6 Textdatierung für schriftsprachliches Chinesisch

> *„While the majority of diachronic studies focus on change in language, we should also not forget the flipside of language change, stability over time, which is equally interesting."*[1]
>
> Vaclav Brezina

Dieses Kapitel widmet sich computerlinguistischen Methoden zur chronologischen Einordnung chinesischsprachiger Texte. Wie in Kapitel 3.3 (ab S. 45) skizziert, lassen sich im Wesentlichen zwei Arten von Datierungsaufgaben unterscheiden: **1.** Die Klärung der Frage, *wann* ein Text etwa verfasst wurde und **2.** eine temporale Einordnung des *Inhalts*, z. B. in welcher Zeit sich die Handlung eines literarischen Texts abspielt, bzw. welcher Zeitraum in einem historiographischen Text beschrieben wird. Hier soll es primär um erstere Herausforderung gehen.

Dass bislang keine Arbeiten zur Datierung chinesischsprachiger Texte vorliegen, in denen die in Kapitel 3.3 (ab S. 45) beschriebenen Methoden eingesetzt werden, ist vermutlich auf den Mangel an geeigneten (Trainings-)korpora zurückzuführen.[2] Für den Zeitraum von 1475 bis ca. 1925 steht mit dem von CrossAsia veröffentlichten *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌) ein großer Datensatz mit den Häufigkeiten der 1–3-Gramme von 11.081 Lokalchroniken (*difangzhi* 地方誌, *DFZ*) zur Verfügung,[3] deren Entstehungszeit bekannt ist und aus dem sich dadurch entsprechende *Statistical Language Models* (*SLM*) berechnen lassen.

In Kapitel 6.1 (ab S. 156) implementiere ich die wesentlichen Ansätze aus den in Kapitel 3.3 vorgestellten Methoden, die auf solchen *chronon*-Sprachmodellen basieren, um ihre Eignung für die (domänenspezifische) Datierung chinesischsprachiger Texte zu testen.

Um eine genreübergreifende Datierung von Texten über einen größeren Zeitraum hinweg zu ermöglichen, arbeite ich zudem zur Erzeugung der temporalen Sprachmodelle mit dem Behelfskorpus aus den Beispielsätzen des *DHYDCD*.[4]

Da die teilweise Rigidität einiger Genres der chinesischen Schriftsprache und das Fehlen umfangreicherer Trainingskorpora die Zuverlässigkeit der Datierung bei der Verwendung dieser Methoden auf Basis statistischer Sprachmodelle stark einschränken können, wird zudem ein innovativer, datenbankgestützter Ansatz zur chronologischen Einordnung von Texten vorgestellt, dem die historischen Lexikalisierungsdaten aus dem *DHYDCD* zugrunde liegen.[5] Werden

1 Vaclav Brezina 2018: *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge & New York: Cambridge University Press, S. 220.

2 Siehe auch Kapitel 4.2, S. 62.

3 *DFZ*, siehe auch Kapitel 4.2, S. 62. Zwar sind in diesem Datensatz vereinzelt deutlich frühere Texte aus der Zeit der Tang 唐-Dynastie (618–907) enthalten, jedoch ist erst für den Zeitraum ab dem Ende des 15. Jahrhunderts eine ausreichende Menge an Texten vorhanden, um sowohl Trainings-, als auch Testdatensätze zur Verfügung zu haben. Die Veröffentlichung erfolgt als *n*-Gramme, da eine Bereitstellung der Volltexte lizenzrechtlich problematisch wäre.

4 Siehe Kapitel 5.6, ab S. 137.

5 Siehe Kapitel 5, S. 107–154.

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

diese für alle in einem Text enthaltenen Lexeme herangezogen, lässt sich ein *Neologismusprofil* erzeugen, welches das (stilistische) Alter des Textes visualisiert (Kapitel 6.2, ab S. 179) und von Philolog:innen zu Analysezwecken herangezogen werden kann, um die zeitliche Einordnung von Texten zu erleichtern. Mittels Regression auf die Daten solcher Neologismusprofile lassen sich überdies für die automatisierte Datierung schriftsprachlicher chinesischer Texte teilweise bessere Ergebnisse erzielen, als dies mit den für die Datierung westlichsprachiger Texte gängigen *SLMs* möglich ist.

Auch mit dem experimentellen Ansatz von Berechnungen auf Basis des durchschnittlichen Jahres der Lexikalisierung der in einem Text enthaltenen Lexeme (*Average Year of Lexicalization*, *AYL*) lassen sich überraschende Ergebnisse erzielen. Mittels Regression kann so ebenfalls das Alter eines Textes geschätzt werden, allerdings – wegen der stilistischen Unterschiede zwischen unterschiedlichen Textgattungen – nur in einem eng gesteckten Rahmen (Kapitel 6.3, ab S. 210).

