5 Das *Hanyu da cidian* 漢語大詞典 als Datenquelle

„Was ist eines wörterbuchs zweck?
nach seiner umfassenden allgemeinheit kann ihm
nur ein groszes, weites ziel gesteckt sein.“[1]

Jacob Grimm

Das einsprachige *Hanyu da cidian* 漢語大詞典 (*Großes Wörterbuch der chinesischen Sprache*, *HYDCD*)[2] dient als wesentliche Datengrundlage für einen Teil der im Rahmen dieser Arbeit entwickelten Datierungsmethoden und deren Evaluation. Es enthält mit insgesamt etwa 370.000 Einträgen einen umfassenden Wortschatz, der die schriftlich überlieferten Anfänge der chinesischen Schriftsprache (*Shangshu* 尚書, *Shijing* 詩經) aus dem ersten Jahrtausend vor unserer Zeitrechnung bis hin zu Neologismen aus den 1990er Jahren abdeckt. Bei der Auswahl der Worteinträge lautete die Vorgabe, Altes und Neues solle gleichermaßen aufgenommen werden, und dabei [sprachlichem] Ursprung und Entwicklung dieselbe Bedeutung beigemessen werden.[3] Zusätzlich zu den unterschiedlichen Bedeutungen der enthaltenen Lexeme werden zumeist Belegstellen dafür aus Primärquellen zitiert, wobei die Herausgeber in der Regel *versuchen*, die jeweils früheste überlieferte Textstelle am Anfang einer Reihe solcher *attestations* anzugeben.[4] Aus bibliographischen Angaben zu diesen Zitaten lassen sich also chronologische Informationen über die Verwendung dieser Lexeme gewinnen. Das *HYDCD* ist zudem – ungeachtet zahlreicher Bemühungen, es als unzuverlässig darzustellen – das bisher umfangreichste *digital verfügbare* historische Wörterbuch der chinesischen Sprache.

Die digitale Ausgabe des *HYDCD*[5] kann also sowohl als Grundlage für die Erzeugung einer diachronen Lexemdatenbank (Kapitel 5.5, ab S. 120), als auch zur Erzeugung von *chronon*-Korpora auf Basis dieser Datenbank und den im *DHYDCD* als Belegstellen angegebenen Textzitaten (Kapitel 5.6, ab S. 137) verwendet werden.

Einige Sinolog:innen sehen im *HYDCD* eine Art kleineres Plagiat des ähnlich aufgebauten *Zhongwen da cidian* 中文大辭典 *(Großes Wörterbuch des Chinesischen)*[6] das selbst wiederum als Über-

1 Jacob und Wilhelm Grimm 1854: *Deutsches Wörterbuch*. Bd. I. A–Biermolke. Leipzig: S. Hirzel, S. XII.

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

3 „古今兼收，源流并重“ *HYDCD*, Bd. 1, S. 1; Yu Zhangrui 余章瑞 1988: „为伊消得人憔悴——记《汉语大词典》的编纂及为其辛勤工作的人们 (Zum Gedenken an Yi Xiao, Erinnerung an die Herausgabe und die Menschen, die hart am *HYDCD* gearbeitet haben)“. In: *Renmin ribao* 人民日报 06.23.

4 Siehe z. B. Henning Klöter 2013: „Chinese lexicography“. In: *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Hrsg. von Rufus H. Gouws et al. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: DeGruyter Mouton, S. 884–893, S. 887.

5 *DHYDCD*.

6 Zhang Qiyun 張其昀 et. al., Hrsg. 1973–1979: *Zhongwen da cidian* 中文大辭典 *(Großes Wörterbuch des Chinesischen)*. Bd. 1–10. Yangmingshan 陽明山: Zhongguo wenhua xueyuan 中國文化學院.

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

setzung des *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)*[7] betrachtet werden kann.[8] Beide früher erschienenen, konkurrierenden Wörterbücher enthalten deutlich mehr Zeichen- und Worteinträge als das *HYDCD*, obwohl letzterem ein „unprecedented scope" nachgesagt wird.[9] Eine entsprechende Diskussion[10] bleibt an dieser Stelle müßig: Trotz bekannter Schwächen (s. u.) bleibt das *HYDCD* als Datenquelle alternativlos, da vergleichbare Wörterbücher wie das *Zhongwen da cidian* und *Dai Kan-wa jiten* nicht in digitaler Fassung vorliegen.[11] Dem *HYDCD* wird ferner eine „greater awareness of historical principles operative in the development of language"[12] bescheinigt.

Ein Vorzug des *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)* dürfte in der Aufnahme von Personen- und Ortsnamen liegen. Durch eine Ergänzung der *DHYDCD*-Daten um Einträge aus der *China Biographical Database Project (CBDB)*[13] lässt sich dieser Nachteil abschwächen bzw. gänzlich eliminieren.

Die im *HYDCD* implizit enthaltenen Lexikalisierungsdaten auf diese Weise zu nutzen, macht es erforderlich, ein tieferes Verständnis dieser Daten zu gewinnen, um Vor- und Nachteile, sowie Einschränkungen der damit möglichen Ergebnisse sichtbar zu machen. Eine Analyse der erzeugten Daten (Kapitel 5.7, ab S. 138) kann dabei nicht nur weitere Einblicke in die Machart des *HYDCD* selbst geben, sondern auch einige kultur- und vor allem sprachgeschichtliche Entwicklungen sichtbar machen.

Obwohl das *HYDCD* für viele Sinolog:innen ein wichtiges Nachschlagewerk darstellt,[14] ist seine Entstehungsgeschichte bislang – zumal in westlichen Sprachen – kaum bearbeitet worden. Sowohl Wilkinson[15] als auch Hargett[16] und Mair[17] gehen zwar kurz darauf ein, geben aber im Wesentlichen die Informationen wieder, die auch im Vorwort des *HYDCD* selbst zu lesen sind.[18]

Wie die Herausgeber des *HYDCD* ihre Quellen für die sprachgeschichtlich so wesentlichen Belegstellen ausgewählt haben, bleibt darin relativ obskur. Hinzu kommt, dass eine Literaturliste der verwendeten Texte fehlt. Auch Hinweise auf die verwendete Ausgabe des jeweiligen Textes

7 Morohashi Tetsuji 諸橋轍次 1955–1960.
8 Siehe Wilkinson 2000, S. 73; siehe auch Victor H. Mair, Hrsg. 2003: *An Alphabetical Index to the Hanyu da cidian*. Honolulu: University of Hawai'i Press, S. 3.
9 Mair 2003, S. 3.
10 Eine vergleichende Analyse der Abdeckung und Qualität der Einträge findet sich z. B. in James M. Hargett 1990: „Review: Hanyu dacidian 漢語大詞典 by Luo Zhufeng 羅竹風". In: *Chinese Literature: Essays, Articles, Reviews (CLEAR)* 12, S. 138–143. DOI: 10.2307/495232, S. 140–142; Über Fachausdrücke etwa schreibt Hargett: „the glosses dealing with such words in the *Hanyu* were superior in any way to those in the *Zhongwen* and *Daikanwa*" Hargett 1990, S. 142.
11 Tatsächlich liegt das auch das *Dai Kan-wa jiten* seit April 2021 als digitale Ausgabe vor. Bei dieser handelt es sich jedoch lediglich um einen hochwertigen Scan, der nach den 51.110 *Kanji* 漢字 Einzelzeichen-Einträgen durchsucht werden kann. Ein digitaler *Plain*-Volltext existiert weiterhin nicht und eine Suchfunktion für mehrsilbige Lexeme fehlt. Vgl. Morohashi Tetsuji 諸橋轍次 (Komp.), Kamata Tadashi 鎌田正 und Yoneyama Torataro 米山寅太郎 (Rev.), Hrsg. 2021 [1955, 1990, 2000]: *Dai Kan-Wa jiten* 大漢和辞典 *Web* 版 *(Großes Chinesisch-Japanisches Wörterbuch, Online-Ausgabe)*. Bd. 1–13; 15. Tokyo 東京: Taishukan shoten 大修館書店, JapanKnowledge.
12 Mair 2003, S. 3.
13 *CBDB*, siehe Kapitel 5.5.3, ab S. 132, siehe auch Kapitel 4.7, ab S. 97.
14 Die Unverzichtbarkeit des *HYDCD* als sinologisches Hilfsmittel zeigt sich etwa darin, dass Mair einen aufwändig kompilierten *Pinyin*-Index zur Unterstützung des Nachschlageprozesses, den er als „agonizingly protracted" beschreibt, herausgegeben hat. Mair 2003, S. 4; siehe auch Wilkinson 2000, S. 69–73.
15 Siehe Wilkinson 2000, S. 69–71.
16 Hargett 1990, Siehe.
17 Siehe Mair 2003, S. 3–10.
18 Siehe *HYDCD*, Bd. 1, S. 1.

sucht man vergeblich, obwohl sich gerade bei älteren Texten wegen ihrer „fluid nature“[19] unterschiedliche Ausgaben massiv voneinander unterscheiden können.

Um ein präziseres Verständnis für die Datengrundlage zu schaffen, wird im Folgenden auch auf die Entstehungsgeschichte des *HYDCD*, sowie seinen Aufbau und Inhalt eingegangen (Abschnitt 5.3, ab S. 113). Um die Entstehung des *HYDCD* nachzuvollziehen, ist außerdem der Exkurs in die Geschichte eines wichtigen Vorbilds ‚[20] des *Oxford English Dictionary* (*OED*, Abschnitt 5.2, ab S. 111) aufschlussreich, dessen Genese sehr gut erforscht ist.

Da die Umformung der Inhalte zu einer relationalen Datenbank nicht auf Basis der gedruckten Ausgabe geschehen kann, muss für dieses Unterfangen eine digitale Ausgabe verwendet werden, die nicht exakt mit der gedruckten Version übereinstimmt. Ihr Inhalt und ihre Qualität werden deswegen im Stichprobenverfahren auf Übereinstimmung mit der Originalausgabe geprüft (Abschnitt 5.4, ab S. 115).

5.1 Eine kurze Geschichte des *HYDCD*

Das „historische Wörterbuch der chinesischen Sprache“[21] wurde zwischen 1986 und 1993 sukzessive in insgesamt zwölf Bänden veröffentlicht, hinzu kommt ein 1994 erschienener Indexband. Aus dem Wörterbuch selbst geht nur wenig über seine Entstehungsgeschichte und die Herangehensweise der Herausgeber und ihrer Mitarbeiter:innen hervor, sie lässt sich jedoch teilweise aus in der *Renmin ribao* 人民日報 (*RMRB*) erschienenen Zeitungsartikeln nachvollziehen. Das dort gezeichnete Bild ist mit Vorsicht zu genießen, da die Autor:innen der *RMRB* mit teils überzogen verherrlichenden, „patridiotischen“ Formulierungen dem Ruf der Zeitung als Organ der Kommunistischen Partei Chinas (KPCh) an vielen Stellen mehr als gerecht werden.[22] Der Leserschaft wird dabei eine Entstehungslegende vermittelt, die um den früheren Premierminister Zhou Enlai 周恩來 (1898–1976)[23] gewoben wird.[24] In einem Artikel aus dem Jahr 1997 wird sogar noch der gerade verstorbene „Genosse Xiaoping“ (Deng Xiaoping 鄧小平, 1904–1997) als einer der Auftraggeber für das *HYDCD* bemüht.[25] Übereinstimmenden Berichten der *RMRB* zufolge soll Zhou wiederholt in Situationen gekommen sein, in denen er beim Treffen mit anderen Staatsoberhäuptern eindrucksvolle, umfassende Wörterbücher aus dem jeweiligen Land geschenkt bekam, aber kein adäquates Gegengeschenk vorzuweisen hatte.[26] So auch beim Besuch eines Gesandten des Fürstentums Monaco, bei dem von chinesischer Seite als Staatsgeschenk lediglich das *Xinhua zidian* 新華字典 überreicht wurde. Zhou soll diesen Umstand mit den Worten „小國送大書，大國送小書“ („Kleines Land schenkt großes

19 Tao Hongyin 2015.

20 Mair bezeichnet das *HYDCD* sogar als „closest approximation to the *Oxford English Dictionary* that is available“. Mair 2003, S. 3.

21 „*lishixing de Hanyu yuwen cidian* , 历史性的汉语语文词典‘“ *HYDCD*, Bd. 1, S. 1.

22 Man sieht sich unbescheiden in der über 2.000-jährigen Tradition von *Erya* 爾雅, *Shuo wen jie zi* 說文解字 und *Kangxi zidian* 康熙字典, die Mitwirkenden werden für ihre Opferbereitschaft glorifiziert und dramatisch geschildert, wie „mit letzter Kraft“, „mit zitternder Hand“, im Krankenhaus usw. gearbeitet wurde. Siehe z. B. Yu Zhangrui 余章瑞 1988; Guan Xi 冠西 1997: „难忘罗老风范 (Schwer, das Gebaren des alten Luo [Zhufeng 羅竹風] zu vergessen)“. In: *Renmin ribao* 人民日报 11.10.

23 Kai Vogelsang 2012: *Geschichte Chinas*. Stuttgart: Reclam, S. 644.

24 Zhao Lanying 趙蘭英 1986: „《Hanyu da cidian》 bianzuan wancheng《汉语大词典》编纂完成 (Die Kompilation des *HYDCD* ist abgeschlossen)“. In: *Renmin ribao* 人民日报 01.11; Yu Zhangrui 余章瑞 1988; Li Hongbing 李泓冰 1994: „龙飞在天 —— 《汉语大词典》编纂前前后后 (Der Drache fliegt – Die ganze Geschichte hinter der Kompilation des HYDCD)“. In: *Renmin ribao* 人民日报 05.11.

25 Siehe Guan Xi 冠西 1997.

26 Yu Zhangrui 余章瑞 1988.

Buch, großes Land schenkt kleines Buch") kommentiert haben.[27] Vor diesem Hintergrund soll er dann – bereits schwer erkrankt – die nötigen Schritte in die Wege geleitet haben, ein Planungssymposium durchzuführen, aus dem ein entsprechendes Arbeitskomitee unter der Leitung von Chen Hanbo 陳翰伯 (1914–1988)[28] hervorging.[29] Einen entscheidenden Motivationsimpuls für die Politik gab allerdings sicherlich vor allem die Tatsache, dass in Japan, Südkorea und Taiwan bereits umfangreiche Chinesischwörterbücher herausgebracht worden waren.[30]

Ähnliche Planungen für ein umfassendes einsprachiges Wörterbuch der chinesischen Sprache gab es in Festlandchina allerdings bereits deutlich früher. Ab 1928 planten einige Linguisten um Li Jinxi 黎錦熙 (1890–1978)[31] und Chao Yuan Ren 趙元任 (1892–1982)[32] die Herausgabe eines *Zhongguo da cidian* 中國大辭典.[33] Das Projekt verzögerte sich jedoch durch den zweiten sino-japanischen Krieg (*Kang-Ri zhanzheng* 抗日戰爭, 1937–1945) und später durch weitere militärische Auseinandersetzungen und gesellschaftliche Umbrüche. In einem 12-Jahresplan für wissenschaftliche Forschung von 1956 wurde die Idee wieder aufgegriffen, der Sprachwissenschaftler Lü Shuxiang 呂淑湘 (1904–1998)[34] hatte darin bereits die Herausgabe des *HYDCD* vorgesehen. Erneut wurden die Pläne durch politische Kampagnen wie den „Großen Sprung nach Vorne" (*dayuejin* 大躍進, 1958–1962) und die Kulturrevolution (*wenhua da geming* 文化大革命, 1966–1976)[35] vereitelt.[36]

Nach dem tatsächlichen Projektstart in den 1970er Jahren erhielten bekannte Größen der chinesischen Linguistik wie Wang Li 王力 (1900–1986)[37] und Lü Shuxiang wichtige Beraterfunktionen und Luo Zhufeng 羅竹風 (1911–1996)[38] konnte als leitender Herausgeber gewonnen werden.[39] An den ersten beiden Bänden sollen 458 Personen aus mehreren Provinzen in 34 Gruppen gearbeitet haben,[40] andere Artikel sprechen sogar von über 1.000 Sprachwissenschaftler:innen und 43 *danwei*s 單位 („Arbeitseinheiten").[41],[42]

27 Li Hongbing 李泓冰 1994.
28 OCLC 2019: *oclc.org – Worldcat Identities*. Website. URL: https://www.worldcat.org/identities (besucht am 19. 05. 2019), 陈翰伯 1914-1988 (lccn-nr92017850).
29 Siehe Yu Zhangrui 余章瑞 1988; Li Hongbing 李泓冰 1994.
30 Vgl. auch Yu Zhangrui 余章瑞 1988.
31 OCLC 2019, 黎锦熙 1890–1978 (lccn-n82156948).
32 Ebd., Zhao Yuanren, 1892-1982 (lccn-n50036317).
33 Li Hongbing 李泓冰 1994.
34 OCLC 2019, 呂叔湘 (lccn-n79034713).
35 Die Dauer der Kulturrevolution wird von Historikern unterschiedlich bewertet – während sie offiziell schon 1969 (und erneut im August 1977 durch Hua Guofeng 華國鋒, 1921–2008) für beendet erklärt wurde, wird das eigentliche Ende der Bewegung eher auf den Tod von Lin Biao 林彪 (1907–1971) oder Mao Zedong 毛澤東 (1983–1976) datiert. Siehe z. B. Vogelsang 2012, S. 570–577.
36 Yu Zhangrui 余章瑞 1988.
37 OCLC 2019, 王力 1900-1986 (lccn-n81021999).
38 Siehe Guan Xi 冠西 1997.
39 Siehe Yu Zhangrui 余章瑞 1988.
40 Siehe ebd.
41 Der Begriff *danwei* bezeichnet im System der Volksrepublik China eine Art städtische Wohn- und v. a. Arbeitseinheit. In diesem Kontext dürften hier im weiteren Sinne Arbeitsstätten, z. B. Hochschulen gemeint sein. Siehe z. B. Vogelsang 2012, S. 644.
42 Zhao Lanying 趙蘭英 1986; He Jiazheng 何加正 und Li Hongbing 李泓冰 1994: „中华民族五千年文化的结晶中国辞书出版史上的壮举《汉语大词典》大功告成首都隆重举行庆功会江泽民李鹏等到会祝贺全书13卷，收词语37.5万余条，约5000万字，是千余专家学者18年艰苦努力的结果 (Die Quintessenz der 5.000-jährigen Kultur des chinesischen Volkes, die Höchstleistung der chinesischen Geschichte der Herausgabe von Wörterbüchern, das *HYDCD*, wurde endlich abgeschlossen und zu diesem Anlass in der Hauptstadt eine große Feier ausgerichtet. An der Veranstaltung nahmen Jiang Zemin, Li Peng und andere teil. Das Werk hat insgesamt 13 Bände, 375.000 Wörter wurden aufgenommen, etwa 50 Mio. Zeichen, das Ergebnis der harten 18 Jahre dauernden Arbeit von über 1.000 Spezialisten und Gelehrten)". In: *Renmin ribao* 人民日报 05.11.

Angespornt von den zuvor erschienenen *Dai Kan-Wa jiten* und *Zhongwen da cidian*,[43] und dank dem staatlich geförderten immensen personellen und finanziellen Aufwand benötigte man vom „Startschuss“ 1975 bis zur Fertigstellung des letzten Bandes 1994[44] weniger als 20 Jahre. In direkter Konkurrenz mit vergleichbaren Unternehmungen ist das ein verhältnismäßig kurzer Zeitraum. Die *RMRB* prahlt, dass man – gewissermaßen als Ausgleich für das späte Angehen des Projekts – deutlich schneller fertig geworden sei als die Engländer, die Deutschen oder die Russen.[45] Luo Zhufeng 羅竹風 wird zitiert, das *HYDCD* sei das Ergebnis einer groß angelegten sozialistischen Kooperation und verkörpere konkret die Überlegenheit des sozialistischen Systems.[46] Möglich, dass Luo diese Worte posthum in den Mund gelegt wurden – sie zeigen jedenfalls, dass das *HYDCD* eine gewisse Rolle für das festland-chinesische kulturelle Selbstbild spielt und als Vorzeigeprojekt staatlicher Kulturförderung und -propaganda gesehen werden kann. Hierbei scheut man sich auch nicht, die früher erschienene und doch umfangreichere[47] Konkurrenz aus Japan und Taiwan zu diskreditieren – nicht nur sei man viel schneller gewesen, viele Einträge im *Dai Kan-Wa jiten* oder dem *Zhongwen da cidian* seien gar keine richtigen Wörter (*„xuduo bucheng “ci” de ci* 許多不成“詞”的詞“, eine Aussage, die aus linguistischer Sicht sicherlich für das *HYDCD* ebenfalls zutrifft) und der Inhalt jener Werke sei „diffus“.[48]

Ein 2010 in der *RMRB* veröffentlichter Artikel schlägt wieder etwas differenziertere Töne an und betont, dass andere Nationen wie Frankreich, Deutschland, Russland und die Vereinigten Staaten mit vergleichbaren Wörterbuchprojekten durchschnittlich 50 Jahre früher fertig waren, außerdem habe das 100 Jahre früher erschienene *OED* deutlich mehr Einträge.[49]

5.2 Das Vorbild: *Oxford English Dictionary*

> *„...where every pains has been taken to ascertain the earliest occurrence of each word and of each signification...“*[50]
>
> Otto Jespersen

Wegen seines Vorbildcharakters für historische Wörterbücher sei an dieser Stelle auch auf das *Oxford English Dictionary* (*OED*) eingegangen. Seine Geschichte kann zwar schwerlich offene Fragen über die Entstehung des *HYDCD* beantworten, wegen der Ähnlichkeit beider Projekte kann sie aber zumindest zum Verständnis der Konzeption beitragen, denn im Gegensatz zum *HYDCD* ist die Entstehungsgeschichte des *OED* sehr gut dokumentiert und erforscht.[51] Die Planung für

43 Siehe Yu Zhangrui 余章瑞 1988.

44 *HYDCD*, Bd. 13, S. ii.

45 Li Hongbing 李泓冰 1994; Konkret bezieht man sich hier v. a. auf das *OED*, bei dem die Planung der ersten Ausgabe im Jahr 1858 begann, dessen letzter, zehnter Band aber erst 1928 fertiggestellt wurde. Siehe John Willinsky 1994: *Empire of Words: The Reign of the OED*. Princeton, New Jersey: Princeton University Press, S. 11.

46 „《汉语大词典》是社会主义大协作的产物，是社会主义制度优越性的具体体现.“ Siehe Guan Xi 冠西 1997.

47 Eine den Autoren der *RMRB* vorliegende Ausgabe von Morohashis Wörterbuch wird mit 550.000 Einträgen angegeben.

48 *Pang za* 龐雜, wörtlich etwa „riesig und divers“. Siehe Yu Zhangrui 余章瑞 1988.

49 Siehe Zhang Zhiyi 张志毅 2010: „, 辞书强国‘究竟有多远 (Wie weit es noch bis zum ‚starken Wörterbuchland‘ ist)“. In: *Renmin ribao* 人民日报 10.12.

50 Otto Jespersen 1912 [1905]: *Growth and Structure of the English Language*. Leipzig: B. G. Teubner, S. 222; zitiert in Willinsky 1994, S. 57.

51 Abgesehen von den im Folgenden zitierten Werken seien erwähnt: Katherine M. Elisabeth Murray 1977: *Caught in the Web of Words: James Murray and the Oxford English Dictionary*. New Haven & London: Yale University Press, eine Biographie über den Herausgeber James Murray; sowie ein Roman über die Beziehung zwischen einem der Bei-

dieses Mammutprojekt nahm ihren Anfang 1857 mit einem Vorschlag des Poeten und Bischofs von Dublin, Richard Trench auf einer Sitzung der Philological Society of London. Mehrere Jahrzehnte vergingen bis zur Veröffentlichung des ersten Faszikels, das 1884 noch unter dem Titel *A new English Dictionary* erschien, sowie der ersten vollständigen Ausgabe 1933.[52]

Das erklärte Ziel der Herausgeber des *OED* war es, „eine angemessene Darstellung der Bedeutung, des Ursprungs und der Geschichte der englischen Wörter zu liefern, die allgemein gebräuchlich sind oder bekanntermaßen zu irgendeinem Zeitpunkt während der letzten siebenhundert Jahre gebräuchlich waren“[53] – sehr ähnlich den Zielen der Urheber des *HYDCD*. Zur Legitimierung der Einträge wurden zusammen mit unzähligen Freiwilligen mehr als fünf Millionen Zitate gesammelt, von denen 1.827.306 in der Ausgabe von 1933 Verwendung fanden.[54]

Bei der Auswahl eben dieser Belegstellen liegt ein offensichtliches *Bias* vor: William Shakespeare (1564–1616) als am meistzitierter Autor wird in 14 Prozent aller Einträge herangezogen, insgesamt werden seine Werke 32.868 mal zitiert, was fast 2 Prozent aller Belegstellen entspricht.[55] Dieses *Bias*, das sich in ähnlicher Form auch im *HYDCD* beobachten lässt,[56] wurde früh bemerkt und kritisiert:[57]

> […] one is struck by the frequency with which Shakespeare's name is found affixed to the earliest quotation for words or meanings. […] this is no doubt due to the fact that Shakespeare's vocabulary has been registered with greater care in Concordances […] than that of any other author, so that his words cannot escape notice, while the same words may occur unnoticed in the pages of many an earlier author.[58]

Das Konsultieren von Konkordanzen vereinfacht natürlich die Identifikation von Belegstellen für die Kompilator:innen, bringt aber eine gewisse Unausgewogenheit mit sich. Wahrscheinlich ebenfalls im Vorhandensein entsprechender Konkordanzen begründet ist die Häufigkeit, mit der Übersetzungen europäischer Klassiker wie der *Aeneis*, der Bibel und anderer nicht originär englischsprachiger Texte herangezogen werden.[59] Darin liegt zugleich ein auffälliger Unterschied zwischen *OED* und *HYDCD*. Die Einträge in letzterem werden fast ausschließlich mit Texten belegt, die ursprünglich in chinesischer Sprache verfasst wurden.

---

tragenden und Murray. Simon Winchester 1998: *The Professor and the Madman: A Tale of Murder, Insanity, and the Making of The Oxford English Dictionary*. New York: HarperCollins, – Winchester ist ebenfalls bekannt für seinen Roman über den Sinologen, Biochemiker und Wissenschaftshistoriker Joseph Needham.

52 Trenchs ursprüngliche Idee der Bildung eines Komitees zur Sammlung unbekannter Lexeme mit dem Ziel der Erweiterung bestehender Wörterbücher erschien den Mitgliedern der *Philological Society* nach seiner Vorstellung systematischer Mängel eben jener vorhandenen Nachschlagewerke als nicht ausreichend. Daher wurde beschlossen, stattdessen die Arbeiten an einem *New English Dictionary* zu beginnen. Siehe Willinsky 1994, S. 3–16.

53 James A. H. Murray et al., Hrsg. 1913–1933: *Oxford English Dictionary*. Bd. 1–13. London: Oxford University Press (im Folgenden zit. als *OED*), Bd. 1, S. vi, übersetzt durch den Verfasser. zitiert in Willinsky 1994, S. 76.

54 Siehe Willinsky 1994, S. 3–4, S. 58.

55 Siehe ebd., S. 57–58. Im Anhang (ab S. 211) finden sich hier Tabellen mit genauen Zählungen der am häufigsten zitierten Autoren, Werke, Zeitschriften und Tageszeitungen unterschiedlicher Ausgaben.

56 Siehe Kapitel 5.7, ab S. 138.

57 Als Gegenargument sei an dieser Stelle aus Jacob Grimms Vorwort zum *Deutschen Wörterbuch* zitiert, das die große Zahl Shakespeare-Zitate in besserem Licht erscheinen lässt: „Hin und wieder wird man der belege zu viel angebracht meinen, namentlich aus Luther und Göthe. doch jenes einflusz auf die sprache, Göthes macht über sie müssen reich und anschaulich vorgeführt werden […]“ Grimm 1854, S. xxxvii.

