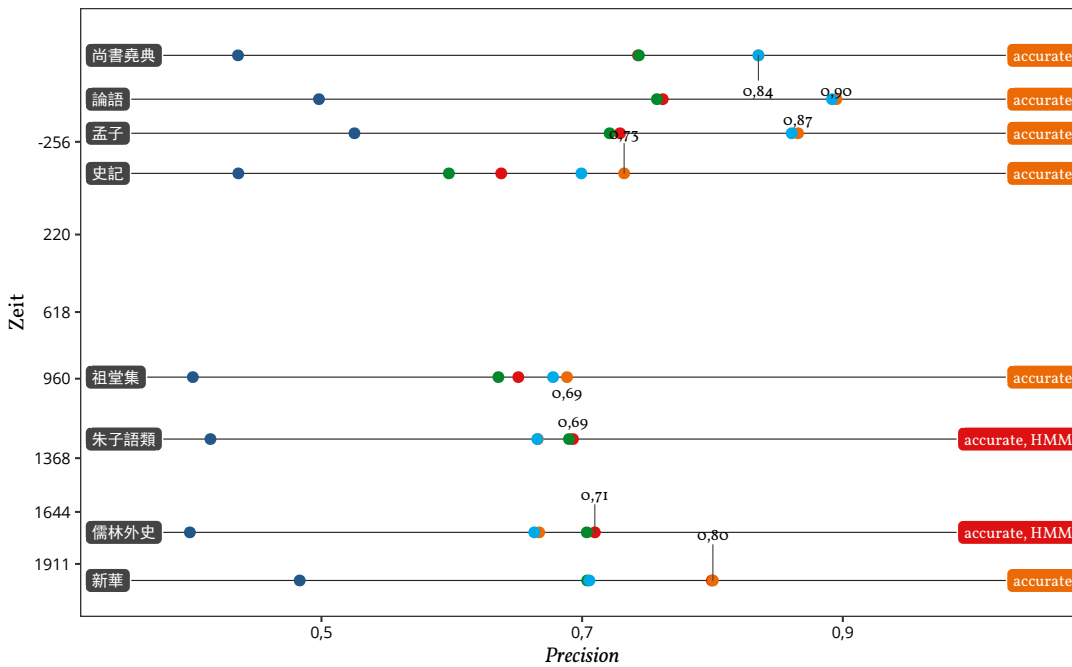

Abbildung 4.3 *Precision* der *Jieba* Modi, diachrone Goldstandards

/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.