Zuletzt werden die wesentlichen Ergebnisse und Unterschiede zwischen den untersuchten Datierungsmethoden, sowie sich daraus ergebende Herausforderungen und Limitationen zusammengefasst und eine Benutzer:innenoberfläche für die erarbeiteten Möglichkeiten präsentiert (Kapitel 6.4, ab S. 229).

6.1 Datierung als Kategorisierungsproblem

„What am I gonna do?
I'm gonna train myself a classifier*!"*
Paul Vierthaler

Für die Textdatierung auf Basis statistischer Sprachmodelle werden zunächst aggregierte Wort- bzw. *n*-Gramm Häufigkeiten für bestimmte Zeitabschnitte (*chronons*) berechnet. Ein zu datierender Text *d* wird dann dem *chronon* mit der größten Übereinstimmung zugeordnet.[6] Anhand des *DFZ*-Datensatzes,[7] sollen zunächst folgende Fragen ergründet werden:

1. Sind statistische Sprachmodelle aus Trainingsdaten dieses Korpus geeignet, um andere Texte innerhalb desselben Korpus zu datieren?
2. Mit welchen Ähnlichkeitsmaßen lässt sich die beste Performance für die Datierung chinesischsprachiger Texte erzielen?
3. Wie wirkt sich eine Reduktion auf „Lexemdimensionen" und die Erweiterung dieser um Namen, Ortsnamen und temporale Ausdrücke auf die Performance aus?
4. Wie wirken sich weitere Parameter, z. B. Mindesthäufigkeiten als Möglichkeit der Reduktion der zu betrachtenden Dimensionen (*features*) aus?
5. Wie verhält es sich mit einer Erweiterung oder Reduktion des verwendeten *n*-Gramm-Raums auf 1-Gramme, 1–2-, 1–3-Gramme?
6. Kann die Performance der Modelle durch Glättung (*smoothing*) verbessert werden?

Die Performance der Modelle und Ähnlichkeitsmaße wird dabei am Anteil der „richtig" datierten Texte (*Accuracy*), sowie am mittleren Abstand der jeweiligen Datierung von den tatsächlichen

6 Ausführlicher in Kapitel 3.3, ab S. 45.
7 Siehe auch Kapitel 4.2, S. 66.

Jahresangaben aus den Metadaten der Texte (*mean error*) gemessen. Als *richtig* werden diejenigen Datierungen angesehen, bei denen das Jahr der Veröffentlichung bzw. das mittlere Jahr des Erscheinungszeitraums (`cleanyear`) in das datierte *chronon* fällt. Durch die Überlappung der *chronons* können also in der Regel bis zu zwei *chronons* als richtig gewertet werden.

Der *mean error* D_{mean} der Datierung eines Textes ist dabei definiert als arithmetisches Mittel des Intervalls der Jahre von Beginn und Ende des datierten *chronons* zum `cleanyear` des zu datierenden Textes:

$$D_{mean} = \frac{(|D_{start}| + |D_{end}|)}{2}$$

Ist also z. B. die Veröffentlichung eines Texts mit 1525–1527 angegeben, der korrekt auf das *chronon* 1500–1550 datiert wird, so beträgt D_{mean} 25 Jahre. Dieser minimale Wert ergibt sich bei jeder *richtigen* Datierung eines Textes:

$$D_{mean} = \frac{(|\frac{1527+1525}{2} - 1500| + |\frac{1527+1525}{2} - 1550|)}{2} = \frac{(|1526 - 1500| + |1526 - 1550|)}{2} = \frac{(26 + 24)}{2} = 25$$

Für eine Gesamtanzahl *N* an Testtexten ist der *mean average error* (*MAE*) also angegeben als

$$MAE = \frac{1}{N}\sum_{n=1}^{N} D_{mean}$$

und spiegelt so die ungefähre Genauigkeit einer Datierung wider.

Als Ähnlichkeitsmaße werden *Cosine similarity* (*CS*) ohne und mit Gewichtung mittels *inverse document frequency* (*idf*),[8] Jaccard *similarity*, sowie *Normalized-Log-Likelihood-Ratio* (*NLLR*) und Kullback-Leibler-Divergenz (*KLD*) getestet, jeweils ohne und mit Gewichtung mittels temporaler Entropie (TE).[9]

Ebenfalls anhand des *DFZ*-Datensatzes soll zudem erörtert werden, ob eine Co-Datierung von Dokumenten gegenüber der Verwendung von *chronon*-Modellen vorzuziehen ist.[10] Um den Vergleich auf eine solidere Grundlage zu stellen, wird dieses Experiment mit einem weiteren Datensatz, dem *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*)[11] wiederholt.[12]

Um in der Lage zu sein, Texte unabhängig von ihrer Zugehörigkeit zu einem stilistisch und vor allem temporal stark eingeschränkten Korpus zeitlich einordnen zu können, werden zudem Experimente mit Sprachmodellen durchgeführt, die aus dem Behelfskorpus der *DHYDCD*-Belegstellen erzeugt werden.[13] Anhand derselben Daten lassen sich zudem grobe quantitative Beobachtungen über Veränderungen der Wortnutzung und des Wortschatzes in dem von diesem Korpus abgedeckten Zeitraum von 700 v. u. Z. bis zum 20. Jh. machen.