58 Jespersen 1912 [1905], S. 222; zitiert in Willinsky 1994, S. 57.

59 Siehe Willinsky 1994, S. 115.

5.3 Aufbau und Inhalt des *HYDCD*

Eine kurze Beschreibung von Struktur, Aufbau und Machart des *HYDCD* soll über Inhalt, Umfang und Qualität der Daten Aufschluss geben, die für die Textdatierung nutzbar gemacht werden sollen. Die Einträge lassen sich in zwei Hauptkategorien unterteilen:

— 1. **Einträge für einzelne Schriftzeichen**, *dan zi tiaomu* 單字条目, sind gewissermaßen Haupt- oder Übereinträge.[60] Diese sind in 200 Radikale unterteilt,[61] sowie nach Radikal- und Zusatzstrichzahl (*residual strokes*) sortiert.[62] Diese Einträge bezeichne ich im Folgenden als „Zeicheneinträge". In der digitalen Ausgabe werden sie von einem Asterisk (*) eingeleitet. Bei Schriftzeichen, für die nicht nur mehrere Bedeutungen, sondern auch unterschiedliche Aussprachen angegeben sind (*duoyinzi* 多音字), werden die Einträge mit (in der gedruckten Ausgabe hochgestellten) arabischen Ziffern durchnummeriert, z. B. *zǐ* 仔 ¹, *zī* 仔 ² und *zǎi* 仔 ³.[63]

Im Unterschied zum *Hanyu da zidian* 漢語大字典 *(Großes Lexikon chinesischer Schriftzeichen)*[64] liegt der Fokus des *HYDCD* auf *ci* 詞, so dass deutlich weniger Einzelzeicheneinträge bestehen als Schriftzeichen bekannt sind. Das Auswahlkriterium für monosyllabische Einträge soll die fortwährende Verwendung der Zeichen gewesen sein, wohingegen im *Hanyu da zidian* auch *si zi* 死字, also „ausgestorbene" Zeichen, aufgenommen wurden.[65]

— 2. **Einträge mit mehreren Schriftzeichen**, *duo zi tiaomu* 多字条目 (wörtlich etwa: „Mehrzeicheneinträge")[66] sind den Zeicheneinträgen des jeweils ersten Zeichens untergeordnet. Im Folgenden bezeichne ich diese Einträge als „Unter-" oder „Worteinträge". Sie enthalten in erster Linie mehrsilbige „Wörter", d. h. *ci*, bzw. „strings of monosyllabic morphemes".[67] In diese Kategorie fallen zum Teil aber auch lexikalisierte Phrasen.[68]

Das *HYDCD* folgt damit der grundlegenden Terminologie der chinesischen Lexikographie, einer Dichotomie zwischen *zi* 字 und *ci* 詞.[69] Aus morphologischer Sicht können beides „Wörter" sein. Aufgrund der „inherent fuzziness" des in der Linguistik allgemein und für das Chinesische im Besonderen problematischen Wortbegriffs[70] wird hier für alle Zeichen- und Zeichenfolgen mit Einträgen im *HYDCD* der Begriff *Lexeme* verwendet.

In einem typischen Eintrag folgt auf das Stichwort eine Auflistung unterschiedlicher Bedeutungen. Ähnlich wie im *OED* und vergleichbaren Wörterbüchern wie *Dai Kan-Wa jiten* 大漢和辞典 *(Großes Chinesisch-Japanisches Wörterbuch)*, wird meist die Bedeutung erklärt, z. B. durch Angabe

60 Vgl. *HYDCD*, Bd. 1, S. 7.
61 Siehe *HYDCD*, Bd. 1, S. 12.
62 Vgl. *HYDCD*, Bd. 1, S. 7. Im *HYDCD* etwas umständlich: „bei gleichem Radikal nach Strichzahl (abzüglich der Strichzahl des Radikals selbst)" („部首相同的按画数(减去部首本身画数)").
63 *HYDCD*, Bd. 1, S. 7.
64 Xu Zhongshu 徐中舒, Hrsg. 1986–1990: *Hanyu da zidian* 漢語大字典 *(Großes Lexikon chinesischer Schriftzeichen)*. 3 Bde. Wuhan 武漢: Sichuan cishu chubanshe 四川辭書出版社, Hubei cishu chubanshe 湖北辭書出版社 (im Folgenden zit. als *HYDZD*).
65 Vgl. Yu Zhangrui 余章瑞 1988.
66 Vgl. *HYDCD*, Bd. 1, 7.
67 Norman 1988, S. 24.
68 Sprichwörter, sowie Phraseologismen wie *chengyu* 成語, *xiehouyu* 歇後語, und *suyu* 俗語. Siehe z. B. *HYDCD*, Bd. 1, S. 428, *bushi dongfeng yaliao xifeng, jiushi xifeng yaliao dongfeng* 【不是東風壓了西風，就是西風壓了東風】 od. Bd. 6, S. 1311, *zhao san mu si* 【朝三暮四】.
69 Eine aktuelle, ausführliche Diskussion dieser Problematik findet sich in Jiang Shaoyu 蔣绍愚 2015, S. 1–4; siehe auch Klöter 2013, S. 885.
70 Vgl. Lukáš Zádrapa 2015: „Word and Wordhood in Classical Chinese". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

von Synonymen, gefolgt von einer diachronen Reihe von Textbeispielen, die als Belegstellen zitiert werden – eine Struktur, die so erstmals im ab 1836 erschienenen *New Dictionary of the English Language* Verwendung findet.[71] In Wörterbüchern Bedeutungen mit illustrativen Textbeispielen zu belegen, ist keineswegs eine aus Großbritannien übernommene Innovation: in der chinesischen Lexikographie hat diese Praxis eine deutlich längere Tradition – bereits im *Erya* 爾雅[72] lassen sich solche Zitate aus Klassikern finden, auch wenn die Quellen nicht explizit angegeben werden.[73] Das 1710–1716 kompilierte *Kangxi Zidian* (*KXZD*) bietet bereits Belegreihen an, die aber kaum chronologisch geordnet sind und überwiegend frühe, kanonisierte Texte zitieren.[74] Aufgeführt werden dabei zudem Verwendungen einzelner Zeichen (*zi* 字), Zeichenverbindungen werden nicht systematisch berücksichtigt.[75]

Die Belegstellen im *HYDCD* stammen aus den unterschiedlichsten Textgattungen, darunter kanonisierte philosophische Klassiker, Geschichtstexte, Gedichte, Romane, Zeitungsartikel – bis hin zu Veröffentlichungen der kommunistischen Partei. Im Unterschied zum *OED* werden die Textbeispiele weder kontinuierlich für jedes Jahrhundert, in dem sich ein Wort nachweisen lässt, gegeben, noch die Erscheinungsjahre der Erstausgabe der zitierten Texte angegeben. Die Kompilator:innen des *HYDCD* begnügen sich mit Angabe von Dynastie und Autor:in, bei Kanontexten sogar mit dem Titel des zitierten Textes.[76] Trotz der Vorbildfunktion des *OED* sollte nicht vergessen werden, welches Sprachverständnis die Kompilator:innen des *HYDCD* hatten und vor allem, in welcher Tradition sie stehen. Die Selbstverständlichkeit, mit der Texte wie *Shangshu* oder auch die Dynastiegeschichten hier ohne Angabe von Autor:in, Ausgabe oder Dynastie zitiert werden, kann in diesem Licht auch mehr als Traditionsbewusstheit, denn als sprachwissenschaftliche Verfehlung erscheinen.

Wie im *OED* versucht man dabei, den *Locus classicus* ausfindig zu machen und als Beleg anzugeben.[77] Es liegt auf der Hand, dass dieses Unterfangen nicht immer gelingen kann. So hat sich in der chinesischen Lexikographie inzwischen gewissermaßen ein eigenes Aufsatzgenre etabliert, dessen Hauptinhalt die Ergänzung noch früherer Belegstellen (*ante-dating*) zu *HYDCD*-Einträgen ist.[78] Auch Li Shens 李申 Monographie *Hanyu da cidian yanjiu* 《汉语大词典》研究 *(A*

71 Siehe Charles Richardson 1836: *A New Dictionary of the English Language*. London: W. Pickering; erwähnt in Willinsky 1994, S. 21, S. 29, S. 94. Während Richardson in der ersten vollständigen Ausgabe von 1836 noch sehr spärlich mit Belegen umgeht, sind sie in der mehrbändigen, ab 1851 erschienenen Ausgabe bei einem Großteil der Einträge vorhanden. Vgl. Charles Richardson 1851: *A New Dictionary of the English Language. 2 vols.* Philadelphia: E. H. Butler & Co.; Eine sorgfältigere Zitierweise mit Jahres- und Seitenzahlen findet sich allerdings in diesem Kontext zuerst im *Deutschen Wörterbuch*, dessen erste Lieferung nur kurze Zeit später, 1852 erfolgte. Im Vorwort des ersten, vollständigen Bandes erläutert Jacob Grimm zu den Belegen: „[…] der name ihres urhebers reicht nicht aus, sie müssen aufgeschlagen werden können […]". Grimm 1854, S. xxxvi.

72 Der Titel wird im Englischen mit „approaching what is correct, proper, refined" wiedergegeben. South W. Coblin 1993: „Erh ya 爾雅". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 94–99, S. 94.

73 Siehe z. B. Yong Heming und Peng Jing 2008: *Chinese Lexicography: A History from 1046 BC to AD 1911*. Oxford: Oxford University Press, S. 90–91.

74 Nach-hanzeitliche Quellen werden eher vereinzelt zitiert, z. B. ein *ci* 詞 des tangzeitlichen Dichters Li Bai (701–762) und die 1343 veröffentlichte *Yuan shi* 元史 im Eintrag zu *jin* 金. Siehe Aisin-Gioro Xuanye 愛新覺羅 · 玄燁 (als Shengzu ren huangdi 聖祖仁皇帝) 1922 [1716], S. 1295.

75 Zum Aufbau des *KXZD* siehe z. B. auch Marc Winter 2015: „Kāngxī zìdiǎn 康熙字典". In: *Encyclopedia of Chinese Language and Linguistics Online*. Hrsg. von Rint Sybesma et al. Leiden: Brill.

76 Zur Veranschaulichung von Inhalt und Struktur sei auf die vom Verfasser kommentierten Beispiele weiter unten (Abschnitt 5.5.2, ab S. 120) verwiesen.

77 In 2.450 Einträgen des *DHYDCD*, also bei einem Anteil von unter einem Prozent, findet sich explizit die Angabe *yuben* „ 语本…" („die Redewendung / das Wort hat seinen Ursprung bei / in…"). Da diese Markierung aber selten und nicht systematisch eingesetzt wird, widme ich ihr hier keine weitere Aufmerksamkeit. Vgl. *DHYDCD*, *passim*.

78 Vgl. z. B. Liu Bing 劉冰 2009: „《汉语大词典》书证迟后例补——以《先秦漢魏晉南北朝詩 (梁詩)》为例 (Ergänzungen für späte Belegstellen im *HYDCD* - anhand von Gedichten der Prä-Qin, Han, Wei, Jin und Nanbei-Zeit [Liang

study on Hanyu da cidian) widmet ein langes Kapitel der Ergänzung früherer Belegstellen.[79] Selbst wenn der wirkliche *Locus classicus* von den Herausgebern nicht gefunden wurde, lässt sich dennoch mit Sicherheit sagen: wenn eine Bedeutung z. B. mit einem Han-zeitlichen Text belegt wird, kann angenommen werden, dass das Lexem mindestens zu dieser Zeit so verwendet wurde. Die früheste Belegstelle impliziert damit ein Mindestwortalter.

Etliche Aufsätze befassen sich zudem mit der Identifizierung von Wörtern, die nicht als Einträge aufgenommen wurden, sowie der Ergänzung und Korrektur von Wortbedeutungen.[80] Bei aller berechtigten Kritik gilt das *HYDCD* dennoch als „most authoritative dictionary of the Chinese language."[81]

Eine herausfordernde Eigenheit des *HYDCD* ist die gemischte Verwendung von Lang- und Kurzzeichen. Die Herausgeber rechtfertigen diese Maßnahme mit der Historizität des Wörterbuchs. Die Worterklärungen sind grundsätzlich in Kurzzeichen (*jiantizi* 簡 [简] 體 [体] 字) verfasst, doch Zeichen- und Worteinträge selbst werden in Langzeichen (*fantizi* 繁體字) angegeben.[82] Vereinfachte Zeichenformen erhalten einen zusätzlichen Zeicheneintrag, der auf das entsprechende Langzeichen verweist.[83] Nicht nur Glossen, auch Namen von Autor:innen, Herausgeber:innen, sowie Titel von zitierten Werken und Dynastien, sind in Kurzzeichen gesetzt. Eine Ausnahme davon bilden Namen wie Gao Shi 高適, bei denen es so zu Mehrdeutigkeiten kommen kann – hier wird z. B. nicht 高适 geschrieben, da 适 *kuò* und 適 *shì* ein gemeinsames Kurzzeichen (适) teilen. Die Belege werden wiederum in Langzeichen gesetzt, sofern der zitierte Primärtext vor 1912 entstanden ist, oder selbst in Langzeichen verfasst wurde.[84]

5.4 Digitale Ausgaben des *HYDCD*

Voraussetzung für die Nutzung des *HYDCD* als Datengrundlage für Softwareprojekte ist eine digitale Ausgabe. Es kursieren drei unterschiedliche, zwischen 1997 und 2007 veröffentlichte CD-Rom Versionen, sowie mehrere Online-Veröffentlichungen.[85] Da keine der offiziellen Ausgaben als Volltext verfügbar ist, greife ich auf eine Online-Version (im Folgenden *DHYDCD*) zurück, die inhaltlich weitestgehend der gedruckten Ausgabe entspricht und in der eben-

Gedichte])". In: *Yuwen Xuekan* 语文学刊 *(Journal of language and literature studies)* 19, S. 72–73, S. 72–73; Zheng Xianzhang 郑贤章 2000: „《Hanyu da cidian》shuzheng chushi li shi bu《汉语大词典》书证初始例试补 (Supplementing some Earlier Citations to Hanyu da cidian)". In: *Gu Hanyu yanjiu* 古汉语研究 *(Research in Ancient Chinese Language)* 2, S. 94–96, S. 94–96; Siehe auch Jing-Schmidt und Hsieh 2019, S. 516 „Studies of the translated religious texts and the vernacular materials preserved at Dunhuang have enabled lexicographers to trace the attestations of many words to an earlier time point than indicated in the standard dictionary Hànyǔ Dàcídiǎn".

79 Li Shen 李申 und Wang Benling 王本靈 2015: *Hanyu da cidian yanjiu*《汉语大词典》研究 *(A study on Hanyu da cidian)*. Beijing 北京: Shangwu yinshuguan 商務印書館 (The Commercial Press), S. 96–172.

80 Siehe z. B. Hu Shaowen 胡绍文 2002: „The Shortages of Hanyu Da Cidian (汉语大词典) From the View of Yi Jian Zhi (夷坚志) – 从《夷坚志》看《汉语大词典》的若干阙失". In: *Research In Ancient Chinese Language* 古汉语研究 4, S. 87–89; Chai Hongmai 柴红梅 2005: „To Remedy Some Flaws of Hanyu da cidian (汉语大词典) – On the Basis of Entry C in Xiandai Hanyu cidian (现代汉语词典)《汉语大词典》瑕疵补正 —- 以《现代汉语词典》C 字条为例". In: *Research In Ancient Chinese Language* 古汉语研究 3.

81 Klöter 2013, S. 887.

82 Vgl. *HYDCD*, Bd. 1, S. 8.

83 Vgl. *HYDCD*, Bd. 1, S. 9.

84 Vgl. *HYDCD*, Bd. 1, S. 8; Siehe auch Hargett 1990, S. 139.

85 Für einen Überblick und genaueren Vergleich siehe Yang Lin 杨琳 2011: *"Hanyu da cidian"guangpan ban yu zhizhiban de qubie*《汉语大词典》光盘版与纸质版的区别 *(Unterschiede zwischen der CD-Rom und der Papierausgabe des Hanyu da cidian)*. URL: http://www.guoxue.com/?p=4453 (besucht am 22. 07. 2018), *passim*. Der Autor rät von einer leichtgläubigen Verwendung zumindest der CD-Rom Versionen zu wissenschaftlichen Zwecken ab („*bu neng qing xin guangpan ban* 不能轻信光盘版"), wohingegen die Online-Ausgabe 2.0 (汉语大词典》网络版 2.0) gelobt wird.

falls Kurz- und Langzeichen gemischt verwendet werden.[86] Sie enthält insgesamt 365.102 Wort- und Zeicheneinträgen (davon 22.327 Zeicheneinträge mit insgesamt 16.361 graphisch unterschiedlichen Schriftzeichen-*types*).

Diese Version entspricht keiner der genannten CD-Rom Versionen: Die 1997 erschienene **Version 1.0** kommt mit nur 18.000 Zeichen- und 336.000 Worteinträgen nicht in Frage, auch scheinen in dieser Ausgabe etliche Belegstellen zu fehlen, was bei der verwendeten Ausgabe kaum der Fall ist.[87] Die zuerst 2003 veröffentlichte **Version 2.0** verwendet zwar, wie die vorliegende Version, eine Mischung von Lang- und Kurzzeichen, es wurden jedoch Einträge hinzugefügt, die in der gedruckten Ausgabe fehlen, z. B. die Zeicheneinträge zu *qiao* / *jiao* 釙, *wang* 亟 und *fa* 泼. Alle drei fehlen in der verwendeten Version. Die aus Kompatibilitätsgründen schwieriger zu betreibende **Version 3.0** aus dem Jahr 2007 kommt mit 336.706 Worteinträgen ebenfalls nicht infrage.[88]

Auch wenn die tatsächliche Provenienz der verwendeten Daten schwer feststellbar ist, entspricht sie mit hoher Wahrscheinlichkeit im Wesentlichen der Online-Ausgabe 2.0 des *HYDCD*.[89] Die Tatsache, dass die neueste zitierte Primärquelle aus dem Jahr 1992 stammt,[90] lässt zudem darauf schließen, dass keine rezenteren Lexikalisierungen hinzugekommen sind bzw. keine Einträge vorhanden sind, die in der gedruckten Version, deren 12. und letzter Inhaltsband 1993 erschien, fehlen.

5.4.1 Qualitätssicherung: Abgleich mit der gedruckten Ausgabe

Um die Verlässlichkeit der verwendeten Daten sicherzustellen und das Ausmaß von Abweichungen in angemessenem Umfang zu prüfen, wird ein vereinfachtes Stichprobenverfahren angewandt.[91]

Dabei soll die Übereinstimmung der Einträge mit der gedruckten Ausgabe als Merkmal untersucht werden, Merkmalsträger für die Stichprobe sind die Wort- und Zeicheneinträge *mit Belegstellen*, mit einer Grundgesamtheit von 323.321 Einträgen.[92] Wesentlich für das hier verfolgte Unterfangen sind die erste bzw. älteste darin angegebene Primärquelle. Wie bei solchen Verfahren üblich soll eine Sicherheit von 95 % erreicht werden, woraus sich ein Wert für das Quantil der Standardnormalverteilung z von 1.96 ergibt.[93] Bei einer ersten Sichtung von 36 Einträgen weisen nur zwei Einträge (ca. 6 %) minimale Abweichungen auf, der Anteilswert P liegt

86 *DHYDCD*.
87 Siehe S. 117.
88 Zu den Angaben siehe den Vergleich von Yang Lin 杨琳 2011, *passim*.
89 Luo Zhufeng 羅竹風, Hrsg. 2005: *Hanyu da cidian* 漢語大詞典 *UTF-8 (Großes Wörterbuch der chinesischen Sprache, Unicode-Version)*. Shanghai 上海. URL: `http://bbs.gxsd.com.cn/forum.php?mod=viewthread&tid=498015` (besucht am 13. 01. 2013).
90 z. B. *Renmin ribao* 人民日报 1992. Siehe *HYDCD*, Bd. 12, S. 421, 體壇; bzw. *DHYDCD*, 體壇.
91 Die Methodik ist den Prognosen nach Bundes- oder Landtagswahlen entlehnt, die auf Befragungen der Wähler:innen basieren. Der Vorteil einer solchen Herangehensweise gegenüber einer Vollerhebung liegt nicht nur im deutlich reduzierten Aufwand, sondern auch in der Vermeidung von Erhebungsfehlern, die bei letzterer „bedingt durch den hohen Aufwand häufig" auftreten. Siehe Göran Kauermann und Helmut Küchenhoff 2010: *Stichproben – Methoden und praktische Umsetzung mit R*. Berlin & Heidelberg: Springer. DOI: 10.1007/978-3-642-12318-4, S. 1, S. 6.
92 Der Begriff der Grundgesamtheit beschreibt die „Menge aller Individuen oder Objekte, über die eine Aussage getroffen werden soll". Merkmalsträger sind „die Einheiten oder Objekte, an denen Untersuchungen, Messungen oder Beobachtungen vorgenommen werden"; Merkmale „die Eigenschaften […], die untersucht, […] werden sollen." ebd., S. 5, vgl. auch S. 29.
93 Siehe ebd., S. 28.

daher bei 10 % oder weniger.[94] Strebt man eine Genauigkeit e von wenigstens 0,05 an, ergibt sich folgende Berechnung der benötigten Stichprobengröße:[95]

$$n \geq \frac{P(1-P)}{e^2/z^2 + P(1-P)/N}$$

$$\frac{0,1 \times 0,9}{0,05^2/1,96^2 + 0,1 \times 0,9/323321} \approx 138,24$$

Eine Stichprobe von 139 Einträgen reicht also aus, um mit 95 %iger Sicherheit den Anteil der von der gedruckten Ausgabe abweichenden Einträge mit einer Genauigkeit von 5 % bestimmen zu können.[96] Die zur Durchführung verglichenen Einträge werden zufällig ausgewählt.[97] Der manuelle Abgleich wird anhand der folgenden Einträge durchgeführt (hier sortiert nach ihrem Vorkommen im *DHYDCD*):

> 一箭道、不期、不論、乾精、亂坟崗、伐棠、休光、佚怠、保佐、偏惰、偏比、傾顛、僞戾、六神無主、出入人罪、剛鍱、剪剪、劍鐔、南沃沮、吹火、周盈、噦吐、圓紗、外轉、大朝覲、奭發、好勇、季王、宸注、尊便、小半、尼軻、層、幔城、庵、弓勢、彝典、怒譴、思如湧泉、怨憎、恃怙、悲筑、慕容、掩滅、捷毒、收煞、收身、收過、放大炮、斃倫、舅、昌羊、春風面、昭飾、智將、曲致、本情、杜口裹足、杯水候、柮、桃葉女、械索、梳雲、椎擊、榆莢錢、機勇、機暇、款、此以、殷劉、每事問、毛胚、氣苦、混一、清狂、溪蓀、無人問津、煙册、牆頭馬上、王鳩、異謀、瞠惑、矚、積秀、突地吼、第恐、笳管、絶腸、綱梢、老是、肉裏錢、胡顏、膽悸、茁茂、茹古涵今、荒率、虞曹、蟻垤、蠢然、要實、覆地翻天、規約、親妮、訓言、講切、警俊、資、賠餉、賡歌、跳鱗、踏蹬、輪轉椅、辟摽、過動、選序、遺寇、配匹、酸溜溜、金衣丹、鉅萬、閔隔、陵誑、雅奏、集裝箱、難分難捨、雷火、靈湖、青粱、鞴靼、食官、餔糜、餘印、馬隊、驍銳、鶴瘦、鶻、鷙膺、麗象、龍馭。

Ergebnisse der Stichprobenanalyse

Insgesamt weisen 8 Einträge aus der Stichprobe (5,76 %) nennenswerte inhaltliche Abweichungen auf, meist in Form von fehlenden Belegstellen. Bedeutsam für die chronologische Einordnung ist dabei lediglich ein einziges fehlendes Zitat im Eintrag 資 ² (*zi*),[98] da hierdurch die älteste Belegstelle abweicht (siehe unten). Die relevante Abweichung liegt also unter einem Prozent. Die verwendete Ausgabe kann damit als hinreichend verlässlich angesehen werden.

Tabelle 5.1 gibt einen Überblick über das Ergebnis der Stichprobenanalyse. In der Spalte *Hochrechnung* ist dabei die theoretische, hochgerechnete Gesamtanzahl der Einträge angegeben, auf die das jeweilige Merkmal zutrifft, wenn man von 323.321 relevanten Einträgen ausgeht.
Im Verlauf der Analyse können im Detail folgende formale bzw. typographische und inhaltliche Unterschiede zwischen *HYDCD* und *DHYDCD* festgestellt werden:

94 Die Bestimmung des Anteilswerts mittels einer Pilotstichprobe ist in der Statistik nicht unüblich. Siehe dazu ebd., S. 39.
95 Ebd., S. 41.
96 Die Repräsentativität „typischer" Einträge kann hier außer Acht gelassen werden, da keine Anzeichen vorliegen, dass bestimmte Typen von Einträgen sich stärker als andere zwischen den verglichenen Ausgaben unterscheiden. Vgl dazu. ebd., S. 8f.
97 Für die Auswahl wurden Einträge aus der *SQL*-Datenbank (siehe dazu Kapitel 5.5, ab S. 120) selektiert und zufällig „sortiert" (`[…] order by rand() limit 139`).
98 Zur Nummerierung von Einträgen zu Zeichen mit unterschiedlichen Aussprachen siehe auch den Abschnitt zur Struktur (5.5.1, ab S. 121).

Tabelle 5.1 Qualität der digitalen Ausgabe – Ergebnisse der Stichprobenanalyse

Merkmal	Einträge	Anteil (von 139)	Hochrechnung
Älteste Belegstelle weicht ab	1	0,72 %	2.326
Fehlende Belegstellen	7	5,03 %	16.282
Inhaltliche Abweichung	8	5,76 %	18.608

— 1. **Belegstellen**. Wie bereits angedeutet fehlen in der digitalen Ausgabe einzelne Quellenzitate. Dies trifft gleichermaßen auf Zeichen-[99] und Worteinträge zu.[100] Auch inhaltliche Unterschiede in zitierten Belegstellen können festgestellt werden.[101]

Lediglich in einem einzigen untersuchten Eintrag weicht – wie bereits erwähnt – die älteste Quellenangabe ab: Im Eintrag von *zì* 資 2 gibt die digitale Ausgabe lediglich eine Belegstelle aus dem Gedicht *Huashan nü* 華山女 von Han Yu 韩愈 (768–824) an,[102] während die gedruckte Ausgabe noch zwei deutlich ältere Belegstellen aus dem *Han shu* 漢書 und dem *Shiji* 史記 enthält.[103] Allerdings finden sich identische, ältere Belegstellen im Eintrag *zì* 資 1,[104] so dass dieser konkrete Fall [zufälligerweise] keine Auswirkungen auf die erzeugten Daten hätte, da graphisch gleiche Zeichen in *Plain Text*-Daten nicht unterschieden werden können.

— 2. **Typographische Markierungen**. In der gedruckten Ausgabe sind Personen-, Dynastie- und Ortsnamen unterstrichen. Durch Unterbrechungen können Dynastie- und Personennamen hier klar unterschieden werden, z. B. 宋 穆休.[105] Da *Plain-Text* keine Formatierungen enthalten kann, fehlen diese Informationen in der hier verwendeten digitalen Ausgabe vollständig, während sie in der offiziellen CD-Rom Version vorhanden sind. Dadurch wird an einigen Stellen beim Parsen der Daten die Unterscheidung, ob eine Quellenangabe im Format *DynastieNachnameVorname*, oder lediglich *NachnameVorname* vorliegt, erschwert.[106]

— 3. **Zitierweise**. Wird mehrmals in Folge dieselbe Quelle zitiert, wird das Zitat in der gedruckten Ausgabe bei den Folgeangaben mit *you* 又 („erneut“) eingeführt, etwa im Eintrag zu *yidai* 佚怠, in welchem zwei Stellen aus dem *Yanzi Chunqiu* 晏子春秋 zitiert werden.[107] Die digitale Ausgabe wiederholt die vollständige Quellenangabe, was die Extraktion dieser Daten erleichtert.