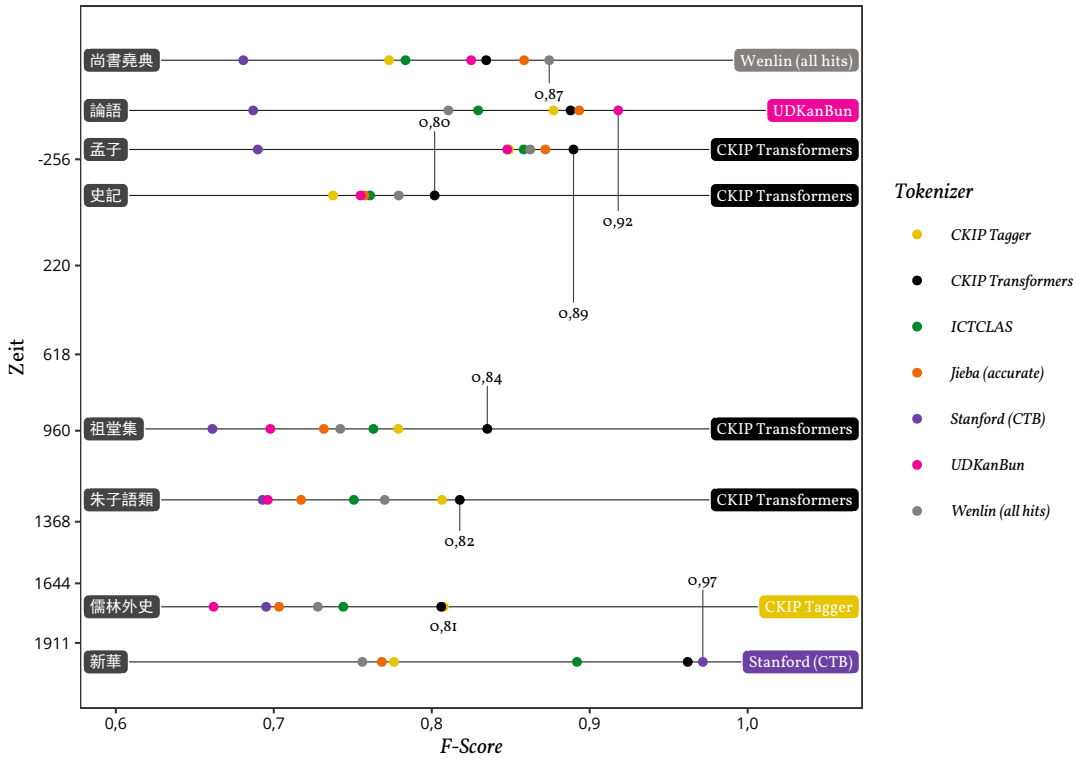

Abbildung 4.4 *F-Scores* aller getesten Tokenizer für alle Goldstandard-Texte. Links ist jeweils der Titel des Texts angegeben, rechts der Name des Tokenizers mit der jeweils besten Performance.

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:

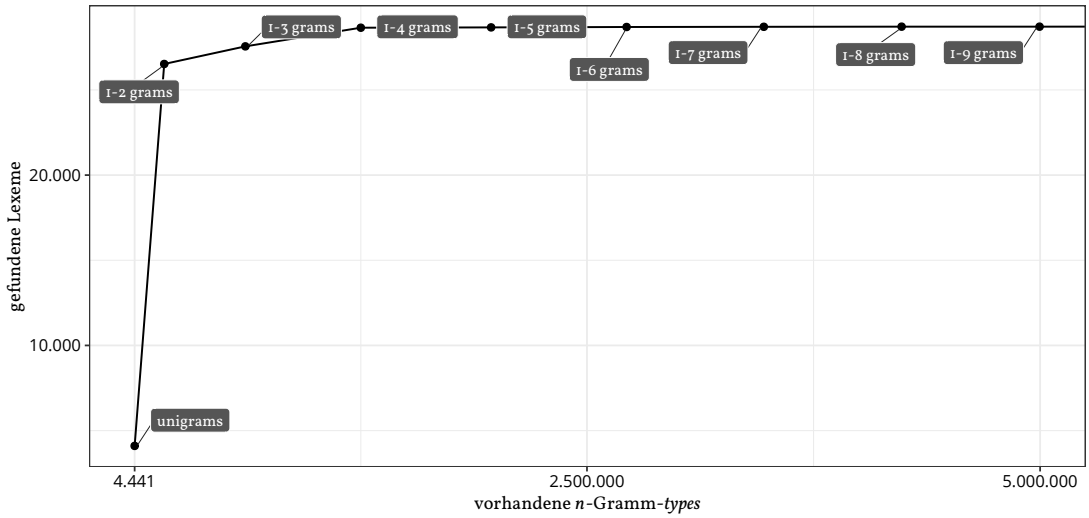

Abbildung 4.5 *n*-Gramm Effizienz am Beispiel von *Hong lou meng* 紅樓夢

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.