— 4. **Gruppierung und Nummerierung von Untereinträgen**. In der gedruckten Ausgabe werden unterschiedliche Wortbedeutungen, sofern zutreffend, nach syntaktischen Kategorien, z. B. *lianci* 連詞 (Konjunktion), gruppiert. Die Kategorien werden dabei mit eingekreister Nummerierung (❶, ❷, ❸…) markiert, die Unter-untereinträge mit einfachen Klammern (3).

99 z. B. zu *rui* 枘 gibt die gedruckte Ausgabe zur zweiten Bedeutung ein *Tang*-zeitliches Zitat an, das in der digitalen Ausgabe fehlt. *HYDCD*, Bd. 4, S. 854; *DHYDCD*, 枘. Ebenfalls nur in der Papierversion enthält der Eintrag *fu* 甫 ein *Qing*-zeitliches Zitat. *HYDCD*, Bd. 6, S. 1599; *DHYDCD*, 甫.

100 Im digitalen Eintrag zu *gengge* 賡歌 fehlen zwei Quellenzitate. *DHYDCD*, 賡歌; *HYDCD*, Bd. 10, S. 275.

101 In beiden Ausgaben wird im Eintrag zu *longyu* 龍馭 eine Stelle aus dem Gedicht *Yu dong jun* 喻東軍 von Wei Zhuang 韋莊 (*ca.* 836–910) in jeweils unterschiedlicher Fassung wiedergegeben. „四年龍馭守峨眉，到此躊躇不能去“ lautet in der digitalen Ausgabe „四年龍馭守峨嵋，鐵馬西來步步遲“ – vermutlich handelt es sich um eine Korrektur in der neueren Ausgabe. *HYDCD*, Bd. 12, S. 1481; *DHYDCD*, 龍馭.

102 *DHYDCD*, 資 2.

103 *HYDCD*, Bd. 10, S. 200.

104 *DHYDCD*, 資 1; *HYDCD*, Bd. 10, S. 199.

105 Siehe *HYDCD*, Bd. 2, S. 1224.

106 Auf diese Problematik wird in Abschnitt 5.5.2, S. 130, genauer eingegangen.

107 Siehe *HYDCD*, Bd. 1, S. 1244.

Die digitale Ausgabe nimmt nur eine einstufige Nummerierung vor, die dadurch häufig abweicht.[108] Bei manchen Einträgen entfällt in der gedruckten Ausgabe die Nummerierung und die zweite bzw. weitere Bedeutungen werden mit *yi zhi* 亦指…(„deutet auch auf…“) oder *yinshen wei* 引申为 (etwa: „eine erweiterte Bedeutung ist…“) eingeleitet.[109] In der digitalen Ausgabe wird in allen Fällen konsequent mit „1., 2., 3.…“ nummeriert, so dass die Abschnitte mit den unterschiedlichen Bedeutungen einfach zu segmentieren sind.

— 5. **Strichzahl des zweiten Zeichens**. In der gedruckten Ausgabe wird diese bei jeder Erhöhung durch eine hochgestellte Zahl ausgewiesen, also z. B. [12]【伐棠】,[110] wobei 12 die Anzahl der Striche von *tang* 棠 angibt. In der digitalen Ausgabe fehlen solche Angaben vollständig.

— 6. **Querverweise** auf andere Worteinträge zeigen in der gedruckten Ausgabe stets auf den relevanten Untereintrag.[111] Auch kleine inhaltliche Unterschiede in den Querverweisen kommen vor – dabei scheint aber nicht eine der beiden Ausgaben genauer zu sein, sondern schlicht beide minimal unterschiedlich.[112]

— 7. In der gedruckten Ausgabe wird bei nicht eindeutiger **Aussprache** des zweiten (dritten, usw.) Zeichens eines Worteintrages die Lesung dieser Zeichen explizit in *Hanyu Pinyin* 漢語拼音 angegeben.[113] In der digitalen Ausgabe fehlen solche Angaben leider.

— 8. Im *DHYDCD* wird ***Zhuyin Fuhao*** 注音符號 („*Bopomofo*“) zur Angabe der Aussprache zusätzlich angebeben, in der gedruckten Ausgabe lediglich das festlandchinesische *Hanyu Pinyin*.

Durch die einfachere, konsequentere Struktur der Untereinträge und die stets vollständigen Quellenangaben ist der Text der digitalen Ausgabe insgesamt leichter maschinenlesbar und damit sogar besser für die Analyse geeignet als der ursprüngliche Text. Für eine Extraktion der Wortklassen aus den Kategorien (siehe 3.) – als Möglichkeit zur Gewinnung von Daten zum *Part-of-Speech Tagging* – wären auch die Angaben in der gedruckten Ausgabe zu unvollständig und unsystematisch.

Zur Veranschaulichung sei an dieser Stelle ein Beispiel aus der gedruckten Ausgabe wiedergegeben:[114]

108 Vgl. z. B. die Einträge zu *bulun* 不論 *HYDCD*, Bd. 1, S. 468; sowie *ceng* 層 *HYDCD*, Bd. 4, S. 60.

109 Siehe z. B. in den Einträgen zu *jianyin* 劍鐔 und *kuan* 款. *HYDCD*, Bd. 2, S. 753, Bd. 6, S. 1444.

110 *HYDCD*, Bd. 1, S. 1190, *fatang* 伐棠.

111 Siehe z. B. im Eintrag *yujiaqian* 榆莢錢: „参见“榆莢❷”“ („siehe *yujia* ❷“) – in der digitalen Ausgabe wird lediglich auf 榆莢 verwiesen. Vgl. *HYDCD*, Bd. 4, S. 1188.

112 Nur die digitale Ausgabe weist im Eintrag zu *ceng* 層 darauf hin, dass dieses Zeichen mit dem homophonen *ceng* 增 austauschbar verwendet werden kann. Siehe *DHYDCD*, 層; *HYDCD*, Bd. 4, S. 60. Umgekehrt verweisen gedruckte wie digitale Ausgabe im zweiten Eintrag zu 增 auf 層. Siehe *DHYDCD*, 增; *HYDCD*, Bd. 2, S. 1222. Während die gedruckte Ausgabe im Eintrag zu *zi* 資 auf das „gleiche“ Zeichen 恣 verweist, fehlt diese Information in der digitalen Ausgabe. Vgl. *HYDCD*, Bd. 19, S. 200; *DHYDCD*, 資.

113 Siehe z. B. im Eintrag zu *zūnbiàn* 尊便 die Angabe „– biàn“ vor der Angabe der Bedeutung. *HYDCD*, Bd. 2, S. 1283. 便 kann, je nach Kontext bzw. Bedeutung auch *pián* oder *biān* gelesen werden.

114 *HYDCD*, Bd. 7, S. 986.

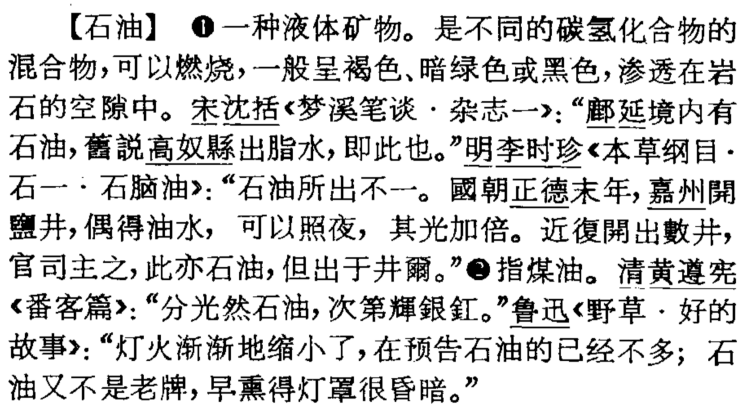

【石油】 ❶一种液体矿物。是不同的碳氢化合物的混合物，可以燃烧，一般呈褐色、暗绿色或黑色，渗透在岩石的空隙中。宋沈括《梦溪笔谈·杂志一》："鄜延境内有石油，舊説高奴縣出脂水，即此也。"明李时珍《本草纲目·石一·石脑油》："石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。近復開出數井，官司主之，此亦石油，但出于井爾。"❷指煤油。清黄遵宪《番客篇》："分光然石油，次第輝銀釭。"鲁迅《野草·好的故事》："灯火渐渐地缩小了，在预告石油的已经不多；石油又不是老牌，早熏得灯罩很昏暗。"

Abbildung 5.1 Eintrag *shiyou* 石油 („Steinöl", Erdöl) in der Originalausgabe des *HYDCD*.

In der digitalen Version ist der gleiche Eintrag enthalten – wobei ein Teil der oben beschriebenen typographischen Vereinfachungen sichtbar wird:

> 【石油】1. 一种液体矿物。是不同的碳氢化合物的混合物，可以燃烧，一般呈褐色、暗绿色或黑色，渗透在岩石的空隙中。宋沈括《梦溪笔谈·杂志一》："鄜延境内有石油，舊説高奴縣出脂水，即此也。"明李时珍《本草纲目·石一·石脑油》："石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。近復開出數井，官司主之，此亦石油，但出于井爾。"2. 指煤油。清黄遵宪《番客篇》："分光然石油，次第輝銀釭。"鲁迅《野草·好的故事》："灯火渐渐地缩小了，在预告石油的已经不多；石油又不是老牌，早熏得灯罩很昏暗。"[115]

Eine offensichtliche Schwäche der hier durchgeführten Stichprobe ist ihre Auswahl aus den Einträgen des *DHYDCD*, denn darin fehlende Zeichen- und Worteinträge bleiben unbemerkt. Das trifft z. B. auf die Zeichen *bing* 丙 und *mei* 美 und die zugehörigen Worteinträge zu.[116] Ob noch weitere Einträge fehlen, die in der gedruckten Ausgabe vorhanden sind, ist nur mit unverhältnismäßigem Aufwand feststellbar.[117] Auf die inhaltliche Qualität der verbleibenden – und davon abgesehen augenscheinlich auch vollständigen – Daten hat dies jedoch keinen Einfluss.

5.5 Erzeugung einer diachronen Lexemdatenbank

Um die Lexikalisierungsdaten aus dem *DHYDCD* nutzbar zu machen, wird dieses in eine *SQL*-Datenbank umgeformt und anschließend mit weiteren Informationen angereichert. Die Strukturierung als relationale Datenbank ermöglicht es, die Daten bei minimaler Redundanz und guter Nachvollziehbarkeit zu erweitern und später zielgenau effizient abzufragen. Dafür werden die Quellen der Belegstellen extrahiert und – soweit ermittelbar – zur chronologischen Einordnung der Lexeme der Entstehungszeitraum bzw. -zeitpunkt des ältesten

115 *DHYDCD*, *shiyou* 石油.

116 Siehe *HYDCD*, Bd. 1, 丙, S. 509–510 , Bd. 9, 美, S. 158–164. 13 Worteinträge zu *bing*, sowie 137, die mit *mei* beginnen, fehlen ebenfalls. Vgl. *DHYDCD*.

117 Das Fehlen der Einträge zu *mei* 美 und *bing* 丙 folgt keiner erkennbaren Logik. Die vorherigen und nachfolgenden Einträge zu *qie/ju/zu/cu* 且 und *qiu* 丘 bzw. *da* 夆 und *qiang* 羌 sind in beiden Ausgaben vorhanden. Eine Möglichkeit, Kandidaten für im *DHYDCD* fehlende Einträge systematisch aufzuspüren ist es, Zeichen zu ermitteln, die zwar in Worteinträgen verwendet werden, aber keinen eigenen Zeicheneintrag haben. Das trifft auf insgesamt 216 Zeichen zu, unter denen sich aber etliche Varianten befinden, z. B. *kuai* 凷 (für 塊) und *chu* 出 (für 出).

zitierten Texts verwendet. Zur strukturierten Extraktion der Inhalte kommen dabei überwiegend **Reguläre Ausdrücke** (*Regular Expressions*, kurz *RegEx*) zum Einsatz. Sie sind ein in vielen Programmiersprachen verbreitetes syntaktisches Konzept zur Beschreibung von Mustern und werden eingesetzt, um bestimmte Informationen aus Texten zu extrahieren, zu suchen oder zu ersetzen.[118] Aufbau und Erstellung dieser historischen Lexemdatenbank sind im Folgenden dokumentiert.

1. Segmentierung der Rohdaten in Zeichen- und Worteinträge.
2. Erkennen der Belegstellen in diesen Einträgen und Interpretation der zugehörigen Metadaten, d. h. Titel, Autor und Entstehungszeit des zitierten Textes.
3. Die oft unvollständigen oder ungenauen Metadaten werden mit externen Datenquellen verdichtet.

5.5.1 Datenstruktur

In der vorliegenden digitalen Ausgabe wird jeder Zeicheneintrag (*dan zi tiaomu* 單字條目) mit einem Asterisk eingeleitet (z. B. „* 漢“, so dass die Einträge mithilfe des regulären Ausdrucks `(\*\p{IsHan})` segmentiert werden können. Die so getrennten Haupteinträge lassen sich in zwei Arten von Untereinträgen unterteilen: mehrsilibige Lexemeinträge, sowie die unterschiedlichen Lesungen der Zeicheneinträge bei *duoyinzi* 多音字. Mehrsilbige Lexemeinträge sind stets an „gefüllten quadratischen Klammern“ 【 】 (*shixin fangtou kuohao* 實心方頭括號) erkennbar.

> 【且 2 末】汉代西域国名。《汉书 · 西域传上 · 且末国》: “且末國，王治且末城，去長安六千八百二十里。”地在今新疆且末县。[119]
> 【且並】并且。清和邦额《夜谭随录 · 诡黄》: “驚惶間已失鞋，且並脱去一襪。”[120]

Die Untereinträge für *duoyinzi* lassen sich an den eckigen Klammern ［ ］erkennen, in welchen die Aussprache angegeben wird, z. B.

> 且 2［jū ㄐㄩ］［《廣韻》子魚切，平魚，精。］1. 多貌。《诗 · 大雅 · 韩奕》: “籩豆有且，侯氏燕胥。” …[121]

Der reguläre Ausdruck `(【[^】]+】|\p{IsHan} [0-9]［.+］)` [122] beschreibt die somit möglichen Markierungen von Untereinträgen.

Da für spätere Verarbeitungsschritte etliche Besonderheiten berücksichtigt werden müssen, werden die Daten mit den obigen regulären Ausdrücken in *Python* segmentiert und direkt in die benötigte Datenbankstruktur geschrieben.[123] In diesem Rahmen finden zusätzliche Verarbeitungsschritte statt:

---

118 *RegEx* finden häufig auch im kommerziellen Kontext Verwendung, z. B. für Formularvalidierungen. Soll zum Beispiel eine Kundin in einem Webshop die IBAN-Nummer einer deutschen Bankverbindung angeben, kann der Anbieter prüfen, ob die Eingabe dem Muster `^DE(?:[ ]?[0-9]){20}$` entspricht. Eine Zeichenkette, die mit „DE“ beginnt, gefolgt von zwanzig Ziffern von 0 bis 9, zwischen denen einzelne Leerzeichen zugelassen sind. Die Existenz oder gar Deckung des Kontos lässt sich damit sicherlich nicht absichern, wohl aber, ob die Kundin passenden Inhalt in das vorgesehene Feld eingibt.

119 *DHYDCD*, 且 2 末. Farbliche Markierungen gemäß Übereinstimmung mit dem Muster der verwendeten regulären Ausdrücke.

120 *DHYDCD*, 且並.

121 *DHYDCD*, 且 2.

122 【, gefolgt von allen Zeichen, die nicht „】“ sind, bis 】 *oder* alternativ: Ein einzelnes *Hanzi* 漢字, eine Ziffer, gefolgt von mind. einem beliebigen Zeichen in eckigen Klammern ［ ］.

123 Ein *Python*-Script verarbeitet die 365.102 Einträge in etwa 90 Sekunden (knapp 4.000 Einträge pro Sekunde).

1. Alle Zeilenumbrüche und Leerzeichen[124] werden entfernt, da sie innerhalb von Einträgen lediglich das Ende einer Zeile von max. 50 Zeichen bzw. etwa 44 chinesischen Schriftzeichen markieren und für die Inhaltsanalyse hinderlich sind.
2. Extraktion der Aussprache- und Reiminformationen aus den Zeicheneinträgen, ebenfalls mit regulären Ausdrücken.
   a) *Pinyin*-Aussprache: `aa\u0101\u00E0\u00E1\u01CEe\u0113\u00E9\u011B\u00E8i\u012B\u00ED\u01D0\u00ECo\u014D\u00F3\u01D2\u00F2uū\u016B\u00FA\u01D4\u00F9v\u00FC\u01D6\u01D8\u01DA\u01DCbcdfggkhjlmmm̄m̀ńňǹpqrstwxyz`.[125]
   b) *Zhuyin*-Aussprache: `([ㄅㄆㄇㄈㄉㄊㄋㄌㄍㄎㄏㄐㄑㄒㄓㄔㄕㄖㄗㄘㄙㄚㄛㄜㄝㄞㄟㄠㄡㄢㄣㄤㄥㄦㄧㄨㄩ市\u02C7\u02CA\u02CB\u02D9]+)`.[126]
   c) Angaben über Reime wie „`［《廣韻》子魚切，平魚，精。］`“[127] entsprechen stets dem regulären Ausdruck `(［《.+》.+］)`.

Segmentierung des *Plain Text*-Wörterbuchs in Datenbanktabellen

Der Aufbau der Datenbank spiegelt die oben beschriebene Struktur des *HYDCD* mit Zeichen- und Worteinträgen wider, die zunächst in zwei entsprechende Tabellen geschrieben werden:

— 1. `hydcd_entries` – Ablage der Zeicheneinträge (Haupteinträge) mit ihrem vollständigen Inhalt (inkl. aller untergeordneten Worteinträge).

- `id` – Eindeutige ID des Zeicheneintrags, nach Radikal und Zusatzstrichzahl sortiert. Der erste Eintrag, *yī* 一 erhält die ID 1, der letzte, *guī* 龜, die ID 16.358.
- `char` – Das Schriftzeichen, auf das sich der Inhalt des Eintrags bezieht.
- `entry` – Der vollständige Inhalt des Eintrags, inklusive aller Untereinträge (Wort- und Zeicheneinträge).

— 2. `hydcd_words` – Eine Datenbankzeile pro Zeichen- *und* Lexemeintrag, d. h. unterschiedliche Lesungen bei Zeicheneinträgen (*duoyinzi* 多音字, Heteronyme) werden gesondert gelistet. So kommen z. B. auf einen Eintrag für *qiě* 且 (ID 30 in der Tabelle `hydcd_entries`) die vier Einträge *qiě* 且, *jū* 且, *zǔ* 且 und *cú* 且, sowie 35 polysyllabische Lexeme, die mit 且 beginnen.

- `id` – Eindeutige ID des Worteintrages.
- `id_internal` – Interner Zähler für Untereinträge des Zeicheneintrages.
- `char_id` – ID des übergeordeten Zeicheneintrags in der Tabelle `hydcd_entries`.
- `word` – gibt das Lemma im ursprünglichen Format in Klammern 【】 mit dem Verweis auf die Lesung des führenden Zeichens an, also z. B. 【且 2 末】, wobei die „2“ signalisiert, dass 且 hier *jū* gelesen wird.
- `cleanword` – das Schlagwort ohne Zusatzinformationen, z. B. 且末.
- `nakedword` – das Schlagwort ohne eventuell vorhandene Interpunktionszeichen.

124 D. h. `\n, \t, \f, \r, \v` und das Leerzeichen.
125 Erweiterung des regulären Ausdrucks aus Thomas Roten 2017: *zhon - Constants used in Chinese text processing*. URL: `https://github.com/tsroten/zhon` (besucht am 10. 11. 2017), `pinyin.py`. Zusätzlich zu den im *Pinyin*-Standard vorgesehenen Zeichen sind hier die im *DHYDCD* verwendeten Sonderformen berücksichtigt.
126 Alle verfügbaren *Zhuyin*-Zeichen und die Unicode-Repräsentationen der verwendeten Diakritika für Töne. Der erste Ton wird in *Zhuyin* nicht graphisch wiedergegeben. Basierend auf ebd., `zhuyin.py`.
127 *DHYDCD*, 且 2.

- `zhuyin`, `pinyin` – gibt die Aussprache jeweils in *Zhuyin* 注音 (z. B. ㄐㄩㄇㄛˋ) und *Hanyu Pinyin* 漢語拼音 (z. B. *jūmò*) an.[128]
- `rhyme` – gibt, sofern vorhanden, für Zeicheneinträge Reim-, Ton und *Fanqie* 反切 Informationen oder Schriftzeichen mit gleicher Lesung, sowie die Quelle der jeweiligen Angabe an, z. B. ［《廣韻》子魚切，平魚，精。］.[129]
- `entry` – Der Inhalt des Worteintrags.
- `entrytype` – *C* für einzelne Zeichen (monosyllabische Wörter), *W* für Worteinträge (polysyllabische Wörter).

Tabellen der diachronen Lexemdatenbank

Aus den so strukturierten Daten werden nun die Lexikalisierungsdaten der Worteinträge extrahiert. Im Beispieleintrag zu *shiyou* 石油[130] werden zu zwei Bedeutungen (unten markiert in rot) Erklärungen gegeben (hier ausgegraut). Zu jeder Bedeutung werden zudem entsprechende Belege aus der Literatur (schwarz) in Anführungsstrichen “” zitiert. Diese werden stets mit einer vereinfachten bibliographischen Angabe eingeführt, bestenfalls im Format DynastieAutor 《Werk · Kapitel》 und sind in der Regel chronologisch sortiert. Da Unterstreichungen im *DHYDCD* fehlen, sind sie zu Illustrationszwecken aus der Originalausgabe übernommen:

> 【石油】
> 1. 一种液体矿物。是不同的碳氢化合物的混合物，可以燃烧，一般呈褐色、暗绿色或黑色，渗透在岩石的空隙中。 宋 沈括《梦溪笔谈 · 杂志一》:“鄜延境内有石油，舊説高奴縣出脂水，即此也。” 明 李时珍《本草纲目 · 石一 · 石脑油》:“石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。近復開出數井，官司主之，此亦石油，但出于井爾。”
> 2. 指煤油。 清 黄遵宪《番客篇》:“分光然石油，次第輝銀釭。” 鲁迅《野草 · 好的故事》:“灯火渐渐地缩小了，在预告石油的已经不多；石油又不是老牌，早熏得灯罩很昏暗。”[131]

Der Begriff *shiyou* 石油 mit der Bedeutung „eine Art flüssiges Mineral“ (*„yi zhong yeti kuangwu* 一种液体矿物“) ist also spätestens in der Song 宋-Zeit (960–1279) belegt und im *Meng xi bi tan* 夢溪筆談 („Pinselunterhaltung am Traumbach“[132]) von Shen Kuo 沈括 (1031–1095) zu verorten. Im Optimalfall handelt es sich bei dieser Angabe um den *Locus classicus*, was aber nicht letztgültig geklärt werden kann. Unabhängig davon, ob Shen Kuo den Begriff wirklich geprägt oder sogar erfunden hat, ist gesichert, dass *spätestens* zu dieser Zeit der Begriff bereits verwendet wurde.

Auch aus dem Ming 明-zeitlichen (1368–1644) *Ben cao gangmu* 本草綱目 („*Materia Medica, Arranged according to Drug Descriptions and Technical Aspects*“[133]) ist ein Zitat angegeben. Die zweite Bedeutung, „Lampenöl“ (Petroleum, *meiyou* 煤油), ist hier erst für die Qing 清-Zeit (1644–1911)

128 Die Aussprachеinformationen für mehrsilbige Einträge werden aus den Angaben in den Zeicheneinträgen des *DHYDCD* zusammengesetzt. Alternative Zeichenlesungen können dabei leider nur für das jeweils erste Zeichen eines mehrsilbigen Ausdrucks berücksichtigt werden, da für die nachfolgenden Zeichen im Gegensatz zur gedruckten Ausgabe kein Hinweis auf die Aussprache gegeben wird (siehe auch Abschnitt 5.4.1, S. 117). Für die „hinteren“ Zeichen wird hier daher immer die erste Lesung angenommen.

129 *Fanqie* ist eine traditionelle Methode zur phonetischen Analyse – die Lesung eines Zeichens wird zu diesem Zweck mit zwei weiteren Zeichen repräsentiert, von denen das erste Zeichen den An-, das zweite den Auslaut angibt. Im Beispiel werden also *zi* 子 und *yu* 魚 zu *ju* „geschnitten“ (*qie* 切). Die Kompilator:innen haben hier zumeist Informationen aus den song-zeitlichen Reimwörterbüchern *Guangyun* 廣韻 bzw. *Jiyun* 集韻 angegeben. Vgl. *HYDCD*, *passim*.

130 Siehe auch S. 120.

131 *DHYDCD*, 石油. Unterstreichung und farbliche Hervorhebungen durch den Verfasser.

132 Shen Kuo 沈括 1997 [1088]: *Pinselunterhaltungen am Traumbach. übs. von Konrad Herrmann.* München: Diederichs.

133 Paul Ulrich Unschuld 1986: *Medicine in China: A History of Pharmaceutics*. Comparative Studies of Health. Berkeley & Los Angeles: University of California Press, S. 145.

nachgewiesen, im *Fanke pian* 番客篇 („The Foreign Guest“[134]) von Huang Zunxian 黃遵憲 (1848–1905)[135], sowie später in Lu Xuns 魯迅 (1881–1936)[136] Prosagedichtsammlung *Ye Cao* 野草 („Wildes Gras“). Für die vorgesehene Anwendung der Datenbank ist die früheste Angabe zur Lexikalisierung über das *Meng xi bi tan* am wichtigsten.

Das Beispiel zeigt, dass chronologische Informationen im *HYDCD* sich zumeist (wenn überhaupt) auf die Angabe der Dynastie beschränken und damit vage bzw. implizit sind. Eine Liste mit Angaben zu den zitierten Werken bzw. der verwendeten Ausgaben fehlt im *HYDCD* zudem völlig, so dass solche Daten aus externen Quellen ergänzt werden müssen.

Zunächst werden die vorhandenen Lexikalisierungsdaten in drei weitere Datenbanktabellen strukturiert, die eine Verknüpfung von Lexem, Belegstellen, sowie zitierten Werken und damit eine (implizite) chronologische Einordnung der Lexeme ohne urheberrechtlich geschützte Inhalte enthalten.

Die Tabelle `the_words` soll alle Schlagworte enthalten, zu denen ein Textbeispiel angegeben ist,[137] sowie den Verweis auf die älteste angegebene Belegstelle. In `the_books` sind die verfügbaren Metadaten zu allen unterscheidbaren, im *DHYDCD* zitierten Texten enthalten und `the_citations` ermöglicht *n : m*-Verknüpfungen für *alle* in den Wörterbucheinträgen zitierten Quellen. Letzteres kann für die diachrone Betrachtung der Wortnutzung und die Datenabdeckung bzw. die Erzeugung eines diachronen Arbeitskorpus herangezogen werden.[138] Die wichtigsten Spalten der genannten Datenbanktabellen werden im Folgenden dokumentiert.

— 1. `the_words` – Alle Zeichen- und Worteinträge, die eine Belegstelle mit Quellenangabe aufweisen, mit Verknüpfung zur Quelle des ältesten zitierten Belegs.

- `id` – ID des Eintrags in der Tabelle `hydcd_words`.[139]
- `cleanword` – Das Lexem bzw. Schlagwort.
- `pinyin` – Die Aussprache in *Hanyu Pinyin* 漢語拼音.
- `firstentry` – Die erste, in dem Eintrag zitierte Primärquelle, inklusive möglicher, mit einem · abgetrennter, Kapitelangaben (z. B. „庄子 · 齐物论“).
- `unordered` – Markiert Einträge, bei in denen eine nicht chronologische Reihenfolge der zitierten Quellen vermutet wird (siehe unten).
- `indirectsource` – Markiert Einträge, in denen bei der ersten Bedeutung keine Quellenangabe gefunden wurde.
- `book` – Der Titel der frühesten zitierten Primärquelle, ohne Kapitelangaben.
- `book_id` – ID der zitierten Quelle in der Tabelle `the_books`.
- `earliest_evidence_id` – `book_id` derjenigen Quelle, die den ältesten in Korpusdaten gefundenen Beleg für die Zeichenkombination in `cleanword` enthält – unabhängig von der Angabe im *DHYDCD*. Diese Spalte wird genutzt, falls ältere Belegstellen gefunden werden können, als im *DHYDCD* angegeben sind.

134 Yang Zhiyi 楊治宜 2015: „The Modernity of the Ancient-Style Verse“. In: *Frontiers of Literary Studies in China* 9.4, S. 551–580, S. 554.
135 Raoul David Findeisen 2004: „Literatur im 20. Jahrhundert“. In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 288–395, S. 293.
136 Reinhard Emmerich 2004: „östliche Han bis Tang“. In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 88–186, S. 136.
137 Damit die erzeugten Daten ohne lizenzrechtliche Bedenken veröffentlicht werden können, sind die Einträge selbst nicht enthalten.
138 Siehe Kapitel 5.7, ab S. 138 bzw. 5.6, ab S. 137.
139 Vgl. Abschnitt 5.5.1, S. 122.

— 2. `the_books` – Alle unterscheidbaren im *DHYDCD* aufgeführten Primärquellen der Zitate aus den Einträgen in `the_words`. Unterscheidbar bedeutet dabei, dass sich eines der Kriterien Titel, Autor, Dynastie *oder* Erscheinungsjahr bzw. einer der Werte in den Datenspalten `clearbook`, `author`, `dynasty` oder `startyear` unterscheidet. Nur wenn *alle* diese Angaben identisch sind, wird davon ausgegangen, dass aus derselben Quelle zitiert wird.[140]

Datenmodell: $1:n$ – jede Primärquelle kann in n Worteinträgen die erste bzw. älteste Quellenangabe (*Locus classicus*) sein. Aufgrund der uneinheitlichen Zitierweise im *HYDCD* ist es bei mehrfach zitierten Quellen teilweise unvermeidlich, dass unter einer ID in `the_books` nicht alle Nennungen zusammengeführt werden, die sich *de facto* auf dasselbe Werk beziehen.[141]

- `id` – Eindeutige ID der Primärquelle, Reihenfolge wie im *DHYDCD*.
- `clearbook` – Der Titel der zitierten Primärquelle. Hier sind aufgrund der uneinheitlichen Zitierweise Duplikate möglich.
- `cbdb_text_id` – ID des Textes in der *CBDB*, falls ermittelbar.[142]
- `title_py` – Die Aussprache des Titels in *Hanyu Pinyin*.
- `title_western` – Englische Übersetzung des Titels, falls vorhanden.
- `startyear` – Frühestmögliches ermitteltes Entstehungsjahr des Textes.
- `endyear` – Spätestmögliches ermitteltes Entstehungsjahr des Textes.
- `dynasty` – Name der Dynastie, während der das Werk entstanden ist.
- `estimate` – Ungenaue Angaben zu Jahr bzw. Zeitraum der Veröffentlichung, bzw. Schätzungen werden mit 1 gekennzeichnet.
- `usecount` – Zähler, wie häufig der Text insgesamt im *DHYDCD* zitiert wurde.
- `useinfirstcount` – Zähler, wie häufig der Text als ältester Beleg im *DHYDCD* angegeben wurde.
- `source` – Quelle der ermittelten Metadaten in den Feldern `startyear`, `endyear` und `author`.[143]
- `author` – Autor:in des Textes, sofern angegeben bzw. ermittelbar.

— 3. `the_citations` – Verortung *aller* Belegstellen im *DHYDCD*. Datenmodell: $n:m$ – zu jedem Schlagwort in `the_words` können n Zitate mit Quellenangabe vorhanden sein. Jedes dabei zitierte Werk aus `the_books` kann in m Einträgen verwendet werden. Mithilfe dieser Tabelle können sowohl alle in einem Wörterbucheintrag zitierten Texte selektiert werden, als auch alle Einträge gefunden werden, in denen bestimmte Texte zitiert werden.

- `id` – Eindeutige ID des Quellenzitats, Reihenfolge entspricht dem *DHYDCD*.

140 Es wird z. B. aus Gedichten mit dem naheliegenden Titel *Qiu ye* 秋夜 („Herbstnacht“) von insgesamt zwölf Urheber:innen zitiert, die während zehn unterschiedlichen Dynastien über einen Zeitraum von insgesamt 1.646 Jahren gelebt haben, von der Jin 晋 bis zur Qing 清-Zeit. Dies ist zwar ein eher extremes Beispiel, aber es finden sich insgesamt über 3.000 Quellen, bei denen zwei Texte gleichen Namens in unterschiedlichen Dynastien verfasst wurden und unbedingt unterschieden werden müssen. In `the_books` nicht unterschieden werden Quellenangaben, bei denen aus demselben Werk, aber aus unterschiedlichen Kapiteln bzw. Abschnitten oder *juan* 卷 zitiert wird. So werden etwa die Referenzen zum *Liexian zhuan* 列仙传 („Biographien beispielhafter Heiliger“) „ 汉刘向《列仙传 · 骑龙鸣》“ und „ 汉刘向《列仙传 · 方回》“ als eine einzige Primärquelle mit mehreren Verwendungen betrachtet. Vgl. *DHYDCD*, * 一, 一丸泥. Unterscheiden sich hingegen zwei Angaben weder im Titel noch in der Dynastie, aber im Autor, werden zwei separate Quellen registriert. So haben z. B. in der Jin 晋-Zeit zwei Historiker, Sun Chuo 孙绰 und Guo Yuanzu 郭元祖 die Biographien aus dem *Liexian zhuan* kommentiert: „ 晋孙绰《列仙传赞 · 老子》“ Siehe *DHYDCD*, 頤生; „ 晋郭元祖《列仙传赞 · 方回》“ *DHYDCD*, 冥神.

141 Siehe dazu Abschnitt 5.5.2, ab S. 128.

142 Siehe dazu Abschnitt 5.5.3, ab S. 132.

143 Ebd.

- `word_id` – ID des Worteintrags, in dem das Zitat gefunden wurde – referenziert auf `the_words.id`.
- `sub_id` – interne ID des Untereintrags, bei mehreren gelisteten Bedeutungen.
- `book_id` – ID der Primärquelle, aus der das Quellenzitat stammt – referenziert auf `the_books.id`.

Zur Veranschaulichung sind die wichtigsten Datenspalten des oben beschriebenen Modells am Beispiel des Lexemeintrags zu *shiyou* 石油 (ID Nr. 208.670) illustriert (Abb. 5.2). Die Spalte `the_words.book_id` referenziert auf das Werk mit der frühesten Belegstelle: *Meng xi bi tan* (ID Nr. 96) – eine tatsächlich frühere Belegstelle (`earliest_evidence_book_id`) konnte nicht ermittelt werden. In der Tabelle `the_books` finden sich unter der referenzierten ID die entsprechenden Metadaten, inklusive dem Jahr (bzw. Zeitraum) der Veröffentlichung und, sofern vorhanden, der Autor, hier Shen Kuo 沈括. Die Daten konnten in diesem Fall aus der *CBDB* ergänzt werden, die IDs der entsprechenden Einträge stehen in `the_books.cbdb_text_id` bzw. `the_books.cbdb_author_id`.[144] Bei insgesamt 414 Lexemen ist die älteste angegebene Belegstelle ein Zitat aus dem *Meng xi bi tan* (`useinfirstcount`).

Die Tabelle `the_citations` gibt zudem die *n : m* Beziehung zwischen Wörterbucheinträgen und zitierten Texten an, so dass nicht nur das älteste, sondern *alle* in dem Eintrag angegebenen Zitate verortet werden können. Zusätzlich zur *Locus classicus*-Angabe werden noch Stellen aus dem *Ben cao gangmu*, *Fan ke pian* und *Ye cao* zitiert.[145]

the_words

id	208670
cleanword	石油
pinyin	*shíyóu*
book_id	96
earliest_evidence_book_id	*NULL*

the_books

id	96
clearbook	梦溪笔谈
cbdb_text_id	2206
startyear	1095
endyear	1095
dynasty	宋
useinfirstcount	414
source	CBDB
author	沉括
cbdb_author_id	1450

the_citations

word_id	book_id	the_books.clearbook
208670	96	梦溪笔谈
208670	1793	本草纲目
208670	45960	番客篇
208670	67	野草

Abbildung 5.2 Beispielzeilen aus den Tabellen `the_words`, `the_books`, `the_citations`

Bei einer geringen Anzahl von Wort- oder Zeicheneinträgen,[146] bei denen unterschiedliche Bedeutungen angegeben sind, wird im ersten Untereintrag kein Beleg angegeben.[147]

144 Siehe dazu Abschnitt 5.5.3, ab S. 132.
145 Siehe auch S. 123.
146 Betroffen sind ca. 1.400 bzw. 0,3 % der Einträge. (Eigene Berechnung / Zählung mit dem regulärem Ausdruck `(?<!([0-9\.]))1\.[^《]+2\..+《.+)`
147 Ein Beispiel hierfür ist der Eintrag zu *daqing* 大青. Unter „1." ist angegeben, es handele sich um eine Art 1–2 Meter hohen Strauch, dessen äußere Erscheinung dann beschrieben wird – jedoch ohne Angabe einer Belegstelle. Erst unter „2." ist *daqing* als Bezeichnung für eine Art Farbpigmentstein, *bianqing* 扁青, angegeben, wofür dann das Ming-

Zudem kommt es vereinzelt vor, dass die älteste *belegte* Bedeutung nicht zuerst angegeben wird, wodurch solche Lexeme zunächst als „zu neu“ eingestuft würden.[148] Um dem entgegenzuwirken wird für Einträge, bei denen mehrere Bedeutungen angegeben sind und – sofern vorhanden – die chronologische Reihenfolge der verwendeten Dynastieangaben von der tatsächlichen Dynastiefolge abweicht, diejenige Primärquelle als früheste Belegstelle angenommen, die die früheste Dynastieangabe aufweist.[149] Dieses Vorgehen sei anhand des Eintrags zu *muguang* 目光 kurz geschildert.

【目光】
1. 眼睛的光芒。明 高启《猛虎行》：“目光燀燀當路坐，將軍一見弧矢墮。”[…]
2.识见；见解。宋 梅尧臣《梦睹》：“既非由目光，所見定何稟。”[…][150]

Bei der unter „1.“ angegebenen Bedeutung, *„yanjing de guangmang* 眼睛的光芒“ (etwa „[klarer] Blick, [klare] Sicht“ usw.) wird ein Ming-zeitliches Werk, *Meng hu xing* 猛虎行 von Gao Qi 高启 (ca. 1336–1374) als Beleg zitiert, bei der zweiten Bedeutung („Erfahrung, Einsicht, Verstehen“) dann zuerst das Gedicht *Meng du* 夢睹 („Traumbeobachtung“) des Song 宋-zeitlichen (960–1279) Dichters Mei Yaochen 梅堯臣 (1002–1060). Durch die Dynastieangaben ist die nicht-chronologische Anordnung der Glossen erkennbar und die Song-zeitliche Textstelle kann als früheste enthaltene Belegstelle erkannt werden.

Bei einigen Einträgen wird die älteste Bedeutung nicht zuerst aufgeführt bzw. die zuerst genannte Belegstelle ist nicht gleichzeitig die älteste ist. Wenn möglich findet deshalb bei der Extraktion der Lexikalisierungsdaten eine chronologische Reihenfolgenprüfung auf Basis des Dynastiemodells statt. Bei etwa 1.800 Einträgen wird daher nicht die erste im *DHYDCD* genannte Quelle als Zeitpunkt der Lexikalisierung verwendet, sondern stattdessen diejenige mit der frühesten Dynastieangabe.[151]

5.5.2 Verwendung der Metadaten aus dem *DHYDCD*

Zu vielen der Quellenangaben lassen sich, wie oben erklärt, Angaben zu Dynastie und Autor:in direkt aus dem *DHYDCD* ermitteln. Für einige wenige Texte, v. a. Zeitungsartikel, ist der Jahrgang angegeben. Da die Position der *Named Entities* klar definiert ist und ein Abgleich mit der *CBDB* oder die Verwendung von *NER*-Methoden eine unnötige Limitation der erkennbaren Namen zur Folge hätte, wird hier ein regelbasiertes Vorgehen gewählt.[152] Die wesentlichen Herausforderungen, die bei der Extraktion der Daten beachtet werden müssen, sind im Folgenden dokumentiert:

---

zeitliche *Ben cao gangmu* 本草綱目 als Beleg herangezogen wird. Siehe *DHYDCD*, 大青. Tatsächlich ist *daqing* aber in seiner ersten Bedeutung bereits im deutlich älteren *Baopu zi* 抱朴子 zu finden. siehe Ge Hong 葛洪 2020 [Anfang 4. Jh.] *Baopuzi* 抱朴子. Digitalisierte Version der *Sibu congkan*-Ausgabe von *Baopuzi nei wai pian* 《四部叢刊初編》本《抱朴子內外篇》. URL: https://ctext.org/baopuzi (besucht am 20. 09. 2020), *neipian* 內篇, *zhi li* 至理.

148 Davon betroffen sind etwas weniger als 0,5 %, bzw. etwas mehr als 1.800 aller Einträge des *DHYDCD*.

149 Zur Erkennung solcher Einträge wird ein einfacher Sortieralgorithmus genutzt: Beginn und Ende der angegebenen Dynastien werden nachgeladen und die Bedeutungen so nach dem ersten Jahr der jeweils angegebenen Dynastie (*startyear*) sortiert. Weicht die entstandene Sortierung von der ursprünglichen Sortierung im Eintrag ab, wird das Lexem in der Tabelle `the_words` als `unordered` markiert. Nur wenn alle Bedeutungen geeignete Belege haben, kann dieser Eingriff vorgenommen werden.

150 *DHYDCD*, 目光. Unterstreichung und farbliche Hervorhebungen durch den Verfasser.

151 Die betroffenen Einträge werden in der Spalte `the_words.unordered` markiert.

152 Vgl. dazu Kapitel 4.7, ab S. 97. Versuche mit *CKIP Tagger* und *CKIP Transformers* haben zudem Probleme bei DynastieAutor:in-Zeichenfolgen gezeigt.

Grundsätzlich lassen sich die Titel von zitierten Werken mit einem einfachen regulären Ausdruck erkennen: `《[^》]+》`, d. h. beliebig viele beliebige Zeichen innerhalb von *double angle brackets* 《》. Vor und nach den Klammern können jedoch ebenfalls relevante Informationen über Erscheinungsjahr[153] und Textgattung[154] angegeben sein. Diese Informationen helfen zudem, zahlreiche Werke gleichen Titels voneinander zu unterscheiden.

Angaben wie *zhu* 注 vor Werktiteln werden dem Titel des Werkes zugerechnet, wie z. B. bei Zheng Xuans 鄭玄 Kommentar zum *Lunyu* 論語 („东汉末郑玄注《论语》“[155]) Dies ist nicht unproblematisch, da z. B. das Zeichen *zhu* 注 auch in Namen vorkommen kann.[156]

Es ergibt sich folgender regulärer Ausdruck:

```
ur'(?:(?:释文引|注)[^)(*.“’””’。，；？！：、》>…】 () [] ]{0,9})?(?:《[^》]+》)(?:\d{4}|
    注|词|诗|曲|套曲)?(?![^“]*”)'
```

Extrahiert werden also zunächst die erwähnten, den Titel ergänzenden Angaben, sowie bis zu neun Zeichen vor Werksnamen, bis das vorhergehende Satzzeichen erreicht wird. Innerhalb dieser Stelle finden sich, sofern vorhanden, unterschiedliche Angaben zum zitierten Werk, die bestenfalls Dynastie und Autor:in enthalten.

Zitierweise und Dynastiesystem

Eine möglichst vollständige Erkennung der Dynastieangaben ist wichtig, da sie oft die einzige chronologische Angabe an den Quellenzitaten darstellen. Dabei verwendet das *HYDCD* – wie bereits angesprochen – ein eigenwilliges System, das im Sprachverständnis der Herausgeber begründet sein dürfte: Antike Werke, wie etwa das *Yijing* 易經,[157] das *Shijing* 詩經,[158] das *Shujing* 書經 bzw. *Shangshu* 尚書[159] oder *Zhuangzi* 莊子[160] werden fast grundsätzlich ohne Angabe von Dynastie und Autor:in zitiert, wenn man sich nicht auf eine bestimmte Ausgabe bezieht. Eine Ausnahme davon sind Werke, die Song Yu 宋玉 (ca. 319–298 v. u. Z.)[161] zugeschrieben werden.[162]

Bei Texten deren Autor:in unbekannt ist, die von einem Autor:innenkollektiv stammen oder unter chinesischen Gelehrten allgemein bekannt sind, wird die Angabe von Dynastie und Autor üblicherweise ebenfalls weggelassen. Ein typisches Beispiel dafür sind die offiziellen Dynastiegeschichten (*zhengshi* 正史). So wird etwa die *Neue Geschichte der Fünf Dynastien* (*Xin Wudai shi* 新五代史) konsequent nur als Werktitel zitiert, obwohl es ebenso unproblematisch wäre, den Text Ouyang Xiu 歐陽修 (1007–1072)[163] zuzuschreiben und in die Song 宋-Zeit zu datieren.[164]

153 z. B. im Format „《人民日报》1982.3.14“; „《人民日报》1957.10.29“. *DHYDCD*, 交售, 交議.
154 Beispiele dafür sind Angaben wie „李善注引《广雅》“ und „陆德明释文引《广雅》“. Ebenfalls nachgestellt finden sich Hinweise auf Kommentare (*zhu* 注), sowie Gedichte und Lieder (*ci* 詞, *shi* 詩, *qu* 曲, *taoqu* 套曲). *DHYDCD*, 壘 2 壘, 廉劌.
155 *DHYDCD*, 張侯論.
156 Der Name von genau 50 Personen aus der *CBDB* endet mit *zhu* 注.
157 Siehe z. B. *DHYDCD*, 左右.
158 Siehe z. B. *DHYDCD*, 左右.
159 Siehe z. B. *DHYDCD*, 一心. Das *Buch der Urkunden* wird stets als 《*Shu*》《书》 zitiert.
160 Siehe z. B. *DHYDCD*, 朝三暮四.
161 Shih Hsiang-lin 施祥林 und David R. Knechtges 2014: „Song Yu 宋玉“. In: *Ancient and Early Medieval Chinese Literature. A Reference Guide. Part Two*. Hrsg. von David R. Knechtges und Chang Taiping 張泰平. Handbook of Oriental Studies. Leiden: Brill, S. 1007–1022, S. 1007.
162 Hier wird *Zhanguo Chu* 战国楚, *Zhanguo shi* 战国时 („die Zeit der Streitenden Reiche“), oder manchmal nur *Zhanguo* 战国 als Dynastie angegeben. Der Grund für diese Ausnahme erschließt sich nicht, vermutlich ist sie auf eine Ungenauigkeit im ursprünglichen Karteikartensystem der *HYDCD*-Herausgeber zurückzuführen. Siehe *DHYDCD*, 對問, 大王風, 更唱迭和.
163 Wilkinson 2000, S. 504.
164 Siehe z. B. *DHYDCD*, 三十六英雄.

Erst bei Werken ab der Han 漢-Zeit finden sich regelmäßig Angaben im Stil DynastieAutor:in《Werksname》.[165] Wie schon die aus dem Werk Song Yus 宋玉 zitierten Stellen zeigen, sind die Dynastienennungen nicht konsequent einheitlich gehalten. Eine Quelle aus der östlichen Han (*Dong Han* 東漢, 25–220) kann etwa mit der Angabe *Dong Han* 东汉, oder lediglich *Han* 汉 versehen sein.[166]

Eine weitere Kuriosität in der Zitierweise stellt der Umgang mit Texten dar, die nach Ende der Qing 清-Zeit (1644–1911) erschienen sind. Bei republikzeitlichen Werken oder Werken aus der Volksrepublik wird grundsätzlich nur der Autor genannt.[167] Da – wie oben erläutert – auch bei ganz frühen oder als allgemein bekannt geltenden Werken die Zeitangabe fehlt, lässt sich diese Erkenntnis leider nicht ohne Weiteres für die Verortung solcher Werke ins 20. Jh nutzen.

Insgesamt sind die chronologischen Angaben im *DHYDCD* für den Zeitraum vom Beginn der Han-Dynastie im Jahr 206 v. u. Z. bis 1911 am vollständigsten. Die unzähligen Quellen aus der Zeit davor und danach, sowie weitere allgemein bekannte Texte, lassen sich nur mit zusätzlichen Daten einordnen.

Aus den teils unorthodoxen oder inkonsequenten Angaben des *DHYDCD* ergibt sich folgendes Dynastiesystem (Tabelle 5.2),[168] das zur Erkennung der Angaben bzw. zeitlichen Einordnung der zitierten Werke genutzt wird.[169] Im *DHYDCD* uneinheitliche Angaben, z. B. *Han* 漢 und *Han dai* 漢代 bzw. *Nan Qi* 南齊 und *Nanchao Qi* 南朝齊 führen darin zu entsprechenden Mehrfacheinträgen.

Tabelle 5.2 Ergänztes Dynastiesystem des *HYDCD*, chronologisch nach Anfangsjahr

Dynastie	正體	*DHYDCD* 简体	von	bis	# zit. Werke
Streitende Reiche[170]	戰國	战国	-1030	-223	354
Chu [Streitende Reiche] (*Zhanguo Chu*)	戰國楚	战国楚	-1030	-223	19
Yan [Streitende Reiche] (*Zhanguo Yan*)	戰國燕	战国燕	-1030	-223	12
Qin	秦	秦	-221	-206	57
Han	漢	汉	-206	220	1.522
= Han (*Han dai*)	漢代	汉代	-206	220	45
Östliche Han (*Dong Han*)	東漢	东汉	25	220	84
Wei [Drei Reiche] (*Sanguo Wei*)	三國魏	三国魏	220	265	927
Shu [Drei Reiche] (*Sanguo Shu*)	三國蜀	三国蜀	221	263	70
Wu [Drei Reiche] (*Sanguo Wu*)	三國吳	三国吴	222	280	58
Jin	晉	晋	265	420	2.035
Frühere Qin [16 Reiche] (*Qian Qin*)	前秦	前秦	350	394	10
Nördliche Wei (*Bei Wei*)	北魏	北魏	386	534	138
Nördliche Liang [16 Reiche] (*Bei Liang*)	北涼	北凉	401	439	35
Song [Südliche Dynastien] (*Nanchao Song*)	南朝宋	南朝宋	420	479	1.117

165 So z. B. im Eintrag zu *Mao Nü* 毛女 („Haarfrau"; eine besonders behaarte Heilige beschrieben, die am *Huàshan* 華山 beheimatet sein soll). „传说中得道于华山的仙女。汉刘向《列仙传·毛女》：“毛女者[…]” *DHYDCD*, 毛女.

166 Xu Shens 許慎 *Shuo wen jie zi* 說文解字 etwa wird einmal im Eintrag zu *xuxue* 鄦學 in der östlichen Hanzeit verortet, im Eintrag zu *shuo wen* 說文 lediglich allgemeiner in der Hanzeit. Doch damit nicht genug: vereinzelt wird auch noch die Angabe *Han Dai* 汉代 („Han-Dynastie") verwendet, wie im Eintrag zu *xingfa zhi*: „汉代班固的《汉书》“. *DHYDCD*, 鄦學, 說文, 刑法志.

167 Eine Unterscheidung in Republik und Volksrepublik wurde möglicherweise aus politischen Gründen vermieden, da eine Zuordnung von Werken aus der Zeit von 1912–1949 zur danach in Taiwan 台灣 weitergeführten Republik (*Minguo* 民國) als Anerkennung ihrer Legitimität gedeutet werden könnte.

168 Zeitangaben übernommen aus Vogelsang 2012, S. 24. Die letzte Spalte gibt an, bei wie vielen unterschiedenen zitierten Werken die jeweilige Dynastieangabe verwendet wurde.

169 Zwar enthält der Indexband des *HYDCD* selbst eine Dynastietabelle, die dort verwendeten Bezeichnungen stimmen aber nicht zuverlässig mit der tatsächlich verwendeten (inkonsequenten) Zitierweise überein. Siehe *HYDCD*, Bd. 13, S. 3–7.

Tabelle 5.2 (Fortsetzung)

Dynastie	正體	*DHYDCD* 简体	von	bis	# zit. Werke
Qi [Südliche Dynastien] (*Nan Qi*)	南齊	南齐	479	502	2
= Qi [Südliche Dynastien] (*Nanchao Qi*)	南朝齊	南朝齐	479	502	523
Liang [Südliche Dynastien] (*Nanchao Liang*)	南朝梁	南朝梁	502	587	2.954
Qi [Nördliche Dynastien] (*Bei Qi*)	北齊	北齐	550	578	110
Zhou [Nördliche Dynastien] (*Bei Zhou*)	北周	北周	557	581	515
Chen [Südliche Dynastien] (*Nanchao Chen*)	南朝陳	南朝陈	557	589	448
Sui	隋	隋	581	618	462
Tang	唐	唐	618	907	31.921
Frühere Shu [Zehn Reiche] (*Qian Shu*)	前蜀	前蜀	903	925	1.095
Fünf Dynastien *Wudai*	五代	五代	907	960	647
Liao	遼	辽	947	1115	33
Song	宋	宋	960	1279	31.678
Jin	金	金	1115	1234	1.962
Yuan	元	元	1234	1367	6.949
Ming	明	明	1368	1644	14.991
Qing	清	清	1644	1911	22.028
Taiping Tianguo[171]	太平天國	太平天国	1851	1864	195
Republik (*Minguo*)[172]	民國	民国	1912	1992	[702]

In Anbetracht dieser Erkenntnisse werden folgende möglichen Zitierweisen berücksichtigt. Es wird dabei immer jeweils ein zitiertes Werk unterschieden, wenn sich Angabe von Dynastie, Autor:in *oder* Titel unterscheiden.

— 1. **DynastieAutor:in**, z. B. „宋司马光《乞罢免役钱状》“.[173] In diesem Fall soll Song 宋 als Dynastie, Sima Guang 司马光 als Autor und *Qi ba mianyi qian zhuang* 乞罢免役钱状 als Titel der Primärquelle extrahiert werden. Da kein zuverlässiges Muster für das Erkennen chinesischer Namen existiert, ist die Aufgabe, Dynastie und Autor:in zu trennen nicht trivial.[174] Die meisten Fälle lassen sich mit folgenden Annahmen abdecken:

— 1.1 **Personennamen** haben mindestens zwei und maximal sieben Zeichen: Vornamen bestehen in aller Regel aus 1–2 Zeichen, Nachnamen haben mindestens ein Zeichen und können nicht länger als vier Zeichen sein. Im Extremfall des letzten Kaisers der Qing 清-Dynastie Aixinjueluo Puyi 愛新覺羅 · 溥儀 kommen wir zusammen mit dem sonst im *HYDCD* kaum in Namen verwendeten Mittelpunkt („ · “) auf die Maximallänge von sieben Zeichen. Für jeden gefundenen *String*, der ***keinen*** Dynastienamen aus Tabelle 5.2 enthält und doch länger als sieben Zeichen ist, kann davon ausgegangen werden, dass es sich nicht (oder zumindest nicht nur) um einen Namen handelt.

170 Da im *HYDCD* nicht zwischen westlicher Zhou (*Xi Zhou* 西周, 11. Jh. –771 v. u. Z.), Frühlings- und Herbstzeit (*Chunqiu* 春秋, 722–482 v. u. Z.) und der Zeit der streitenden Reiche (*Zhanguo* 戰國, 453–221 v. u. Z.) unterschieden wird, ist hier der gesamte Zeitraum der *Zhou*-Dynastie(n) angegeben.

171 Das „Himmlische Reich des höchsten Friedens“ (*Taiping Tianguo* 太平天國) wird von Historiker:innen weniger als legitime Unterbrechung der Qing 清-Herrschaft, denn als Aufstand (*luan* 亂) gewertet.