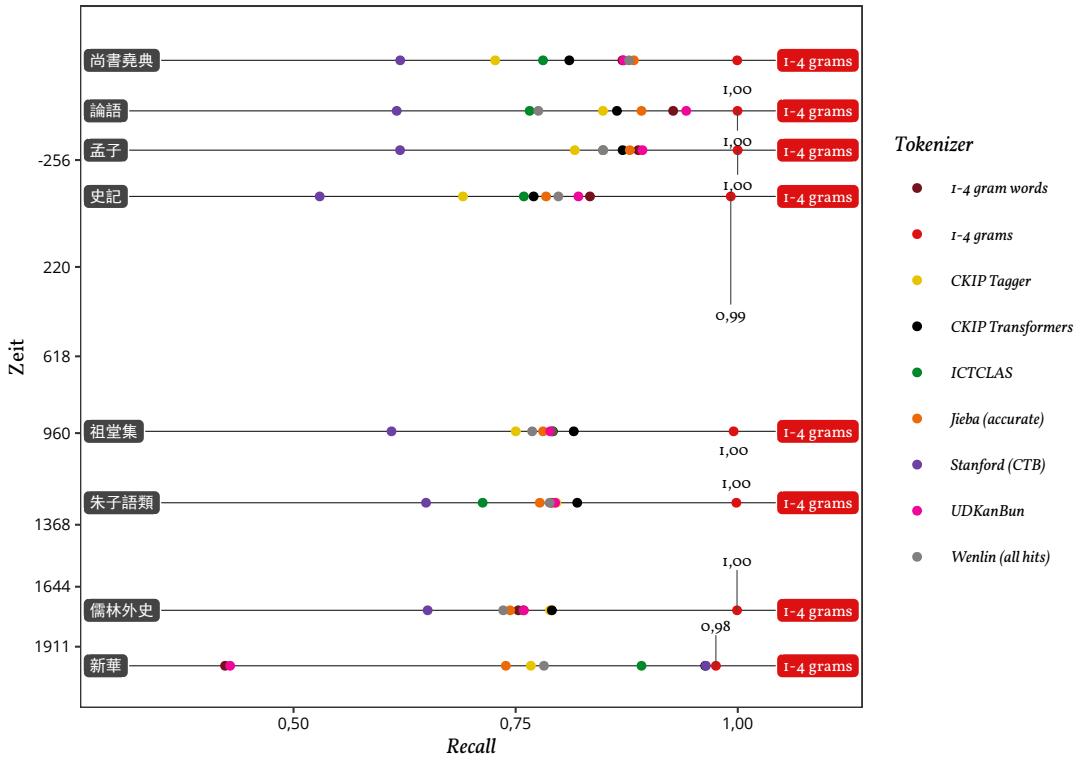

Abbildung 4.6 *Recall* der getesteten Tokenizer vs. Verwendung von 1–4-Grammen

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.

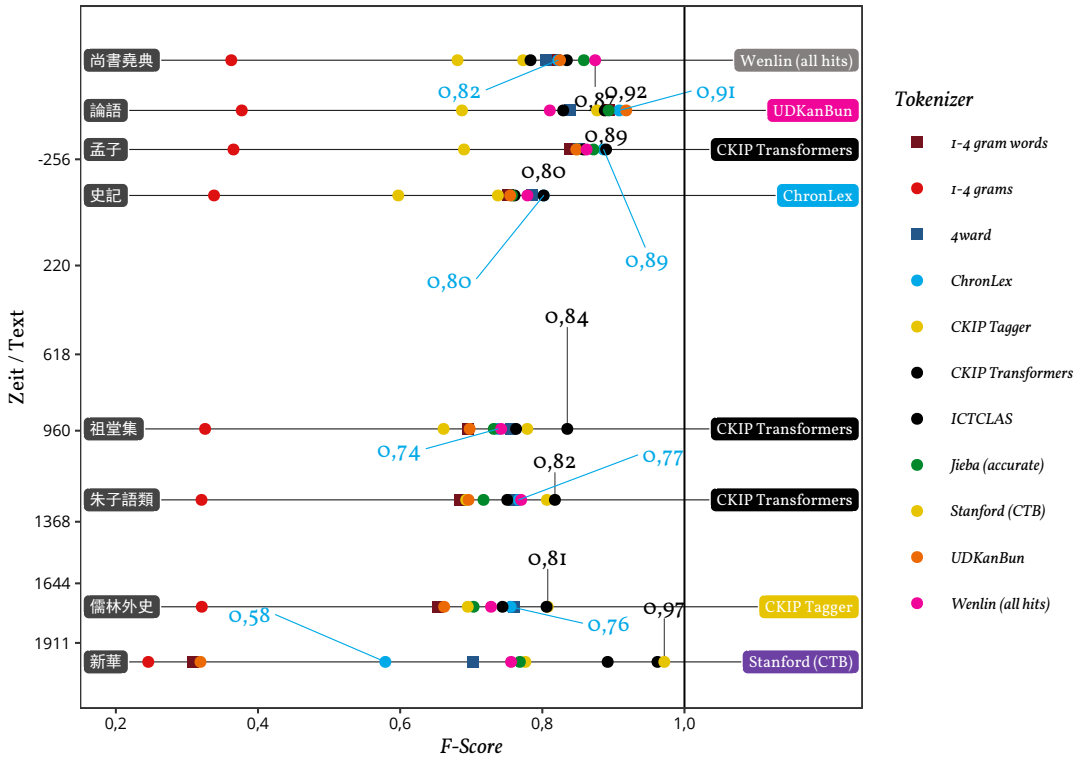

Abbildung 4.7 *F-Score* der getesteten Tokenizer vs. *ChronLex*

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).