172 Die Angabe *Minguo* 民國 (Republik, 1912–) wird im *(D)HYDCD* nicht explizit gemacht, genauso wenig wie *Renmin gongheguo* 人民共和國 (Volksrepublik, 1949–), obwohl durchaus Belege aus Texten angegeben werden, die nach 1912 verfasst wurden. Als Enddatum ist aus praktischen Gründen das Jahr 1992 angegeben, da keine neueren Belege im *DHYDCD* vorhanden sind.

173 *DHYDCD*, 朝三暮四.

174 Siehe dazu auch 4.7, ab S. 97.

— 1.2 Alle **Dynastieangaben** entsprechen der Liste in Tabelle 5.2. Dass die Bezeichnungen teilweise überlappen, kann durch absteigende Sortierung nach Länge umgangen werden.[175]

Einige Dynastiebezeichnungen treten auch als Familienname auf. Bei den *xing* 性 Qing 清, Ming 明, Yuan 元, Jin 金, Song 宋, Tang 唐, Sui 隋, Jin 晉, Han 漢 Qin 秦 ist daher besondere Vorsicht geboten. Es ist dabei etwas wahrscheinlicher, dass jemand den Namen einer vorangegangenen Dynastie als Nachnamen trägt als den Namen einer zukünftigen Dynastie.[176] Zur Minimierung dieser Problematik werden Dynastien mit gleicher Anzahl Zeichen also aufsteigend chronologisch sortiert. Bei Dynastien mit einer Zeichenlänge > 1 ist gesichert, dass es sich nicht um den Familiennamen einer Autor:in handelt.
Wie kann jedoch unterschieden werden, ob es sich bei einer Zeichenfolge wie 元麻革 um den Namen einer Person (Yuan Mage) oder um eine *Yuan* 元-zeitliche Person namens Ma Ge handelt? Absolute Sicherheit kann nur die Recherche des zitierten Werks bzw. der möglichen Namen geben. Mittels einer **Liste bekannter Familiennamen** können zumindest mögliche Familiennamen erkannt werden.[177] Dass Ma 麻 als *xing* nachgewiesen ist, macht es zumindest wahrscheinlich, dass es sich um eine Angabe im Format DynastieAutor:in handelt.

— 1.3 **Aufeinanderfolgende Dynastienamen** wie bei Tang Song Jing 唐宋璟[178] lassen zudem darauf schließen, dass es sich beim zweiten Zeichen um den Familiennamen des Autors handelt. Eine Sicherheit besteht dennoch auch hier nicht, wie der Name Song Qinghai 宋清海 (geb. 1947) beweist.[179]

— 2. **Autor:in** – nur der Name der Autorin oder des Autors.[180]

— 3. **Dynastie** – bloße Angabe des Dynastienamens, z. B. „北魏《元灵耀墓志》：“少傾乾蔭，孤苦自立。”“.[181] Diese Zitierweise kommt selten vor.

— 4. Zeitangabe als **Datum oder Jahreszahl**: „《花城》1981 年第 3 期：“恰好她的父亲是个热心研制中草药的老人[…]“.[182] Solche Angaben finden sich bei Zeitschriften oder Tageszeitungen in der Regel direkt hinter dem Titel der Primärquelle. In selteneren Fällen finden sich Jahreszahlen auch im Titel, z. B. „《<1958 年儿童文学选>序言》“.[183] Ebenfalls möglich sind Jahreszahlen in vollbreiten Unicode-Ziffern (z. B. „１９５７“).[184] Chinesische Jahreszahlen wie „一九二九年“

175 D. h. *Zhanguo Chu* 战国楚 (drei Zeichen) vor *Zhanguo* 战国 (zwei Zeichen).
176 Diese Annahme ist nicht statistisch erforscht. Es finden sich aber in den aus dem *DHYDCD* extrahierten Metadaten z. B. etliche Qing 清-zeitliche Autor:innen mit dem Nachnamen Song 宋 (Song Weipan 宋維藩, Song Qianxu 宋潛虛, Song Yongyue 宋永岳, Song Xiangfeng 宋翔鳳, Song Xuezhu 宋學洙, Song Wan 宋琬, Song Dazun 宋大樽, Song Xian 宋銑 usw.), jedoch nur ein einziger Song-zeitlicher Autor, dessen Name – überdies ein Pseudonym – mit dem Zeichen *qing* 清 beginnt: *Qingyuan Zhenjun* 清源真君, der „wahren Fürst der klaren Quelle“. Allgemeingültig ist diese Annahme jedoch nicht, da sich bereits zur Tang 唐-Zeit Personen mit dem Nachnamen Song 宋 – bereits vor der Tang-Zeit der Name einer Dynaste – finden. Vgl. *DHYDCD*, 一枕, 一線, 井井有法, 屯 2 田, 中 2 率, 升堂拜母, 脩辭, 俯燭.
177 Eine zum Abgleich mit den Inhalten des *DHYDCD* in Kurzzeichen erzeugte Liste aller 1.549 ersten Zeichen von Familiennamen wurde zu diesem Zweck aus der *CBDB* erzeugt. Die Liste wurde mit weiteren Erkenntnisen aus dem *DHYDCD* angereichert und die einsilbigen Dynastienamen (s. o.) entfernt.
178 *DHYDCD*, 養老.
179 *DHYDCD*, 論 2.
180 Siehe dazu auch das Beispiel auf S. 123.
181 *DHYDCD*, 乾 2 蔭.
182 *DHYDCD*, 中草藥.
183 *DHYDCD*, 恰切.
184 „【踏察】勘察，探测。《１９５７散文特写选》序：“你可以认识从北大荒的踏察人员到海南岛的盐业工人[…]“ *DHYDCD*, 踏察.

(1929) werden hingegen als Bestandteil des Titels betrachtet.[185]

— 5. **Sonstige** häufiger angegebene Informationen, z. B. „ 中国近代史资料丛刊“ („Collectanea of Materials on Modern Chinese History“) als Titel einer Buchreihe[186] oder „ 马王堆汉墓帛书甲本“[187] als Hinweis auf einen bestimmten Band bzw. Version des danach angegebenen Werkes. In beiden genannten Fällen kann eine Dynastieangabe aus einer Liste hinzugefügt werden.

— 6. **Keine Metadaten**, wie in der Regel z. B. bei Klassikerzitaten.[188] Hier kann zunächst lediglich der Titel extrahiert werden.

Weitere, seltenere Fälle werden zur Vermeidung von *Over-Engineering* nicht berücksichtigt.[189] Selbst durch ausgefeiltes Parsing lässt sich bei der Extraktion der Metadaten keine Perfektion erreichen, vor allem die Trennung von Dynastie und Autor:in ist problematisch. Angaben wie „ 宋宋祁《授龙图阁谢恩表》[…]“[190] können durch das doppelte Auftreten von Song 宋 als Dynastie- und Familienname verhältnismäßig gut erkannt werden. Bei Angaben wie Tang Wuke 唐无可 ohne zusätzliche Informationen zu erkennen, ob Tang hier tatsächlich der Familienname des Autors ist, bleibt unmöglich.[191] Ein noch komplexeres Regelwerk würde die Nachvollziehbarkeit der erzeugten Daten zudem immer weiter verschlechtern.

5.5.3 Gewinnung von Daten aus der *China Biographical Database*

> *„Dowerjai, no prowerjai! Доверяй, но проверяй!*
> *Vertraue, aber prüfe nach!“*
>
> Russisches Sprichwort

Die *CBDB*[192] eignet sich in zweierlei Hinsicht, um die bereits aus dem *DHYDCD* gewonnenen Metadaten zu verdichten. Aus der Tabelle `text_codes` lässt sich das Erscheinungsjahr für einige der zitierten Primärtexte ermitteln. Wenn dies nicht gelingt, oder der Text nicht verzeichnet ist, können gegebenenfalls die Lebensdaten von Autor:innen ergänzt werden: Deren Lebensspanne ist in der Regel deutlich kürzer, als die – wenn überhaupt – im *DHYDCD* angegebenen Dynastien. In „ungünstigen“ Fällen wie Tang 唐, Song 宋, Ming 明 oder Qing 清 beschreiben sie einen

185 z. B. „[…]殷夫《一九二九年的五月一日》诗:“我们总同盟罢业,[…]“. *DHYDCD*, 罷業. Zitiert wird hier ein Gedicht mit dem Titel „Der 1. Mai 1929“ von Xu Xiaojie 徐孝杰 (1909–1931), der unter dem Pseudonym Yin Fu 殷夫 veröffentlichte. Vgl. OCLC 2019, `1ccn-n82080968` 殷夫 1910-1931. Ob das Gedicht wirklich 1929 verfasst wurde, geht aus dem Titel keineswegs hervor.

186 Siehe z. B. *DHYDCD*, 三點會, „【三點會】天地会的别名。中国近代史资料丛刊《辛亥革命·兴中会革命史要》:“本来在广州的客籍人,多半加入三点会。“ *Zhongguo Jindai Shi Ziliao Congkan* 中国近代史资料丛刊 ist eine 1951–1961 veröffentlichte Buchreihe über moderne chinesische Geschichte. Im hier gezeigten Beispiel wird aus dem Band über die Xinhai-Revolution (*Xinhai Geming* 辛亥革命) zitiert.

187 „Erstes Buch der Manuskripte aus den Han-Gräbern von Mawangdui 馬王堆“. Siehe z. B. *DHYDCD*, 才 2, „[…]通“哉1”。语气词。马王堆汉墓帛书甲本《老子·德经》:“以正之國,以畸用兵,以無事取天下,吾何以知其然也才。” […]“.

188 z. B. „《庄子·齐物论》:“狙公賦芧,曰[…]“ *DHYDCD*, 朝三暮四.

189 Ein Beispiel dafür wären die inkonsequenten Angaben einer Autorenmehrheit, etwa „ 夏丏尊叶圣陶《文心》“ *DHYDCD*, 規約. Zwar wird das *Wenxin* meistens mit genau diesen beiden Autoren, Xia Mianzun und Ye Shengtao, zitiert, vereinzelt wird aber auch die Reihenfolge umgekehrt oder nur Xia wird mit *deng* 等 („et al.“) genannt.

190 *DHYDCD*, 膚屑.

191 Siehe z. B. *DHYDCD*, 朦朧.

192 Siehe dazu auch Kapitel 4.7, ab S. 97.

Zeitraum von etwa 300 Jahren, während die durchschnittliche Lebensspanne einer in der *CBDB* katalogisierten Person 60 Jahre beträgt.[193]

Die Datenübernahme erfolgt in zwei Schritten:

— 1. Das **Erscheinungsjahr der Texte** wird – sofern verfügbar – ermittelt und zugewiesen. Wenn möglich werden auch Lebensdaten von Autor:innen bzw. Herausgeber:innen ergänzt, sofern sie eindeutig zugeordnet werden können. Als Treffer wird dabei die Übereinstimmung von Titel *und* Autor oder Titel *und* Epoche gewertet. Ist beides im *DHYDCD* nicht angegeben, so werden nur Informationen für eineindeutige Werktitel übernommen.[194]

— 2. Wenn in der *CBDB* keine geeigneten Informationen über den Text vorliegen, aber bereits die Autor:in ermittelt werden konnte,[195] werden die **biographischen Daten** zur genaueren chronologischen Einordnung der im *DHYDCD* zitierten Texte verwendet. Da auch Personennamen in beiden Datenbanken mehrfach vorkommen können,[196] werden die Daten nur dann übernommen, wenn eine ein-eindeutige Übereinstimmung besteht, oder wenn die Lebensdaten der Person zu einer bereits aus dem *DHYDCD* extrahierten Dynastieangabe passen.

Die beschriebenen Einschränkungen reduzieren das Potenzial der Datengewinnung aus der *CBDB* zwar, stellen aber sicher, dass deutlich weniger *false positives* übernommen werden. Für mehr als 100.000 der in `the_books` unterschiedenen Quellen können so Metadaten aus der *CBDB* ergänzt bzw. präzisiert werden.[197]

Da die Metadaten hier jedoch „unüberwacht“ gewonnen wurden, kann die Datenübernahme in einzelnen Fällen auch zu falschen bzw. zu späten Lexemdatierungen führen. So wird z. B. das *Xiaojing* 孝經 (*Klassiker der kindlichen Pietät*), ein Text aus dem konfuzianischen Kanon, der ziemlich sicher in die vorchristliche Zeit datiert werden kann,[198] durch einen in der *CBDB* gelisteten, gleichnamigen songzeitlichen Text mehr als 1.000 Jahre zu spät in das Jahr 1098 datiert. Das wirkt sich wiederum auf die zeitliche Einordnung aller 62 *DHYDCD*-Einträge aus, die das *Xiaojing* als ältesten Beleg zitieren.

Ergänzende Recherche von Metadaten

Da einige frühe bzw. kanonisierte Texte im *HYDCD* ohne Angabe von Dynastie oder Autor zitiert werden, kann durch die in den Abschnitten 5.5.2 und 5.5.3 beschriebenen Maßnahmen gerade für einige sehr häufig zitierte Texte keine chronologische Einordnung der zugehörigen Lexeme vorgenommen werden. Für die am häufigsten als *Locus classicus* zitierten Werke lohnt es sich daher – sofern möglich – das Erscheinungsjahr oder zumindest den ungefähren Zeitraum „von Hand“ zu

193 Errechnet aus 33.200 Datensätzen, für die Geburts- und Todesjahr zur Verfügung standen, per `select avg(c_deathyear - c_birthyear) from biog_main where c_birthyear != 0 and c_deathyear != 0 and c_deathyear > c_birthyear and (c_deathyear - c_birthyear) < 200`. (Da z. B. das Todesjahr des ming-zeitlichen Beamten Dong Fan 董璠 (1444–1526) in der *CBDB* mit 4526 angegeben ist (Siehe *CBDB*, 32020, 董璠.) – ist eine einfache Plausibilitätsprüfung angezeigt.)

194 So ist z. B. dem Qing-zeitlichen Kalligraphen Jin Renrui 金人瑞 (1610–1661) ein Werk mit dem Titel *Shiji* 史記 zugeordnet, siehe *CBDB*, 65871, 金人瑞. Im *DHYDCD* ist mit *Shiji* 史记 hingegen in der Regel das Sima Qian 司馬遷 (ca. 145–90 v. u. Z.) zugeschriebene Geschichtswerk gemeint.

195 Siehe dazu Abschnitt 5.5.2, ab S. 128.

196 Siehe dazu Kapitel 4.7, ab S. 97.

197 Siehe dazu auch die Daten und Visualisierungen zur Datenbank in Kapitel 5.7, ab S. 138.

198 Siehe William G. Boltz 1993a: „Hsiao ching 孝經“. In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 141–153, S. 143.

recherchieren und die ermittelten Daten zu ergänzen.[199] Für insgesamt 116.862 Lexeme liegen dadurch ergänzte oder genauere chronologische Daten vor.

5.5.4 Ergänzung um frühere Belegstellen

Eine immer wieder angeführte Kritik am *HYDCD* ist die Unzuverlässigkeit bei der Angabe der frühesten Belegstellen (*Locus classicus*).[200] Diese Problematik kann reduziert werden, indem die Belegstellen um Vorkommen in digital verfügbaren Texten ergänzt werden. Durch Verwendung vorhandener elektronischer Textsammlungen ist dies mit überschaubarem Aufwand möglich. Ein dafür geschriebenes *Python*-Skript verarbeitet die *types* der jeweiligen Texte und ergänzt Belegstellen in der Datenbank. Hierfür werden das Loewe-Korpus,[201] sowie die Volltexte der *zhengshi* 正史 genutzt.[202] Für Lexeme mit einer Länge von 1–3 Zeichen werden zusätzlich Daten des *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌)[203] für 1.000 zufällig ausgewählte Texte[204] herangezogen.

— 1. Für jeden der Korpustexte wird die „beste“ datierte `id` aus der Tabelle `the_books` nachgeschlagen.[205] Texte, die nicht zugeordnet werden können, etwa, weil sie nicht im *DHYDCD* verortet oder nicht datiert sind, werden übersprungen. Für die *Difangzhi* 地方誌 *n*-Gramm-Daten werden Einträge in der Tabelle `the_books` aus den Metadaten des Datensatzes ergänzt.

— 2. Alle 1–4- bzw. 1–3-Gramme der betrachteten Texte werden mit der Liste der Lexeme im *DHYDCD* abgeglichen, für die bereits Belegstellen bekannt sind (Tabelle `the_words`). Die Interpunktion der Korpustexte bleibt dabei unverändert.[206]

— 3. Ist der gerade betrachtete Korpus-Text älter als derjenige, der für ein Lexem im *DHYDCD* als *Locus classicus* angegeben ist, so wird seine in Schritt 1 ermittelte bzw. angelegte ID in die Spalte `earliest_evidence_id` bzw. `earliest_evidence_dfz_id` der Tabelle `the_words` geschrieben.[207] Falls im Verlauf eine noch ältere Belegstelle gefunden wird, wird die `earliest_evidence_id` überschrieben, bis die älteste im Korpus vorhandene Belegstelle dokumentiert ist. Die ursprüngliche Angabe aus dem *DHYDCD* bleibt durch die Verwendung der zusätzlichen Spalten erhalten. Die Verwendung der früheren Belegstellen bleibt so optional und die Nachvollziehbarkeit gewährleistet.

199 Siehe dazu die Auswertungen in Abschnitt 5.7.4, S. 150.
200 Siehe dazu Kapitel 5.3, ab S. 113.
201 Siehe T. Schalmey 2009, S. 104–106, einige Texte dieses Korpus sind nicht genau datierbar, siehe auch Kapitel 4.2, S. 66.
202 Siehe dazu auch Kapitel 2.3, ab S. 20.
203 *DFZ*.
204 Unter Ausschluss der Texte, die in Kapitel 6.1.1 (ab S. 158) bzw. 6.2.5 (ab S. 197) als Testdaten verwendet werden.
205 Falls mehrere Einträge desselben Titels bestehen, wird der mit der frühesten Datierung bevorzugt. Falls auch hier mehrere Einträge bestehen, wird derjenige Text verwendet, der am häufigsten als *Locus classicus* zitiert wurde. In SQL ausgedrückt: `where startyear is not null order by startyear asc, useinfirstcount desc limit 1`. Dass gerade bei häufig zitierten Werken zahlreiche Duplikate in `the_books` vorhanden sind, ist der ungenauen Zitierweise des *DHYDCD* geschuldet. Das Liu Xiang 劉向 zugeschriebene *Liexian zhuan* 列仙傳 („Biographien von Unsterblichen“) wird z. B. auf insgesamt acht unterschiedliche Weisen, teils indirekt (d. h. innerhalb eines Zitats aus einer anderen Quelle), zitiert. Siehe dazu auch Abschnitt 5.5.1, S. 123.
206 D. h. enthält ein Text z. B. die Zeichenfolge „人。人“ oder „人，人“, wird dies nicht als Belegstelle für das Lexem *renren* 人人 gezählt.
207 Um die Wirkung dieser Maßnahmen bewerten zu können (siehe S. 183), werden diese Belege in gesonderte Datenspalten gespeichert (siehe dazu auch Abschnitt 5.5.1, ab S. 123).

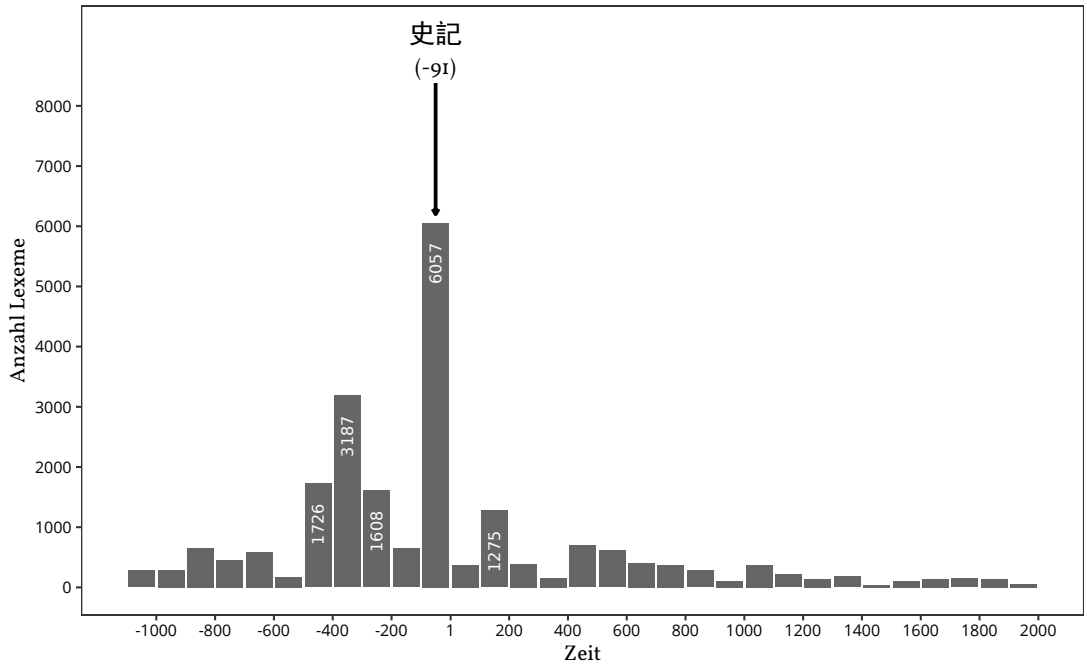


Abbildung 5.3 Neologismusprofil des *Shiji* 史記, ohne Korpusbelegstellen

Für 80.547 (etwa ein Viertel) der im *DHYDCD* lexikalisierten Zeichenkombinationen können so aus den Loewe und *zhengshi*-Korpora frühere Belegstellen ergänzt werden. Aus den *Difangzhi*-Daten werden für 10.325 Zeichenkombinationen frühere Verwendungen aus 731 Texten hinzugefügt.[208] Am Beispiel des *Shiji* 史記-Neologismusprofils (Abb. 5.3, 5.4)[209] wird der Effekt dieser Maßnahme deutlich sichtbar.

Da der Text selbst in den Trainingsdaten enthalten ist, weist die zweite Abb. fast ausschließlich Zeichenkombinationen auf, die *vor* oder auf das 1. Jh. v. u. Z. datiert sind. Trotz der offensichtlich sehr intensiven Rezeption des Texts durch die Herausgeber:innen des *DHYDCD*[210] lassen sich allein im *Shiji* für über 3.000 2–4 Zeichen-Kombinationen frühere Belegstellen finden. Dies ist allerdings nicht ausschließlich auf die Nachlässigkeit der Herausgeber:innen zurückzuführen. Zeichenkombinationen, die in einem der Korpustexte in einer abweichenden Bedeutung auftreten, sollten als *false positives* angesehen werden.[211] Mit 7.468 Lexemen, bei denen bereits im *DHYDCD* das *Shiji* als *Locus classicus* angegeben ist,[212] machen die ergänzten Stellen also 28,7 % der insgesamt mit dem *Shiji* belegbaren Lexeme aus.[213]

Wie viel später die frühesten Belege im *HYDCD* sein können und wie (un-)zutreffend die älteren Belegstellen sind, sei anhand zweier Beispiele veranschaulicht. Die Zeichenfolge *sudi* 宿地,

208 Zu den verwendeten Korpora siehe Kapitel 4.2, ab S. 64.
209 Die hier verwendete Darstellung wird in Kapitel 6.2 (ab S. 179) ausführlich erläutert.
210 Siehe auch Abschnitt 5.7.4, ab S. 150.
211 Im Gegensatz zu den hier ohne jegliche semantische Analyse verglichenen Zeichenfolgen, beziehen sich die Belege im *DHYDCD* auf konkrete Bedeutungen. Entsprechende Beispiele finden sich in Kapitel 6.2.3, ab S. 190.
212 Siehe dazu Tabelle 5.3, S. 150.
213 Der tatsächliche Anteil an *false positives* ist dabei schwer feststellbar. Da in den nachgelagerten Analysen (v. a. Kapitel 6.2, ab S. 179) Zeichenfolgen ebenfalls zunächst ohne jede semantische Analyse verglichen werden, lohnt sich der Aufwand einer genaueren, manuellen Analyse an dieser Stelle nicht.

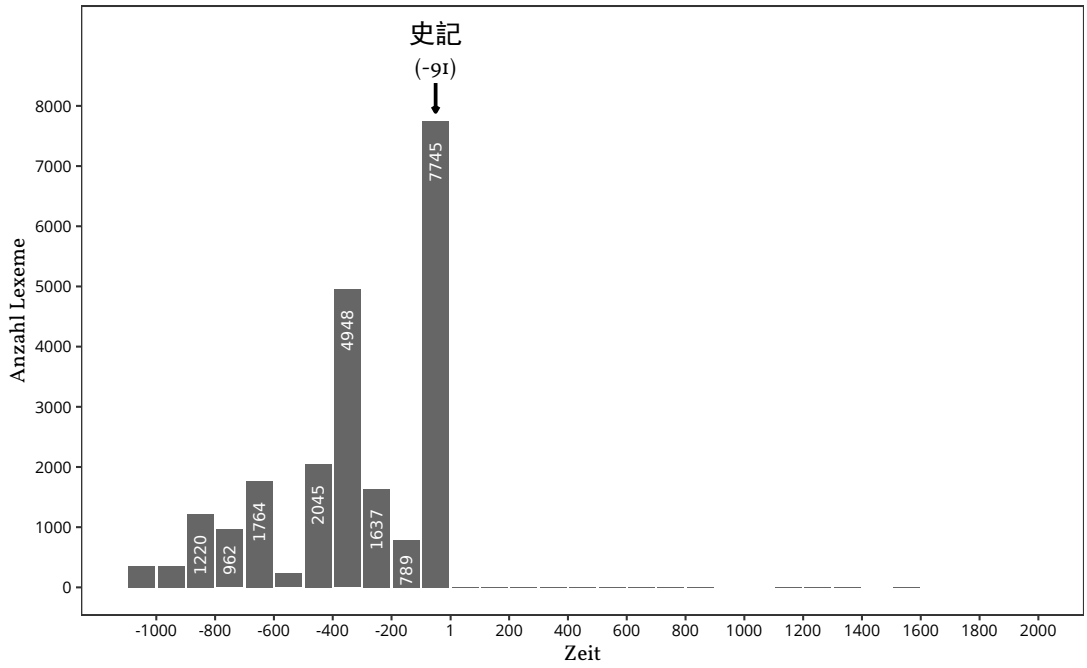


Abbildung 5.4 Neologismusprofil des *Shiji* 史記, mit Korpusbelegstellen

im *HYDCD* erklärt als *zhusu de difang* 住宿的地方, „ein Ort zum Übernachten" und erst mit einer Ausgabe der *Xinmin Evening News* (*Xinmin Wanbao* 新民晚報) vom 29. März 1987 belegt,[214] findet sich im *Shiji* bereits in ähnlicher Bedeutung.[215]

Als *Wan dan jun* 萬石君 („Zehntausend-*dan*-Fürst") bezeichnet Sima Qian 司馬遷 den hanzeitlichen Beamten Shi Fen 石奮 (gest. 124 v. u. Z.), dem im *Shiji* eine Biographie gewidmet ist.[216] Im gleichlautenden *HYDCD*-Eintrag wird zwar auf ihn verwiesen, die früheste Belegstelle für eine angelehnte Bedeutung stammt aber aus dem 1772 veröffentlichten *Gai yu cong kao* 陔餘叢考 von Zhao Yi 趙翼 (1727–1814), obwohl *wan dan* 萬石 bereits für die Han-Zeit belegt wird.[217]

Auch für die Daten aus der *CBDB* (siehe Kapitel 4.7, S. 97) können zeitlich frühere Belege für 12.541 Personennamen und 707 Ortsnamen gefunden werden. Hierzu werden einzigartige Personennamen mit einer Länge von drei Zeichen[218] aus der `biog_main`-Tabelle geladen. Ein Text wird als frühere Belegstelle gewertet, wenn sein `lastyear` früher ist als das angegebene Geburts- bzw. Indexjahr der gefundenen Person. Eindeutige Ortsnamen werden aus der `addresses`-Tabelle geladen und ein Text als frühere Belegstelle gewertet, wenn sein `lastyear` früher ist als die früheste Nennung des Ortsnamens.

214 Siehe *DHYDCD*, 宿地.

215 Vgl. z. B. Sima Qian 司馬遷 1959 [91 v. u. Z.] *Shiji* 史記 *(Records of the Grand Historian)*. Beijing 北京: Zhonghua shuju 中華書局, S. 476: „[…] 古者天子五載一巡狩, 用事泰山, 諸侯有朝宿地 […]". „In alten Zeiten, wenn der Kaiser alle fünf Jahre eine Inspektionsreise machte und am Taishan [Opfer]dienste ausführte, hatten alle Fürsten, die ihm [dort] die Ehre erwiesen einen Ort zum Übernachten […]".

216 Siehe ebd., S. 2763–2768.

217 Siehe *HYDCD*, Bd. 9, S. 462, 萬石, 萬石君. Vgl. auch Soffel 2004, S. 173.

218 Namen mit zwei Zeichen Länge weisen ein sehr hohes Ambiguitätspotenzial auf. Ausführlicher dazu siehe Kapitel 4.7, ab S. 97 und 6.2.2, S. 189.

Theoretisch ließe sich der beschriebene Vorgang mit beliebig vielen Texten wiederholen, um für jede Zeichenkombination die tatsächlich älteste überlieferte Belegstelle zu finden, um Belege für im *HYDCD* unbelegte Lexeme zu finden, oder um Vorkommen von Zeichenkombinationen diachron zu dokumentieren, die nicht als Lexem im *HYDCD* gelistet sind. Letzteres würde den Umfang der erzeugten Daten allerdings dramatisch vergrößern – wahrscheinlich ohne entscheidenden Mehrwert für den Anwendungszweck.[219]

Die nun zur Verfügung stehende diachrone Lexemdatenbank dient als Grundlage für die in Kapitel 6.2 und 6.3 (ab S. 179) behandelten Textdatierungsmethoden. Durch statistische Auswertung dieser Daten können zudem weitere Rückschlüsse auf die Machart des *HYDCD* gezogen werden.[220]

5.6 Das *DHYDCD* als diachrones Behelfskorpus

Die in Kapitel 3.3 vorgestellten Datierungsmethoden werden anhand diachroner Korpora evaluiert.[221] In Ermangelung eines umfangreichen, diachronen Korpus, welches den gesamten Zeitraum der schriftsprachlichen Texttradition abdeckt,[222] wird aus den Belegen im *DHYDCD* ein Behelfskorpus erzeugt.[223] Auch die Entstehung der offiziellen Dynastiegeschichten (*zhengshi* 正史) erstreckt sich zwar über einen großen Zeitraum, mit ingesamt nur 25 Texten bzw. in der Regel einem Text pro Dynastie eignen sie sich jedoch nicht für die Erzeugung von statistischen *chronon*-Sprachmodellen. Wünschenswert ist zudem eine ausgewogene Mischung relevanter Textgattungen.

Die Verwendung eines aus solchen Einzelsätzen erzeugten Korpus wurde bereits am Beispiel des *Oxford English Dictionary* beschrieben.[224] Obwohl es nicht als „vollständig ausgewogen und repräsentativ“[225] gelten kann, stellt es eine umfangreiche Sammlung natürlicher Sprache dar, wobei „die erfasste Zeitspanne von keiner anderen digitalisierten Quelle übertroffen wird.“[226] Dies gilt umso mehr für die insgesamt 919.280 Belegstellen aus dem *DHYDCD*, von denen 612.639 aus etwa 41.436 unterscheidbaren Texten zeitlich eingeordnet werden können.[227]

Um die Methodik der geringen Genauigkeit der zeitlichen Zuordnung der *attestations* anzupassen, wird eine grobe Einteilung in Zeiträume von 100 Jahren (*chronons*) mit einer Überlappung von jeweils 50 Jahren zum nächsten Subkorpus verwendet. Inhaltliche Überschneidungen werden dabei zugelassen. Ein Zeitraum von 100 Jahren erscheint sinnvoll, da im *HYDCD* – im Gegensatz zum *OED* – nicht das Jahr des Erscheinens der zitierten Texte angeben ist, sondern lediglich die Dynastie. Für einen Teil der Texte konnten durch Hinzuziehen externer

219 Zu Unterschieden bei der Nutzung von *n*-Gramm- und wort- bzw. lexembasierten Sprachmodellen siehe v. a. auch Kapitel 6.1, ab S. 156.
220 Siehe Kapitel 5.7, ab S. 138.
221 Siehe Kapitel 3.3, S. 55.
222 Siehe Kapitel 4.2, ab S. 62.
223 Zur Verwendung dieses Korpus siehe Kapitel 6.1.3, ab S. 171.
224 Siehe Kathryn Allan 2012: „Using OED data as evidence“. In: *Current Methods in Historical Semantics*. Hrsg. von Kathryn Allan und Justyna A. Robinson. Topics in English Linguistics. Berlin & Boston: Walter de Gruyter, S. 17–39, S. 19; siehe auch Hoffmann 2004, Hoffmanns Belegstellen aus dem *OED* ergeben ein diachrones Korpus des Englischen von etwa 2,4 Mio. Sätzen bzw. 33–35 Mio. Wörtern aus dem Zeitraum vom 11. bis zum 20. Jh, wobei erst ab dem 15. Jh. eine nennenswerte Menge an Textmaterial vorliegt.
225 Hoffmann 2004, S. 26.
226 Ebd., S. 26, übersetzt durch den Verfasser.
227 Zu Einschränkungen bei der Differenzierung von im *DHYDCD* zitierten Quelltexten siehe Kapitel 5.5.2, S. 127. Siehe auch Kapitel 5.5.3, S. 132.

Quellen wie der *CBDB* die Lebensdaten der Autor:innen oder sogar das Jahr der Veröffentlichung ergänzt werden.[228] Insgesamt wird dadurch eine durchschnittliche Genauigkeit von 76 Jahren erreicht.[229] Durch die Verwendung einer *chronon*-Länge von 100 Jahren ist gleichzeitig sichergestellt, dass für jeden Zeitraum eine für die Erkennung sprachgeschichtlicher Trends ausreichende Menge an Textmaterial extrahiert wird.[230]

Für jeden Zeitraum werden zunächst die relevanten Primärquellen aus der Tabelle `the_books` geladen. Anschließend werden aus der Tabelle `the_citations` die *DHYDCD*-Einträge mit Zitaten aus diesen Werken ermittelt.[231] Daraus werden nun die entsprechenden Belegstellen extrahiert und als chaotisches Pseudotext-Potpourri aneinander gereiht. Zur Veranschaulichung sei hier ein Auszug aus dem Subkorpus für den Zeitraum 1000–1100 gegeben, das sich aus insgesamt 11.191 Belegstellen aus 10.169 Texten zusammensetzt.

> […] 鄜延境内有石油……余疑其煙可用，試掃其煤以爲墨，黑光如漆，松墨不及也。[232]細看落墨皆松瘦，想見掀髯正鶴孤。[233] < 蔚州 > 土貢：熊鞹、豹尾、松實。[234] 罪出其身，不使廢松檟之奉。[235] 爛文章之糾纏，驚節解而流膏……收薄用於桑榆，製中山之松醪。[236]撥置千憂並百慮，且醉一斛松醪春。[237] […]

Aus den so entstandenen 53 Subkorpora lassen sich nun grobe temporale Sprachmodelle berechnen. Dabei kann ein Zeitraum von 700 v. u. Z. bis zum 20. Jh. abgedeckt werden.[238]

5.7 *HYDCD-Data Science*: Erkenntnisse aus der Datenbank

> *„But we are in greater darkness*
> *if we go still further back […]“*[239]
>
> Mario Alinei

Die in Kapitel 5.5 erzeugte Lexemdatenbank erlaubt einige Einblicke in die Machart des *HYDCD* bzw. des *DHYDCD*, sowie in die Entwicklung des chinesischen Wortschatzes.[240] Ein tiefergehendes Verständnis der erzeugten Daten ist zudem für die Entwicklung von Datierungs- bzw. Fälschungserkennungssoftware auf dieser Basis nützlich.

228 Siehe Abschnitt 5.5.3, S. 132
229 Siehe auch Abschnitt 5.7.1, ab S. 139.
230 Vgl. auch Hoffmann 2004, S. 17, siehe auch S. 24.
231 Siehe Kapitel 5.5.1, S. 123. Um den Anteil bei der Extraktion entstandener *false positives* im Korpus zu minimieren, werden nur Primärquellen mit zwei oder mehr *attestations* in Betracht gezogen.
232 *DHYDCD*, 松煙墨. Belegstelle aus dem *Meng xi bi tan* 夢溪筆談 von Shen Kuo 沈括 (1031–1095).
233 *DHYDCD*, 松瘦. Aus *Ciyun* Liu *Jingwen jian ji* 次韻劉景文見寄 von Su Shi 蘇軾 (1031–1101).
234 *DHYDCD*, 松實. Aus dem 1060 fertiggestellten *Xin Tangshu* 新唐書.
235 *DHYDCD*, 松檟. Aus *Xie zhe shou Xiuzhou tuanlian fushi biao* 謝謫授秀州團練副使表 von Shen Kuo.
236 *DHYDCD*, 松醪. Aus *Zhong shan song lao fu* 中山松醪賦 von Su Shi.
237 *DHYDCD*, 松醪春. Aus *Wang Baishuishan ci Hejiang lou yun* 望白水山次合江樓韻 von Li Gang 李綱 (1083–1140). Durch die Art der Datengewinnung aus der *CBDB* wird das Zitat ausgewertet, obwohl Li Gang das Gedicht vermutlich erst nach 1100, also nach *chronon*-Ende verfasst hat, da für diese Quelle nur die biographischen Daten des Autors vorliegen. Erläuterungen dazu siehe Kapitel 5.5.3, S. 132.
238 Siehe dazu Kapitel 6.1.3, ab S. 171.
240 Für eine sprach- und kulturhistorische Betrachtung der hier vorgestellten Daten siehe auch Tilman Schalmey 2020: „Das *Hanyu Da Cidian* 漢語大詞典 als Sprachgedächtnis“. In: *Erinnern und Erinnerung, Gedächtnis und Gedenken*. Hrsg. von Maria Khayutina und Sebastian Eicher. Jahrbuch der Deutschen Vereinigung für Chinastudien. Wiesbaden: Harrassowitz, S. 73–90, *passim*.

5.7.1 Genauigkeit der gewonnenen Daten

Bedingt durch die Zitierweise im *DHYDCD*[241] und die teilweise sehr ungenaue Datierbarkeit älterer Primärquellen[242] können einige Lexeme keinem genauen Jahr zugeordnet werden, sondern unterschiedlich langen Zeiträumen. Abb. 5.5 stellt die Genauigkeit der Datierung aller so eingeordneten Lexeme dar. In Abb. 5.5a wird der Datenstand ohne zusätzliche Belegstellen gezeigt, in 5.5b sind diese berücksichtigt. Die durchschnittliche Genauigkeit der Datierung $\bar{u}$ beträgt 86, bei Berücksichtigung der ergänzten Belege 76 Jahre. Die Darstellung zeigt auch, dass bereits ab der Han-Zeit der überwiegende Anteil der Primärquellen sehr genau bzw. mit einer Genauigkeit von weniger als 100 Jahren datiert werden kann. In Abb. 5.6a zeigt, wie viele Bedeutungen in den datierten Einträgen unterschieden werden. In knapp 80 % der Einträge wird nur eine Bedeutung angegeben und belegt. Mit zunehmender Anzahl an Bedeutungen nimmt dann die Anzahl entsprechender Einträge logarithmisch ab. In einem durchschnittlichen Eintrag werden 1,45 Bedeutungen oder Konnotationen unterschieden und belegt. Ein offensichtlicher Zusammenhang besteht zwischen Äquivokation[243] und der Zeichenlänge der Einträge. Längere, mehrsilibige Wörter bzw. Phrasen weisen tendenziell eine geringere Anzahl an Bedeutungen auf (Abb. 5.6b).

Für einzelne Zeichen können – im Extremfall des Eintrags zu *fa* 發 – bis zu 81 Bedeutungen unterschieden werden.[244] In solchen Fällen werden aber zahlreiche streitbare Konnotationen betrachtet, deren semantischer Unterschied aus den angegebenen Textbelegen oft nicht klar wird.[245] Davon abgesehen lassen die Anteile an mehrdeutigen Lexemen erahnen, dass die Unterscheidung von Wortbedeutungen (*word sense disambiguation*) einen Mehrwert für Datierungsaufgaben bringen würde,[246] gleichzeitig aber insbesondere für klassische Texte eine immense Herausforderung darstellt.[247]

Im Kontext der Mehrdeutigkeit chinesischer Zeichen sei nochmals auf *duoyinzi* 多音字, Zeichen mit unterschiedlichen Aussprachen, eingegangen. Von den 16.361 graphisch unterschiedlichen Schriftzeichen, denen im *DHYDCD* Einträge gewidmet sind, ist für etwas mehr als 80 % nur eine einzige Lesung angegeben, für die restlichen sind zwei oder drei, in einzelnen Fällen sogar bis zu sieben Lesungen bekannt (Abb. 5.7).[248] Fast immer werden unterschiedliche Lesungen mit abweichenden Bedeutungen, häufig auch mit anderen grammatikalischen Kategorien in Verbindung gebracht.

---

240 Alinei 2004, S. 215.

241 Siehe Abschnitt 5.5.2, ab S. 127.

242 Siehe Loewe 1993, S. xi.

243 Da hier nur die graphische Gestalt der Lexeme und die Anzahl der angegebenen Erklärungen untersucht werden kann, ist eine Unterscheidung zwischen Homographie (bei unterschiedlicher Aussprache), Homophonie (unterschiedliche Bedeutung bei gleicher Aussprache) und Polysemie (unterschiedliche, verwandte Bedeutungen) hier nicht abbildbar.

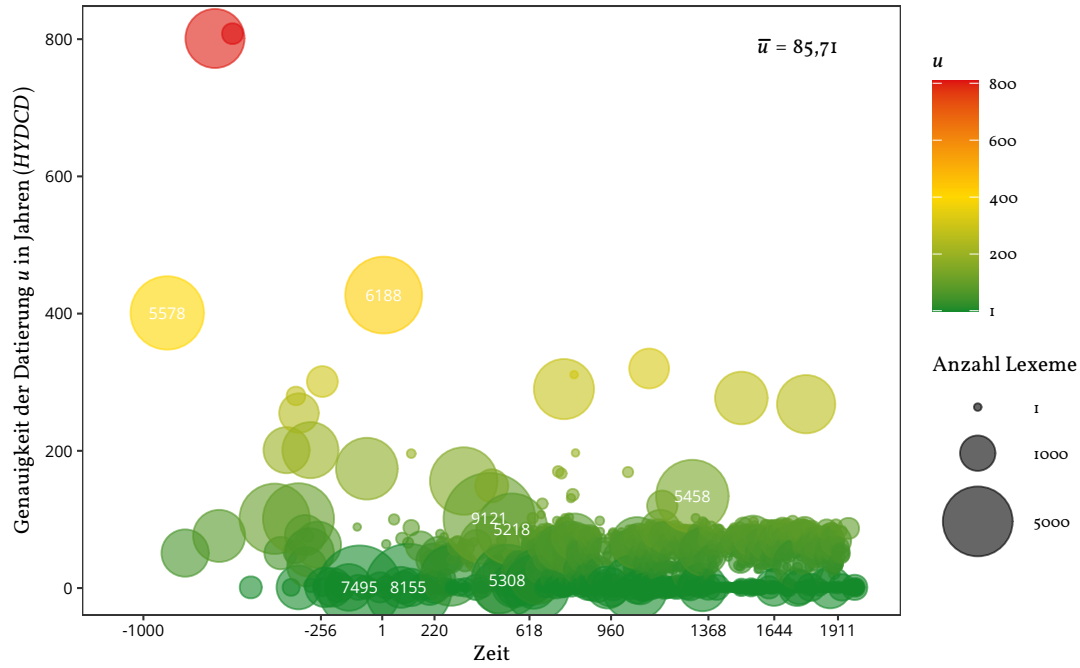
244 Siehe *DHYDCD*, *fa* 發.

245 Im Eintrag zu *zhi* 之 etwa wird zwischen mehreren Fällen unterschieden, in denen *zhi* stets als subordinierende Partikel zwischen zwei Satzgliedern funktioniert, der grammatikalische Unterschied wirkt eher konstruiert. Ähnliches gilt für die Verwendung als Pronomen. Siehe *DHYDCD*, *zhi* 之.

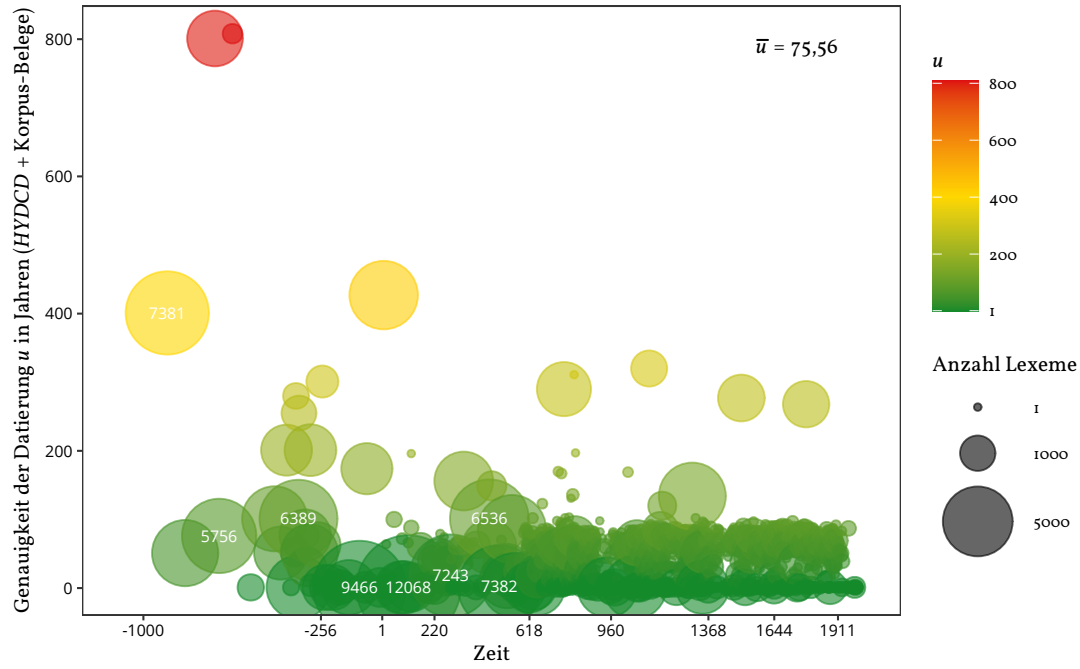
246 Vgl. Kanhabua und Nørvåg 2008, S. 361.

247 Dass eine automatisierte, kontextuelle Erkennung unterschiedlicher Wortbedeutungen und deren Veränderung grundsätzlich möglich ist, wird in Tahmasebi, Borin und Jatowt 2019, erörtert. Siehe S. 1–2. Eine darin vorgestellte Studie über semantische Veränderungen chinesischer Wörter ist Tang Xuri, Qu Weiguang und Chen Xiaohe 2015: „Semantic Change Compuation: A Successive Approach". In: *World Wide Web* 19.3, S. 375–415. doi: 10.1007/s11280-014-0316-y, Die Implementierung vergleichbarer Techniken würde den Rahmen dieser Dissertation sprengen. Für *n*-Gramm-Daten sind sie zudem nicht anwendbar.

248 Für das Zeichen 繆 werden folgende Lesungen gegeben: *móu*, *jiū*, *miù*, *mù*, *miào*, *liáo* und *lù*. Siehe *DHYDCD*, 繆/繆 1 – 繆 7.

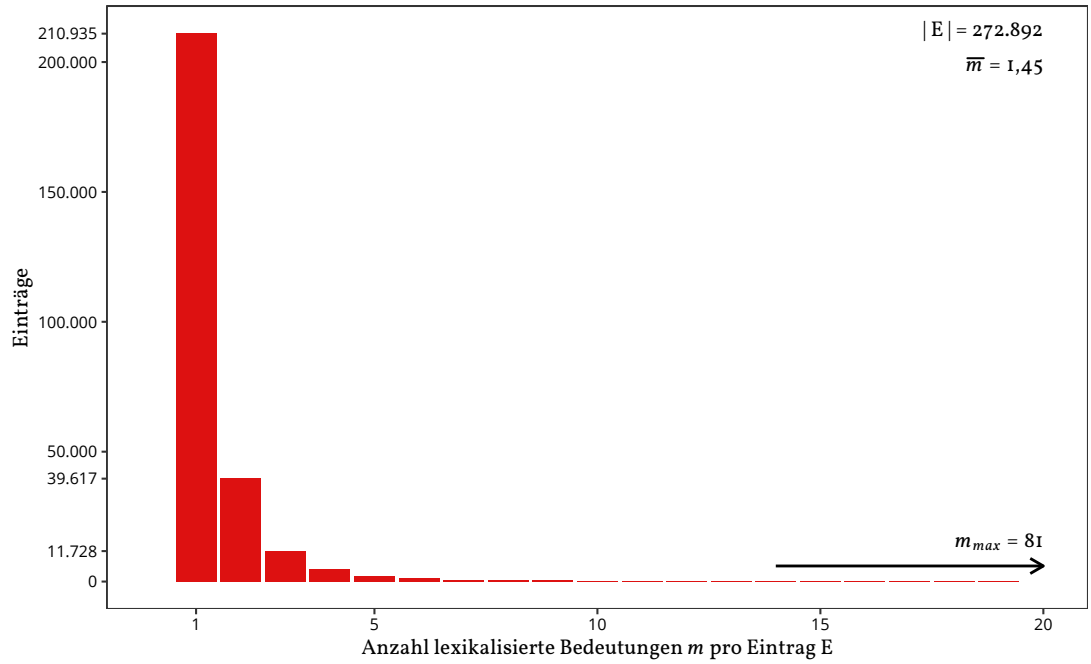


(a) *HYDCD*

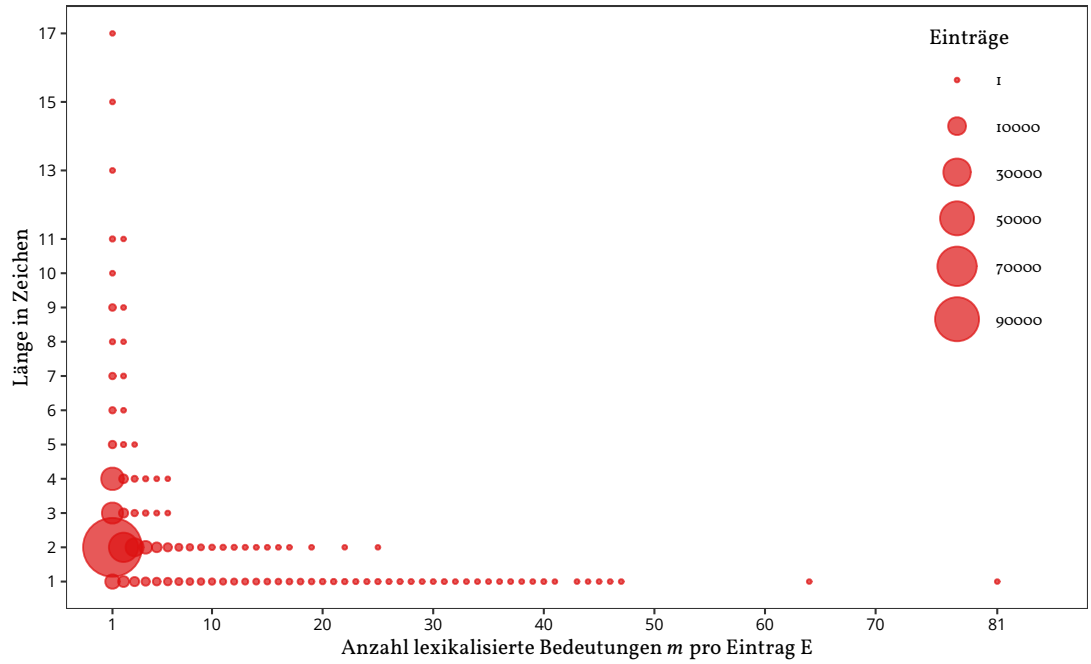


(b) + Korpora-Belege aus Loewe, *zhengshi* 正史

Abbildung 5.5 Genauigkeit der Lexemdatierung



(a) Anzahl angegebener Bedeutungen *vs.* Anzahl Einträge



(b) Anzahl angegebener Bedeutungen *vs.* Länge in Zeichen

Abbildung 5.6 „Unterschiedliche“ Bedeutungen in *HYDCD*-Einträgen

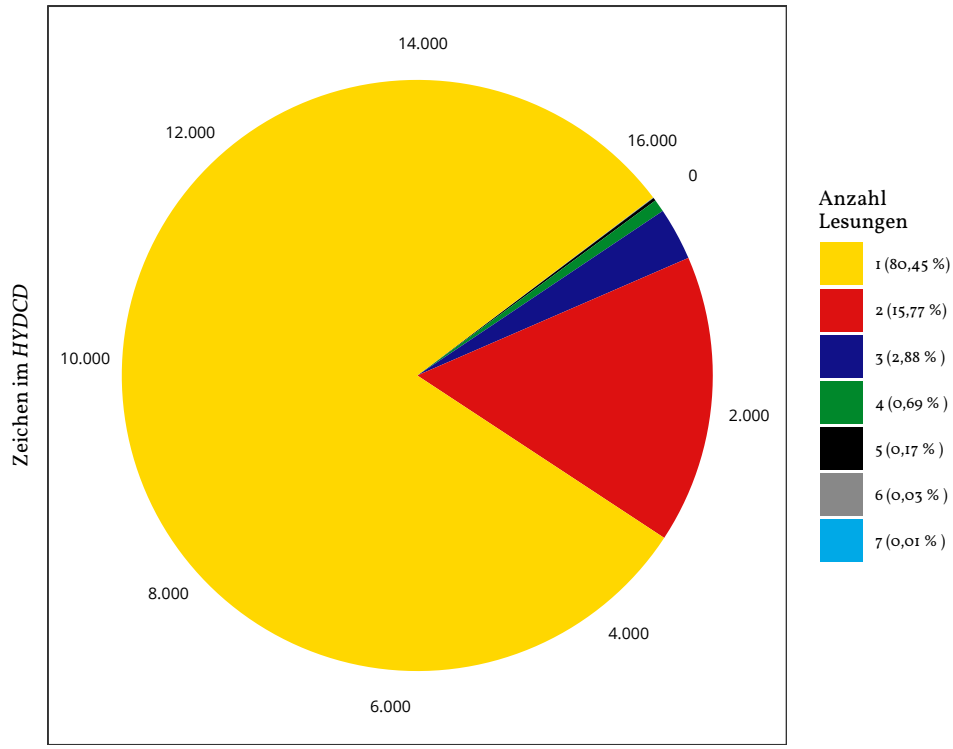


Abbildung 5.7 Lexikalisierte Zeichen im *DHYDCD* nach Anzahl ihrer Lesungen

5.7.2 Lexikalisierung pro Jahrhundert

Betrachtet man anhand der jeweils frühesten Belegstellen der *DHYDCD*-Lexeme aus der Tabelle `the_words` die Anzahl neuer Lexikalisierungen pro Jahrhundert auf einer Zeitachse, wird die chronologische Abdeckung und Gewichtung der Daten sichtbar (Abb. 5.8).[249] Sie lässt unterschiedliche Rückschlüsse und Interpretationen zu.