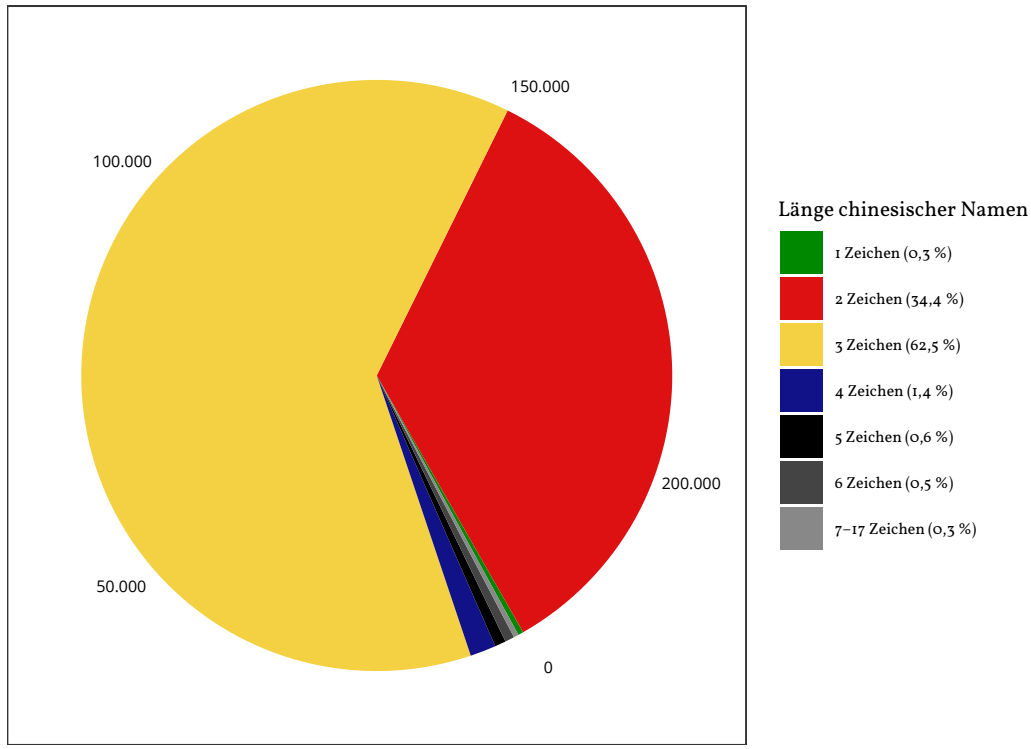

Abbildung 4.8 Länge unterschiedlicher Namen in der *CBDB* in Zeichen, anteilig

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*.

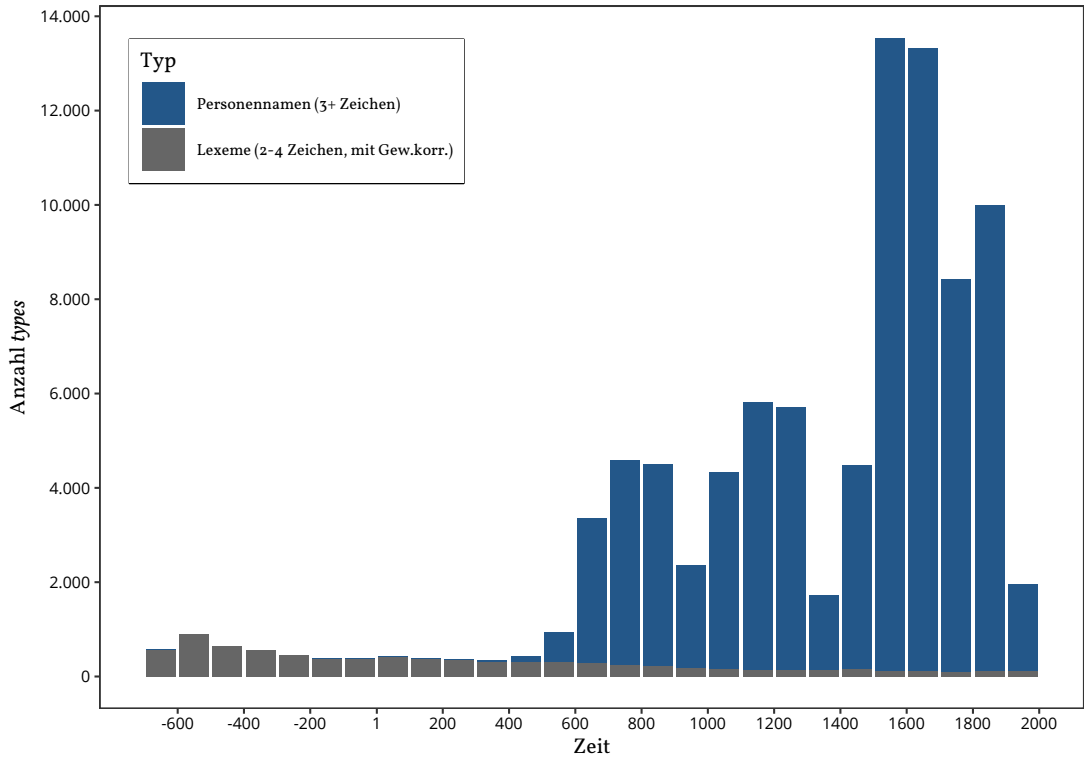

Abbildung 4.9 2–3 Gramme einzigartiger Namen in der *CBDB* nach Jahrhundert

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.