— 1. Die vor dem **4. Jh. v. u. Z.** verhältnismäßig dünne Datenlage zeigt sich klar. Sie ist darauf zurückzuführen, dass aus den ersten Jahrhunderten des Betrachtungszeitraums überhaupt nur wenige (längere) Texte überliefert sind. Zudem ist die Datierung prä-hanzeitlicher Texte meist ungenau, so dass teils auf grobe Schätzungen zurückgegriffen werden muss.[250] Zudem wurde ein Großteil der bis heute überlieferten frühen Texte während der Han-Zeit durch Liu Xiang 劉向 (77–6 v. u. Z.) und seine Mitarbeiter redigiert und standardisiert, so dass sie uns gewissermaßen gefiltert vorliegen.[251]

— 2. Im *DHYDCD* stehen für Quellen aus dem **20. Jh.** meist weniger Metadaten zur Verfügung als sonst.[252] Dadurch fällt die Lexikalisierung auch hier übertrieben gering aus, obwohl gerade

249 Lexeme, die wg. einer zu ungenauen Datierung der Belegstelle nicht eindeutig einem Jahrhundert zugeordnet werden können, werden gemäß ihrer Datierung anteilig zugeordnet. Vorgehensweise und Berechnung dafür werden in Kapitel 6.2.1, ab S. 184 beschrieben.
250 Siehe auch die Angaben in der Tabelle der häufig zitierten Texte in Abschnitt 5.7.4, S. 150.
251 Siehe z. B. Kern 2004, S. 46.
252 Siehe Abschnitt 5.5.2, S. 129.

im 20. Jh. durch den viel intensiveren Kontakt mit dem Westen,[253] den technischen Fortschritt und den Erfolg der geschriebenen Umgangssprache (*baihuawen* 白話文)[254] zahllose Neologismen entstanden sind.

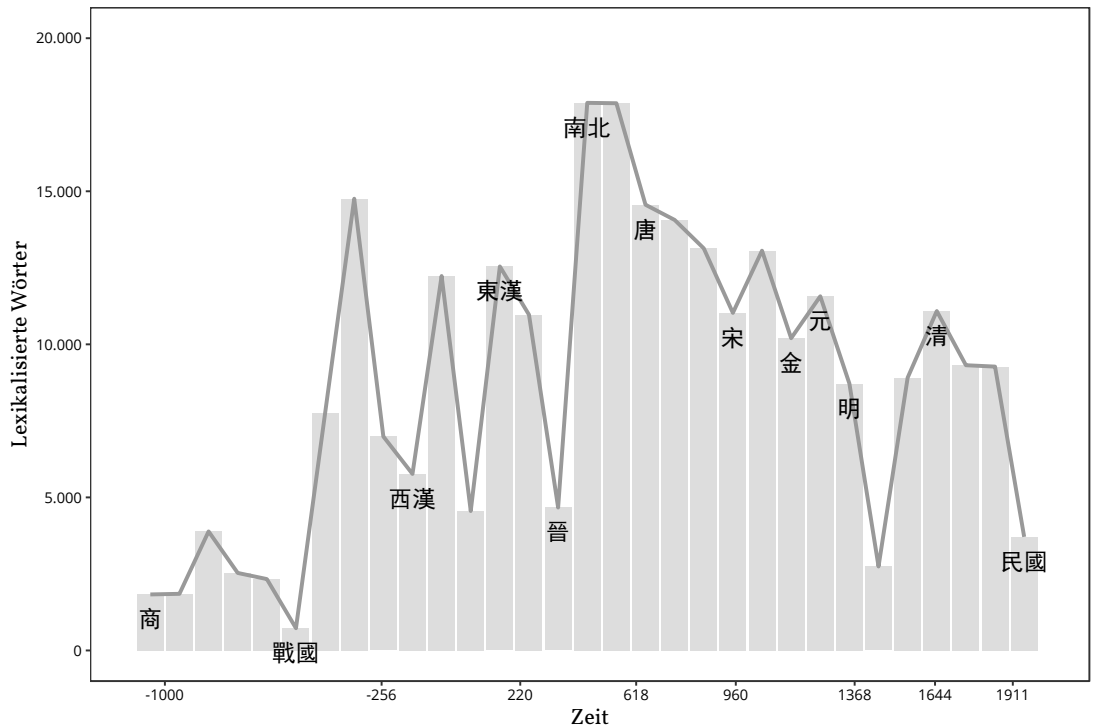


Abbildung 5.8 Lexikalisierung im *HYDCD* nach Jahrhundert. Die Balken stehen für die Anzahl der Lexeme, deren älteste Belegstelle (etwa) aus dem jeweiligen Jahrhundert stammt.

— 3. Auf den ersten Blick scheint die restliche Lexikalisierung (und damit die Entstehung) neuer Wörter einer zyklischen bzw. dynastischen **Schwankung** zu unterliegen, angedeutet durch die Zickzackkurve in Abb. 5.8. Dies ist tatsächlich plausibel, da auch für andere Sprachen größere Schwankungen im Wortschatz im Zusammenhang mit wichtigen historischen Ereignissen in Verbindung gebracht werden konnten.[255] Dazu passt besonders auch der Abwärtsknick in der Lexikalisierung während der gegenüber Einflüssen von außen als besonders verschlossen geltenden Ming 明-Dynastie.[256] Eder mahnt jedoch zurecht, dass „jeder Versuch, direkte Zusammenhänge zwischen historischen Ereignissen und stilistischen Veränderungen

253 Siehe z. B. Lackner, Amelung und Kurtz 2001, S. 2.

254 Vgl. z. B. Elisabeth Kaske 2007: *The Politics of Language in Chinese Education, 1895–1919*. Leiden: Brill, v. a. S. 30–31.

255 Siehe Mikhail V. Arapov 1983: „Word Replacement Rates for Standard Russian (A.D. 1100–1850)". In: *Historical Linguistics*. Hrsg. von Barron Brainerd. Quantitative Linguistics 18. Bochum: Dr. N. Brockmeyer, S. 50–61, S. 60; siehe z. B. auch Eder 2018, S. 364: „Our study corroborated the hypothesis that epochs of substantial stylistic drift are followed by periods of stagnation, rather than forming purely linear trends." ; vgl. auch Bochkarev, Solovyev und Wichmann 2014, S. 4.

256 Siehe z. B. Richard von Glahn 1996: *Fountain of Fortune: Money and Monetary Policy in China, 1000–1700*. Berkeley: University of California Press, S. 90. Hinter der konservativen Isolationspolitik der Ming standen allerdings vor allem wirtschaftspolitische Beweggründe.

zu finden, menschlichen Vorurteilen unterliegt."[257] Dass lexikalischer Wandel durch Krisen beschleunigt werden kann, bestätigt sich aber auch an der Anzahl der Neologismen, die vom Leibniz-Institut für Deutsche Sprache (IDS) für das Deutsche im Kontext der „Coronakrise" aufgezeichnet wurden.[258] Allerdings bleibt abzuwarten, welche und vor allem wie viele der 2.433 gesammelten Wortschöpfungen (Stand: November 2022) mittelfristig im Sprachgebrauch erhalten bleiben bzw. lexikalisiert werden.[259]

— 4. Auffällig ist zudem die besonders hohe Lexikalisierung im **5. und 6. Jh.**, die andeuten könnte, dass zu dieser Zeit besonders viele neue Wörter Eingang in die chinesische Sprache fanden. Ein plausibler historischer Grund hierfür wäre die verstärkte Verbreitung des Buddhismus in China.[260] Durch die Übersetzung von Sutren gelangen damit zahlreiche Sanskrit-Begriffe in die chinesische Sprache.[261] Ein Abgleich der gefundenen Lexeme mit dem buddhistischen chinesischen Wörterbuch von William E. Soothill und Lewis Hodous[262] zeigt jedoch, dass dies nur einen verhältnismäßig kleinen Beitrag leistet, während die Hauptursache in der hohen Gewichtung des *Hou Han shu* 後漢書 als Primärquelle gesehen werden kann.[263] Vor dem Hintergrund, dass ein Kriterium der Herausgeber des *HYDCD* war, dass die Aufnahme von Fachvokabular beschränkt sein sollte auf Begriffe, die zum allgemeingebräuchlichen Wortschatz gezählt werden können,[264] ist die im Verhältnis geringe Lexikalisierung buddhistischer Termini wenig überraschend.

— 5. Die Auswahl der Belegstellen wurde durch Neigungen bzw. Präferenzen der Herausgeber:innen und auch durch das zur Verfügung stehende Material beeinflusst, so dass gut erschlossene Texte unverhältnismäßig häufig zitiert werden. Ein Vergleich mit dem *OED*, für das das Vorhandensein von Konkordanzen offensichtlich die Auswahl der Belegstellen beeinflusst hat, legt dies ebenfalls nahe.[265] Ein dadurch entstehendes ***Bias*** bedingt, dass die Lexikalisierung der betrachteten Jahrhunderte unterschiedlich gut dokumentiert ist.

Berücksichtigt man die zusätzlichen, früheren Belegstellen aus *zhengshi* 正史 und Loewe-Korpora[266] (Abb. 5.9) lässt sich eine teilweise Verschiebung nach links beobachten, am grundsätzlichen Verlauf des Balkendiagramms ändert sich aber kaum etwas. Das 4. Jh. v. u. Z. weist nunmehr die höchste Konzentration von *Loci classici* auf, was klar auf die Gewichtung des verwendeten Loewe-Korpus zurückzuführen ist, aus dem ein hoher Anteil der ergänzten Belegstellen stammt.[267] Ließe man die zyklischen bzw. zufälligen Schwankungen in der

257 Eder 2018, S. 363, übersetzt durch den Verfasser. Eder untersucht Sprachwandel im Englischen mithilfe des *Google n-Gram Service* und entdeckt entsprechende *peaks* im Kontext des amerikanischen Bürgerkriegs in den 1870er Jahren und der *Great Depression* in den 1920er Jahren.

258 Leibniz-Institut für Deutsche Sprache (IDS), Hrsg. 2020: *Neuer Wortschatz rund um die Coronapandemie*. Online-Neologismenwörterbuch OWID. Mannheim. URL: https://www.owid.de/docs/neo/listen/corona.jsp (besucht am 09. 11. 2022).

259 Vgl. auch Jing-Schmidt und Hsieh 2019, S. 523: „[T]he majority of new words in fact fail to become established in language."

260 Siehe z. B. Vogelsang 2012, S. 219.

261 Wang Li 王力 2011 [1958], S. 590–591. Siehe auch Kapitel 2.2, ab S. 16.

262 William E. Soothill und Lewis Hodous 2003 [1937]: *A Dictionary of Chinese Buddhist Terms*. Online Version. URL: http://mahajana.net/texts/soothill-hodous.html (besucht am 28. 11. 2017).

263 Siehe Abschnitt 5.7.4, ab S. 150; ausführlicher dazu siehe auch T. Schalmey 2020, S. 79 u. S. 84–85.

264 „[…]对专科词的收录以进入一般语词范围的为限[…]" Yu Zhangrui 余章瑞 1988.

265 Siehe Kapitel 5.2, 111.

266 Siehe dazu auch Abschnitt 5.5.4, S. 134.

267 Siehe dazu Kapitel 4.2, ab S. 66.

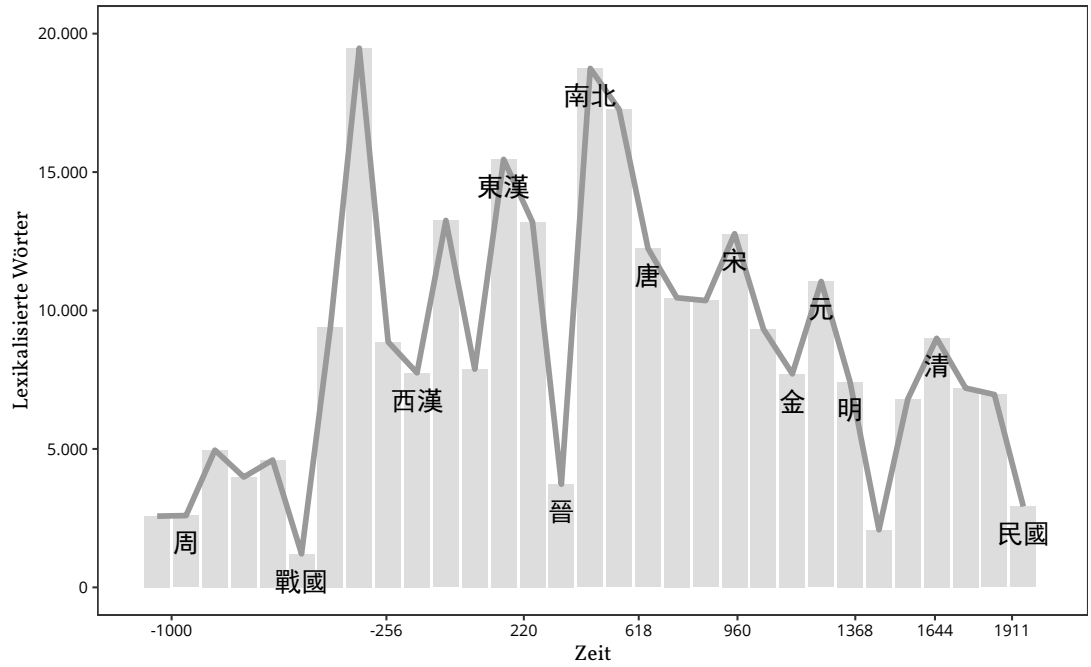


Abbildung 5.9 Lexikalisierung mit zusätzlichen Korpusbelegstellen aus Abschnitt 5.5.4

Lexikalisierung außer Acht, so kann insgesamt (abgesehen von den Randbereichen) eine gleichbleibend hohe Lexikalisierung mit durchschnittlich 8.727 neuen Einträgen pro Jahrhundert im *DHYDCD* beobachtet werden.[268]

Dass gerade zu Beginn und zum Ende des Betrachtungszeitraums eine geringere Lexikalisierung zu beobachten ist, legt einen Blick auf das kumulative Wachstum des Wortschatzes nahe. Besitzt das Piotrowski-Gesetz[269] auch einen Erklärungsgehalt für das Wortschatzwachstum im Chinesischen? Betrachten wir anhand der über die Belegstellen im *DHYDCD* datierbaren Lexikalisierung das kumulative Wortschatzwachstum pro Jahrhundert, so folgt es tatsächlich einer *s*-förmigen Kurve (Abb. 5.10).[270] Vom 5. Jh. v. u. Z. bis zum Ende des 3. Jhs., sowie vom 14. bis 20. Jh. sind zudem kürzere *s*-Kurven innerhalb des Gesamtverlaufs erkennnbar.[271] Auch wenn entsprechende Gesetzmäßigkeiten Spekulation bleiben müssen, spricht vieles dafür, von einem natürlichen, logistischen Wortschatzwachstum auszugehen, welches durch einschneidende historische Ereignisse beeinflusst werden kann.

268 Berücksichtigt werden dabei nur die insgesamt 270.525 Einträge in `the_words`, die sich chronologisch einordnen lassen. Einträge ohne Belege, bzw. mit Belegen ohne ausreichende bibliographische Daten, können nicht gezählt werden. Die *tatsächliche* Lexikalisierung würde bedeutend höher ausfallen. Auch die unvermeidbare Unvollständigkeit des Wörterbuchs sollte nicht vergessen werden.

269 Siehe Kapitel 2.1, ab S. 14.

270 Die Kurve der idealisierten *s*-förmigen Lexikalisierung in Abb. 5.10 wird in *R* mithilfe der Funktion `drm` (*Dose-Response Model*) geschätzt. Sie ist Teil des Pakets `drc`, das sich primär an Epidemiolog:innen richtet. Siehe Christian Ritz 2016: *drc Analysis of Dose-Response Curves, Version 3.0-1. R* package. URL: `rdocumentation.org/packages/drc/versions/3.0-1` (besucht am 10. 02. 2021).

271 Vgl. auch Aitchison 2001 [1991], S. 92: „A closer look at each *S*-curve, however, suggests that many *S*-curves are themselves composed of smaller *S*-curves. Each little *S*-curve covers one particular linguistic environment."

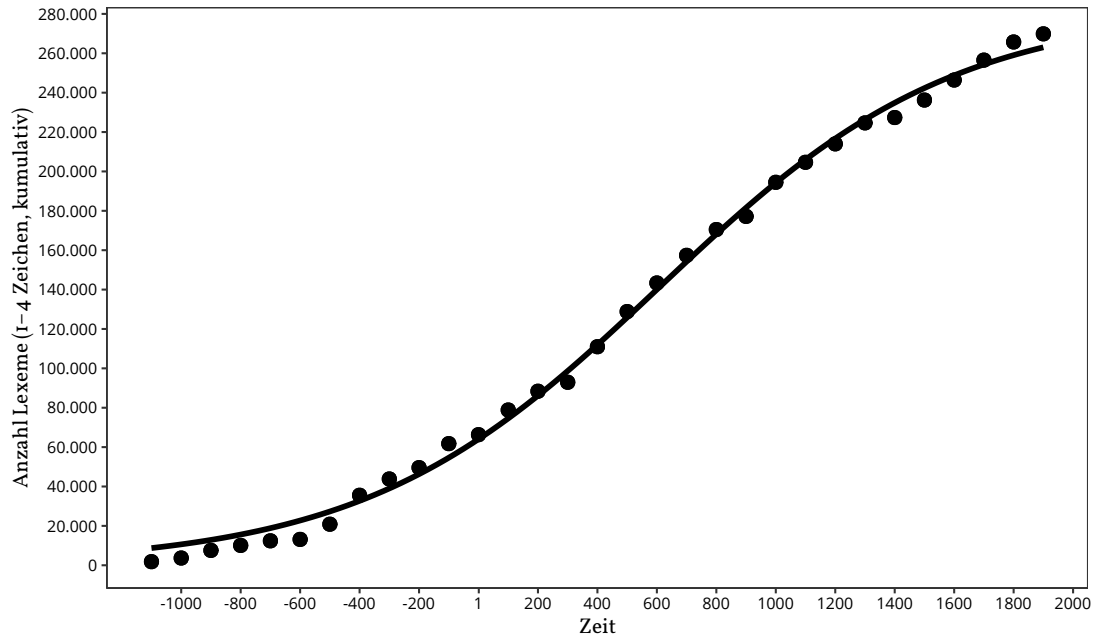


Abbildung 5.10 Lexikalisierung im *DHYDCD* nach Jahrhundert (ohne zusätzliche Belegstellen)

5.7.3 Mono- und Polysyllabizität

Ein Großteil der heute verwendeten Schriftzeichen stand bereits spätestens während der Han 漢-Zeit zur Verfügung. Große Zeichenwörterbücher wie das *Hanyu da zidian* 漢語大字典[272] beweisen zwar eindrucksvoll, dass auch in den folgenden knapp zwanzig Jahrhunderten kontinuierlich neue Zeichen entstanden sind. von denen sich allerdings nur wenige durchsetzen konnten. Die Rigidität der chinesischen Schrift bzw. der tatsächlichen Zeichennutzung zeigt sich anhand der Belege im *DHYDCD*. Die Lexikalisierung neuer Schriftzeichen (Abb. 5.11) nimmt im zeitlichen Verlauf tendenziell klar ab.[273]
Diese Beobachtung steht scheinbar im Widerspruch zu der von Best und Zhu Jinyang beobachteten „Zunahme der Schriftzeichen“[274] – es bleibt dabei allerdings anzumerken, dass im *DHYDCD* kaum historische oder lokale Zeichenvarianten aufgeführt sind.

Auf der anderen Seite ist als Trend erkennbar, dass die Lexikalisierung 3- und 4-silbiger Wörter, die in frühen Texten noch wenig belegt sind, im Laufe des Betrachtungszeitraums insgesamt kontinuierlich zunimmt (Abb. 5.12). Dabei überwiegen zunächst 4-silbige Lexeme klar – im Laufe der Jahrhunderte wird diese Verteilung aber zunehmend gleichmäßiger. Trotz der scheinbaren Prävalenz quadrisyllabischer Ausdrücke wie *chengyu* 成語 in der chinesischen Sprache, werden also ca. ab dem 7. Jh. ungefähr gleich viele trisyllabische Wörter lexikalisiert.[275]

272 *HYDZD*.
273 Ausführlicher dazu siehe auch T. Schalmey 2020, S. 80–82.
274 Best und Zhu Jinyang 2006, S. 208.
275 Siehe dazu auch T. Schalmey 2020, S. 81–82. Über die tatsächliche Häufigkeit dieser Lexeme in Texten kann hier natürlich keine Aussage getroffen werden.

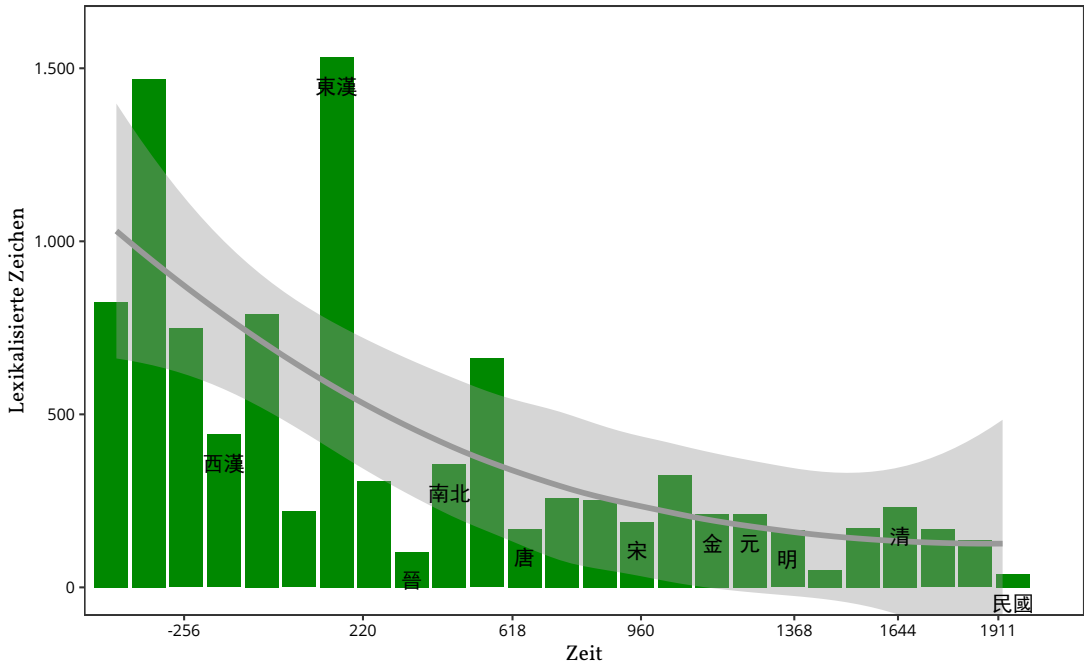


Abbildung 5.11 Lexikalisierung neuer Schriftzeichen im *DHYDCD*

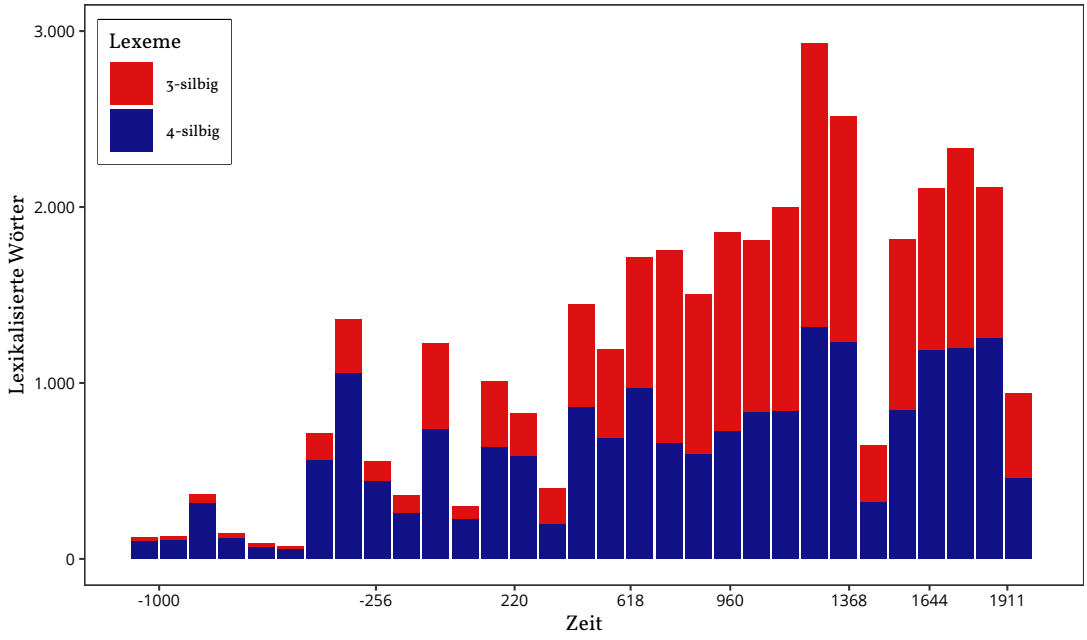


Abbildung 5.12 Lexikalisierung 3- und 4-silbiger Wörter

Die insgesamt klar zunehmende Entstehung mehrsilbiger Wörter mag den steigenden Bedarf daran widerspiegeln, komplexere oder zahlreichere Konzepte sprachlich eindeutiger auszudrücken. Bei langfristiger diachroner Betrachtung spricht sie auch dagegen, dass das allgemein für Sprachwandel typische „crunching“,[276] das Kompakter-werden sprachlicher Ausdrücke für das Chinesische uneingeschränkt zutrifft.[277] Auch im Hinblick auf Übertragungen schriftsprachlicher (*wenyan* 文言) oder gar klassischer (*guwen* 古文) Texte in die schriftliche Form der modernen Umgangssprache *baihua wen* 白話文 lässt sich diese Beobachtung für die chinesische Sprache nicht allgemein bestätigen.

Gleichzeitig stellt eine Länge von vier Zeichen ein typisches Maximum dar. Zwar sind im *DHYDCD* vereinzelt bis zu 17-silbige Ausdrücke lexikalisiert, doch bereits der Anteil an 5-silbigen ist verschwindend gering (Abb. 5.13).[278]

Der zeitliche Verlauf der Aufnahme neuer disyllabischer Lexeme entspricht etwa dem der gesamten Lexikalisierung.[279] Werden die Gesamtanteile *aller* datierbaren *DHYDCD*-Lexeme nach Anzahl der Silben betrachtet, ist das wenig überraschend: über 80 % aller Lexeme sind disyllabisch (Abb. 5.13). Der hohe Anteil erklärt sich durch eine starke Präferenz für Zusammensetzungen.[280] Typische Beispiele umfassen Wortbildungen aus zwei bedeutungsgleichen oder -ähnlichen Morphemen wie *bao+hu* 保護 („beschützen“), *xiao+shou* 銷售 („verkaufen“) oder *gou+mai* 購買 („kaufen“), Abkürzungen wie *Beida* 北大 („Uni Peking“), sowie einsilbige Ortsbezeichnungen, die eine Kategorieangabe erfordern, wie *faguo* 法國 („*Fa*-Land“, Frankreich).[281]

Für die moderne Hochsprache beobachtet Breiter auf Basis des *Xiandai Hanyu pinlü cidian* 現代漢語頻率詞典 *(Häufigkeitswörterbuch der modernen chinesischen Sprache)* eine ähnliche Verteilung, die sich lediglich durch einen deutlich höheren Anteil an monosyllabischen Lexemen unterscheidet (1-silbig 11,90 %, 2-silbig, 73 %, 3-silbig ca. 8,7 %, 4-silbig ca. 6,5 %, längere Lexeme weniger als 1 %). Untersucht man anstatt des Vorhandenseins von Lexemen die Wortlängenverteilung in *Texten*, bzw. in verschiedenen Textgattungen, ergibt sich ein anderes Bild. Sogar im modernen *putonghua* 普通話 dominieren einsilbige Wörter mit 64,3 % des von Breiter untersuchten Korpus.[282] Während in literarischen Texten einsilbige *tokens* den größten Anteil ausmachen, dominieren in juristischen, wissenschaftlichen- und Zeitungstexten zweisilbige Wörter.[283] Auch

276 Aitchison 2001 [1991], S. 116.

277 Vgl. Reinhard Köhler 1986: *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Quantitative linguistics 31. Bochum: Dr. N. Brockmeyer, S. 75–78. Köhler sieht die „Minimierung des Produktionsaufwandes“ als „Systembedürfnis“, welches Sprachwandel bedingt und sich auf die Länge und Komplexität sprachlicher Ausdrücke auswirkt. Ronald Langacker bezeichnet Sprachen sogar als „gigantic expression-compacting machines“, die sprachliche Ausdrücke typischerweise im Laufe der Jahrhunderte kompakter werden lassen. Siehe Ronald W. Langacker 1977: „Syntactic reanalysis“. In: *Mechanisms of Syntactic change*. Hrsg. von Charles N. Li. Austin: University of Texas Press, S. 57–139, S. 106; zitiert in Aitchison 2001 [1991], S. 116; Andererseits vermuten Arapov und Cherc einen Zusammenhang zwischen Silbenlänge und Alter von Wörtern dahingehend, dass neuere Wörter häufiger eine höhere Anzahl an Silben aufweisen. Siehe Mikhail V. Arapov und Maja M. Cherc 1983 [1974]: *Mathematische Methoden in der historischen Linguistik [Matematičeskiye metody v istoričeskoy lingvistike, Математические методы в исторической лингвистике]*. Übers. von Reinhard Köhler und Peter Schmidt. Quantitative Linguistics 17. Bochum [Moskau]: Dr. N. Brockmeyer [Nauka], S. 49–50.

278 Siehe auch Kapitel 4.5.2, S. 92.

279 Siehe Abb. 5.8, S. 143.

280 Ausführlicher dazu siehe Wong Kam-Fai 黃錦輝 et al. 2010, S. 11–18.

281 Siehe z. B. Lü Shuxiang 呂叔湘 1963: „Xiandai Hanyu danshuang yinjie wenti chutan 现代汉语单双音节问题初探 (Vorläufige Studie zum Problem von Mono- und Disyllabizität im modernen Chinesischen)“. In: *Zhongguo yuwen* 中国语文 1, S. 10–22; zitiert in Wong Kam-Fai 黃錦輝 et al. 2010, S. 10.

282 Siehe Breiter 1994, 224ff. zitiert in Schindelin 2005a, S. 959.

283 Siehe Zhu Jinyang und Karl-Heinz Best 1992: „Zum Wort im modernen Chinesisch“. In: *Oriens extremus* 35, S. 45–60, S. 52f. zitiert in Schindelin 2005a, S. 960.

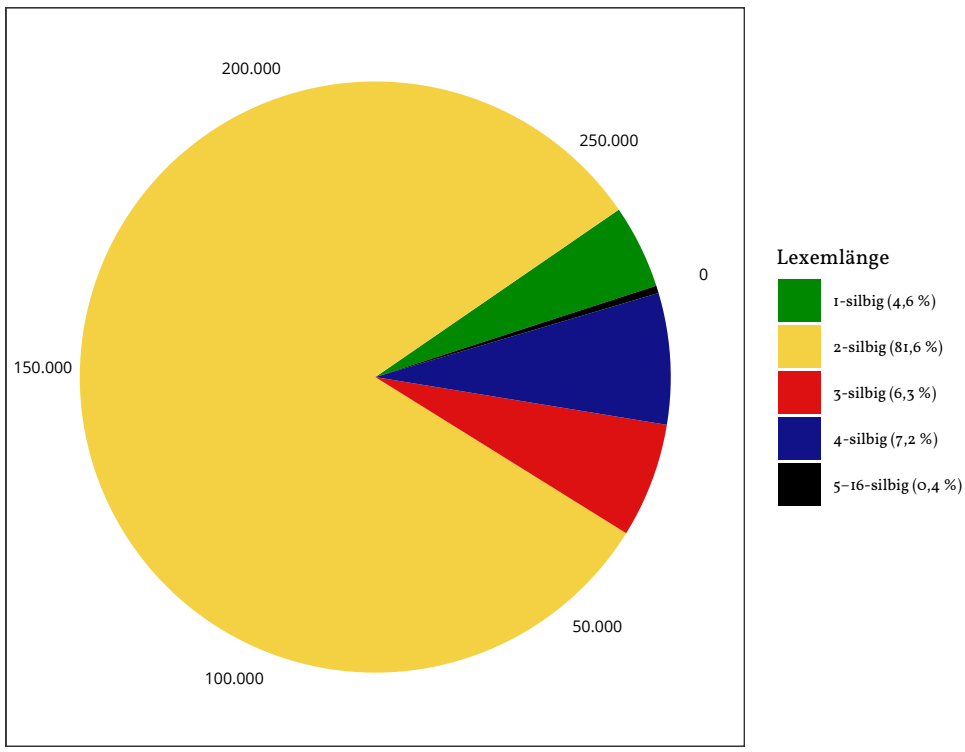


Abbildung 5.13 Chronologisierbare Lexikalisierung im *DHYDCD* nach Länge der Lexeme

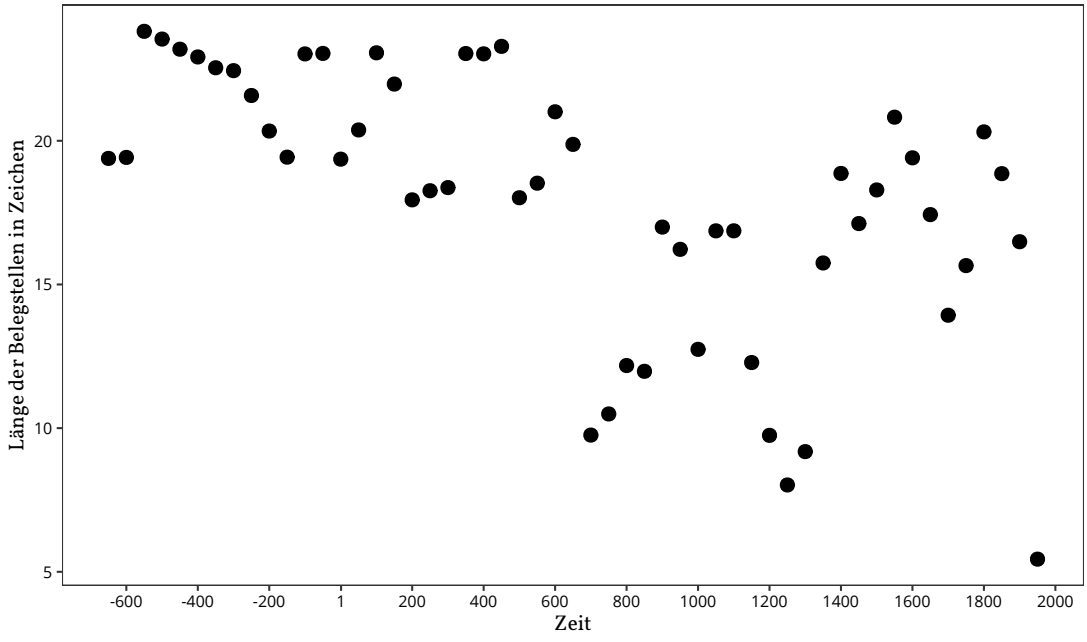


Abbildung 5.14 Länge der Belegstellen im *DHYDCD* nach Jahrhundert (in *zi* 字, inkl. Interpunktion)

über eine diachrone Analyse von Texten aus den vergangenen 2.500 Jahren lässt sich die zunehmende durchschnittliche Wortlänge statistisch nachweisen.[284]

Ungeachtet der gestiegenen Länge lexikalisierter und verwendeter Wörter, nimmt die Länge der Belegstellen im *DHYDCD* bei diachroner Betrachtung nicht zu. Abgesehen von wenigen Ausreißern schwankt sie um ein konstantes Niveau von knapp 20 Zeichen und scheint insgesamt sogar eher abzunehmen (Abb. 5.14). Hoffmann zeigt bei einer Untersuchung der *attestations* aus dem 11.–20. Jh. im englischsprachigen *OED* hingegen durchaus eine Zunahme der Länge der Textbelegstellen.[285] Bei den vorliegenden Daten scheint die Ableitung langfristiger, sprachgeschichtlicher Trends in Bezug auf die Satzlänge zu spekulativ, auch da die Belegstellen gekürzt sein können.

5.7.4 Lexikalisierung nach *Locus classicus*

Auffällig an den *Locus classicus*-Angaben im *DHYDCD* ist, dass einige Werke unverhältnismäßig oft genannt werden. In Tabelle 5.3 sind die 30 am häufigsten als früheste Belegstelle angegebenen aufgelistet. Zusammen werden sie für etwas mehr als 100.000 Lexeme (ca. 30,7 % der belegten Einträge) herangezogen.[286] Dominant sind die frühesten Textzeugnisse und wichtige kanonische und philosophische Texte vertreten, deren Datierung ist aber leider „nicht immer mit der Genauigkeit möglich, die sich Leser vielleicht wünschen würden“[287] Besonders häufig werden auch *zhengshi* 正史 wie *Hou Han shu* 後漢書 (*HHS*), *Han shu* 漢書, *Shiji* 史記, *Jin shu* 晉書 usw. angeführt.

Tabelle 5.3 Die 30 häufigsten *Locus classicus*-Angaben im *DHYDCD*

#	Text (Konkordanz)	Datierung	*Locus classicus*-Angaben	Länge in 1.000 Zeichen
1	*Hou Han shu (H 41)* 後漢書	ca. 400–500	9.091	917,5
2	*Han shu (H 36)* 漢書	92	8.115	712,2
3	*Shiji (H 40)* 史記	-91	7.468	512,6
4	*Shijing (HS 9)* 詩經	ca. -1100–-700	5.547	30,5
5	*Wen xuan [H 26]* 文選	520–530	5.278	996,2
6	*Zuo zhuan (HS 11)* 左傳	ca. -500–-400	5.058	180,5
7	*Liji (H 27)* 禮記	ca. -400–-300	4.513	98
8	*Jin shu [H 32]* 晉書	646	4.351	1.167
9	*Xin Tang shu [H 32]* 新唐書	1060	4.303	1.800,8
10	*Zhou li (H 37)* 周禮	ca. -150–23	3.777	52,8
11	*Sanguo zhi (H 33)* 三國志	ca. 280–297	3.551	390,4
12	*Shangshu (TJ)* 尚書	ca. -1100–-300	3.410	25,7
13	*Zhuangzi (HS 20)* 莊子	ca. -400–-200	3.036	65,1
14	*Song shu* 宋書	492–493	2.698	811,1
15	*Guanzi* 管子	ca. -720–-645	2.546	57,5
16	*Hong lou meng* 紅樓夢	ca. 1730–1764	2.324	731,1

284 Siehe Best und Zhu Jinyang 2006, S. 209–211. Best und Zhu beobachten ein quasi lineares Wachstum der Wortlänge, von etwa 1,15 Zeichen vor- und während der Han 漢-Zeit, auf bis zu 1,8 im 20. Jh., stellen aber ebenfalls eine breite Streuung fest.

285 Siehe Hoffmann 2004, S. 25. Er betont aber, dass die Zitatlänge „proves to be fairly constant, particularly for the time between 1450 and the end of the 19th century“ und begründet eine starke Zunahme im 20. Jh. mit der Präferenz der Herausgeber:innen der 2. Ausgabe für mehr Kontext.

286 Hinter den *Pinyin*-Titeln sind vorhandene Konkordanzen bzw. Indexbände dazu angegeben (*H = Harvard-Yenching*, *HS = Harvard-Yenching Supplement*, *TJ = Shangshu tongjian*), siehe auch Fußnote 293. Datierungen sind im Wesentlichen den entsprechenden Artikeln aus Loewe 1993 bzw. Wilkinson 2000, S. 503–505, entnommen.

287 Loewe 1993, S. xi, übersetzt durch den Verfasser.

Tabelle 5.3 (Fortsetzung)

#	Text (Konkordanz)	Datierung	*Locus classicus*-Angaben	Länge in 1.000 Zeichen
17	*Chu ci* 楚辭	ca. -329–-278	2.154	29,8
18	*Song shi [H 34]* 宋史	1345	2.069	4.037
19	*Shui hu zhuan* 水滸傳	ca. 1320–1372	2.063	437,3
20	*Yijing (HS 10)* 易	ca. -850–-800	2.033	21,6
21	*Huainanzi* 淮南子	-139	1.969	130,8
22	*Xunzi (HS 22)* 荀子	ca. -300–-238	1.943	64,9
23	*Guoyu* 國語	ca. -500–-300	1.918	70,4
24	*Nan shi* 南史	ca. 643–659	1.748	676,2
25	*Wei shu* 魏書	ca. 551–554	1.716	999
26	*Han Feizi* 韓非子	ca. -350	1.703	108,9
27	*Baopuzi* 抱樸子	ca. 265–420	1.534	152,2
28	*Ernü yingxiong zhuan* 兒女英雄傳	1878	1.502	472,1
29	*Jiu Tang shu* 舊唐書	945	1.496	2.001,9
30	*Lun heng* 論衡	80	1.442	164,1

Legt man die Daten aus Tabelle 5.3 und Abb. 5.8 (S. 143) übereinander, wird der Einfluss der meistzitierten Texte noch deutlicher (Abb. 5.15, S. 152). Die grauen Balken stellen wie in Abb. 5.8 die Gesamtlexikalisierung jeweils eines Jahrhunderts dar. Die 30 am häufigsten zitierten Texte sind gemäß ihrer in Tabelle 5.3 angebenen Datierung (x) und der Häufigkeit der *Locus classicus*-Angabe darübergelegt.[288] Die Visualisierung veranschaulicht, wie Zitate aus einzelnen Texten für einen großen Teil der Lexikalisierung des jeweiligen Jahrhunderts „verantwortlich" sein können. Als Extrembeispiel sticht das *Hou Han shu* 後漢書 (*HHS*) heraus: Mit über 9.000 darin belegten Lexemen ist es nicht nur Primärquelle für einen Großteil der Lexikalisierung aus dem 5. Jh., es ist auch der am häufigsten im *DHYDCD* als älteste Belegstelle angeführte Text.

Zwar wurde das *HHS* in seiner heute erhaltenen Form von Fan Ye 范曄 (398–445) erst im 5. Jh. kompiliert, es entstand aber in erster Linie aus überlieferten Materialien der östlichen Han 東漢-Zeit (25–220) wie dem *Dongguan Hanji* 東觀漢記.[289] Die im *HHS* enthaltenen „neuen" Lexeme dürften also größtenteils spätestens Han-zeitlich sein und die Datierung der Lexikalisierung über das *HHS* kann damit als teilweise „verspätet" angesehen werden. Ähnliches gilt sicherlich auch für die anderen prominent vertretenen *zhengshi*-Texte, sowie das im 6. Jh. kompilierte *Wenxuan* 文選 (mehr als 5.000 Angaben), eine heterogene Zusammenstellung von Texten, die teilweise mehrere Jahrhunderte früher datieren.[290]

Hieran zeigt sich einmal mehr, dass sich aus dem *DHYDCD* zwar Belege extrahieren lassen, wann ein Wort *spätestens* sicher belegt ist, ein Teil der so datierten Lexeme aber durchaus bereits mehrere Jahrhunderte früher verwendet worden sein kann. Die Gewichtung einzelner Texte liefert zudem einen wichtigen Erklärungsansatz für die Schwankungen in der Neulexikalisierung pro Jahrhundert[291] – sie können in der Arbeitsweise der Herausgeber:innen des *HYDCD* bei der Auswahl der *attestations* begründet liegen. Vermutlich wurden wichtige, leicht zugängliche Texte

288 Die Datierung früher Texte wie *Shijing* 詩經 oder *Yijing* 易經 kann bestenfalls eine grobe Schätzung sein, die sich über mehrere Jahrhunderte erstreckt. Da die Texte als Mittelpunkte dieser Perioden dargestellt werden, kommt es hier vereinzelt zu einer höheren y-Platzierung als die der Gesamtlexikalisierung des jeweiligen. Jh. Die Zuordnung der dadurch ungenau datierbaren Lexeme erfolgt anteilig auf die Jahrhunderte der Schätzperiode (siehe dazu Kapitel 6.2.1, S. 184)

289 Zur Entstehungsgeschichte des *Hou Han shu* siehe Bielenstein 1954, S. 9–17.

290 Siehe auch T. Schalmey 2020, S. 79.

291 Vgl. Abschnitt 5.7.2, ab S. 142.

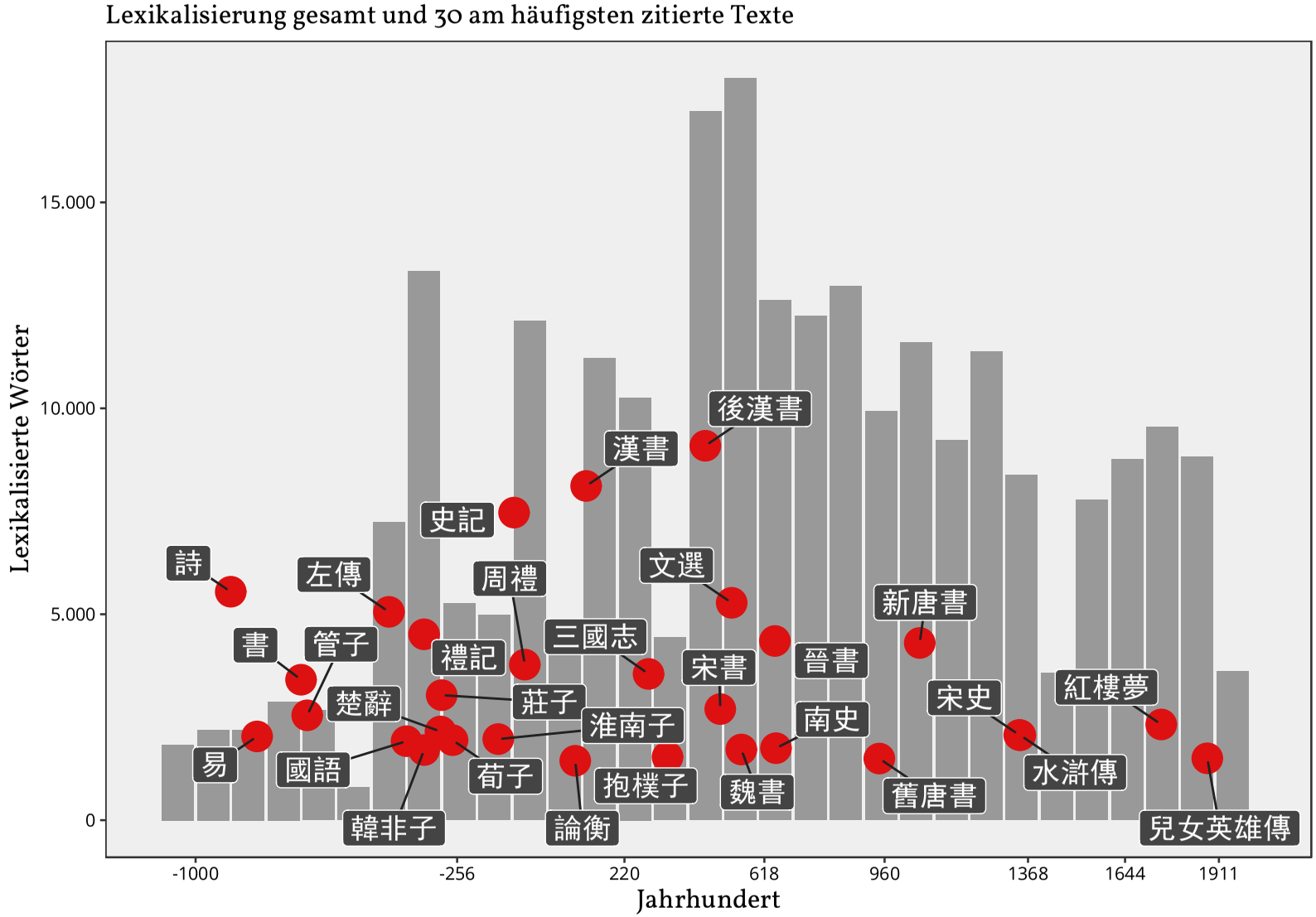


Abbildung 5.15 Lexikalisierung im *DHYDCD* nach Jahrhundert – häufigste *Locus classicus*-Texte

genau gelesen und daraus Belege exzerpiert. Auch wenn man sicher bemüht war, für jedes Lexem das ältestmögliche Textbeispiel zu zitieren, mag die Auswahl von der Versuchung beeinflusst gewesen sein, auf Konkordanzen bzw. Indizes zurückzugreifen. Dass – ähnlich wie beim *OED* – mit solchen Hilfsmitteln gearbeitet wurde, liegt nahe,[292] da gerade für die sehr häufig zitierten Texte wie *HHS*, *Han shu* 漢書, *Shiji* 史記, *Shijing* 詩經, *Sanguo zhi* 三國志 und weitere entsprechende Bände aus der Harvard Yenching-Reihe sinologischer Indizes vorliegen, die ab 1931 von Hong Ye 洪業 (William Hung) et al. herausgegeben wurde.[293] Als Beispiel sei der Eintrag zu *zhong gui ren* 中貴人 genannt. Darin wird die Biographie des Generals Li (*Li jiangjun liezhuan* 李將軍列傳) aus dem *Shiji* zitiert[294] – dieselbe Textstelle, die auch im entsprechenden Index unter *zhong gui ren* als erstes angegeben ist.[295] Die Prominenz früher Texte wie *Yijing* 易經, *Shangshu* 尚書 und *Shijing* 詩經 ist selbstverständlich, da aus der abgedeckten Zeit sonst wenig umfangreiches Textmaterial erhalten ist. Davon abgesehen prägt eine Vorliebe für die offiziellen Dynastiegeschichten (*zhengshi* 正史) die Liste häufig zitierter Texte.

Betrachtet man die 30 unabhängig von der Angabe als früheste Belegstelle meistzitierten Texte (Tabelle 5.4), ergibt sich eine ähnliche Liste. Einige wichtige Romane wie *Hong lou meng* 紅樓夢, *Shui hu zhuan* 水滸傳 und *Ru lin wai shi* 儒林外史 rücken auf, weitere Texte mit umgangssprachlichen Elementen wie die Geschichtensammlung *Liao zhai zhi yi* 聊齋志異, sind ebenfalls häufiger vertreten. Den 30 meistzitierten Texten sind 21,7 % aller Belege entnommen.[296]

Tabelle 5.4 30 meistzitierte Werke im *DHYDCD*

#	Text	Datierung	Belegstellen	Länge in 1.000 Zeichen
1	*Hou Han Shu* 後漢書	ca. 400–500	9.091	917,5
2	*Han shu* 漢書	111	8.115	712,2
3	*Shiji* 史記	-94	7.468	512,6
4	*Xin Tang shu* 新唐書	1060	5.547	1.800,8
5	*Wenxuan* 文選	ca. 520–530	5.278	996,2
6	*Jin shu* 晉書	648	5.058	1.167
7	*Zuo zhuan* 左傳	ca. -500–-400	4.513	180,5
8	*Shijing* 詩經	ca. -1100–-700	4.351	30,5

292 Siehe Kapitel 5.2, ab S. 111.

293 Siehe Tabelle 5.3, S. 150; vgl. u. a. Hong Ye 洪業 (William Hung), Hrsg. 1966 [1949]: *Combined indices to Hou Han shu and the notes of Liu Chao and Li Hsien (*後漢書及注釋綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 41. Taipei 台北 [Beijing 北京]: Harvard-Yenching [Yanjing xue she 燕京學社]; Hong Ye 洪業 (William Hung) et al., Hrsg. 1966 [1940]: *Combined indices to Han Shu and the notes of Yen Shih-ku and Wang Hsien-ch'ien (Hanshu ji buzhu zonghe yinde* 漢書及補註綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京學社引得) 36. Taipei 台北 [Beijing 北京]: Harvard Yenching Institute [Yanjing da xue tu shu guan 燕京大學圖書館]; Hong Ye 洪業 (William Hung), Hrsg. 1955 [1947]: *Combined indices to Shih chi and the notes of P'ei Yin, Ssu-ma Cheng, Chang Shou-chieh, and Takigawa Kametaro (Shi ji ji zhu shi zong he yin de* 史記及注釋綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 40. Cambridge, MA [Beijing 北京]: Harvard University Press [Yenching University Press]; Hong Ye 洪業 (William Hung) et al., Hrsg. 1934: *A concordance to Shih ching (Mao shi yin de* 毛詩引得*)*. Harvard-Yenching Institute Sinological Index Series Supplement (Hafo Yanjing xue she yin de te 哈佛燕京大學圖書館引得特刊) 9. Beijing 北京: Harvard-Yenching (Hafo Yanjing xueshe 哈佛燕京學社); Hong Ye 洪業 (William Hung) et al., Hrsg. 1938: *Combined Indices to San Kuo Chih and the Notes of P'ei Sung-chih (*三國志及裴注綜合引得*)*. Harvard-Yenching Institute Sinological Index Series (Hafo Yanjing xue she yin de 哈佛燕京大學圖書館引得) 33. Beijing 北京: Harvard-Yenching (Hafo Yanjing xueshe 哈佛燕京學社).

294 *DHYDCD*, 中貴人.

295 Siehe Hong Ye 洪業 (William Hung) 1955 [1947], S. 62.

296 Insgesamt werden nach den in Abschnitt 5.5.2 (ab S. 128) beschriebenen Kritierien mehr als 155.000 Quellenangaben unterschieden. Eine Liste der am häufigsten zitierten 20.804 würde 80 %, 63.727 dann 90 % der insgesamt 919.280 identifizierten Belegstellen abdecken.

Tabelle 5.4 (Fortsetzung)

#	Text (Konkordanz)	Datierung	*Locus classicus*-Angaben	Länge in 1.000 Zeichen
9	*Hong lou meng* 紅樓夢	ca. 1730–1764	4.303	731,1
10	*Liji* 禮記	ca. -400–-300	3.777	98
11	*Sanguo zhi* 三國志	ca. 280–297	3.551	390,4
12	*Shui hu zhuan* 水滸傳	ca. 1320–1372	3.410	437,3
13	*Song shi* 宋史	1345	3.036	4.037
14	*Zhou li* 周禮	ca. -150–23	2.698	52,8
15	*Song shu* 宋書	ca. 492–493	2.546	811,1
16	*Liaozhai zhiyi* 聊齋志異	1740	2.324	381,4
17	*Shangshu* 尚書	ca. -1100–-300	2.154	25,7
18	*Jiu Tang shu* 舊唐書	945	2.069	2.001,9
19	*Ernü yingxiong zhuan* 兒女英雄傳	1878	2.063	472,1
20	*Ming shi* 明史	1643	2.033	2.081,9
21	*Nan shi* 南史	ca. 643–659	1.969	676,2
22	*Zhuangzi* 莊子	ca. -400–-200	1.943	65,1
23	*Huainanzi* 淮南子	ca. -139	1.918	130,8
24	*Guanzi* 管子	ca. -720–-645	1.748	57,5
25	*Guoyu* 國語	ca. -500–-300	1.716	70,4
26	*Baopuzi* 抱樸子	ca. 265–420	1.703	152,2
27	*Xunzi* 荀子	ca. -300–-238	1.534	64,9
28	*Zi zhi tong jian* 資治通鑒	1084	1.502	1.933,1
29	*Chu ci* 楚辭	-329–-278	1.496	29,8
30	*Ru lin wai shi* 儒林外史	1749	1.442	231,9

Die Auswahl der am häufigsten im *DHYDCD* zitierten Texte zeigt eine insgesamt stark in der klassischen Tradition verwurzelte Textrezeption durch die Herausgeber:innen, die durchaus an das *KXZD* erinnert. Im Gegensatz zu diesem wird aber zumindest einer Auswahl an moderneren, umgangssprachlichen Quellen ebenfalls Gewicht verliehen.

Der Schwerpunkt dieses Kapitels lag in der Erschließung des *DHYDCD* als diachrone Lexemdatenbank und der Analyse dieser Ressource. Die Datenbank dient als Basis für die Entwicklung der Datierungsmethoden für schriftsprachliche chinesische Texte, die in Kapitel 6.2 (ab S. 179) und 6.3 (ab S. 210) beschrieben werden. Die Erzeugung eines diachronen Behelfskorpus aus dem *DHYDCD* erlaubt es zudem, die in Kapitel 6.1 (ab S. 156) für schriftsprachliche chinesische Texte adaptierten Methoden für einen langen Betrachtungszeitraum zu evaluieren.