8 Siehe dazu S. 168; siehe auch Kapitel 3.3, S. 53.
9 Ausführlich dazu siehe Kapitel 3.3, ab S. 45.
10 Vgl. Kapitel 3.3, S. 50; siehe auch de Jong, Rode und Hiemstra 2005, S. 7; Bamman et al. 2017, S. 4.
11 CrossAsia, Staatsbibliothek zu Berlin 2019b: *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書. Datenset. Version 0.0.1-20190307. DOI: 10.5281/zenodo.2586940.
12 Siehe Abschnitt 6.1.2, ab S. 169.
13 Siehe 6.1.3, ab S. 171.

6.1.1 Datierung mit *difangzhi* 地方誌 Sprachmodellen

Aus dem *DFZ*-Datensatz wird zunächst ein temporales Sprachmodell mit 17 *chronons* für einen Zeitraum von 450 Jahren erzeugt. Hierfür wird eine *chronon*-Dauer von 50 Jahren mit jeweils 25-jähriger Überlappung gewählt.[14] Das erste *chronon* deckt den Zeitraum von 1475–1525 ab, das zweite 1500–1550 und das letzte die Jahre 1875–1925.[15]

Für jedes *chronon* werden jeweils 50 zufällig ausgewählte Lokalchroniken als Trainingsdaten verwendet.[16] Dabei werden nur Texte berücksichtigt, für die beide chronologischen Einordnungen (Angaben zum Inhalt und zeitliche Angaben zu Herausgeber- bzw. Autorschaft) vorliegen. Außerdem werden Texte, bei denen beide Angaben 50 oder mehr Jahre auseinander liegen, ausgeschlossen. So soll sichergestellt werden, dass es sich nicht um eine (überarbeitete) Neuausgabe eines eigentlich deutlich älteren Texts handelt und die Verfasserschaft tatsächlich ungefähr in den angegebenen Veröffentlichungszeitraum fällt. Ist anstatt des Erscheinungsjahres ein Zeitraum angegeben, wird das mittlere Jahr dieses Zeitraums für die chronologische Einordnung verwendet, also z. B. 1909 für 1908–1910. Durch die Überlappung der *chronons* kann sich die Textauswahl für die Trainings-Subkorpora teilweise überschneiden.[17] Für alle *chronons* werden die absoluten (*tokens*) und relativen Häufigkeiten (*tf*), sowie temporale Entropie (TE) und *term frequency inverse document frequency* (*tf-idf*) vorberechnet.[18] Außerdem werden die Gesamthäufigkeiten des Trainingskorpus vorgehalten. Da der *DFZ*-Datensatz zahlreiche Variantenzeichen (*yitizi* 異體字) enthält, wird die in Kapitel 4.3 beschriebene Normalisierung lediglich zur Erkennung mehrsilbiger Ausdrücke als Lexeme verwendet, um einen potenziellen Datenverlust zu minimieren.[19] Als *Baseline* für die unterschiedlichen Ähnlichkeitsmaße dient ein Zufallsgenerator, der eines der 17 *chronons* auswählt und theoretisch mit einer Wahrscheinlichkeit von etwa 11 % die Testdokumente korrekt datieren wird.

Werden solche Sprachmodelle für alle 1–2-Gramme berechnet, entsteht eine Dokumentensammlung mit insgesamt mehr als 7 Mio. *n*-Gramm *types*. Etwa 40 % davon sind *Hapax legomena* mit nur einem einzigen Vorkommen im Korpus, die nur mit sehr geringer Wahrscheinlichkeit in einem *Query*-Dokument *d* vorkommen. Berechnungen mit einer so großen Zahl an Vektoren sind zudem wenig performant.

Obwohl die *chronon*-Sprachmodelle jeweils aus der gleichen Anzahl von Texten aggregiert werden, kann die Anzahl der enthaltenen *types* stark schwanken. So enthält z. B. das *chronon* Sprachmodell von 1550–1600 ca. 1,5 Mio. *n*-Gramm *types*, das Modell von 1850–1900 mit über 2,1 Mio. deutlich mehr.[20] Diese Diskrepanz kann bei der Berechnung einiger Ähnlichkeitsmaße problematisch sein, da große *chronons* so lediglich deswegen „ähnlicher" zu einem *Query*-Dokument sein können, weil mehr Dimensionen verglichen werden können. Dieser Effekt kann durch die Verwendung von ausgeglichenen Modellen reduziert werden, indem alle *chronons* so lange um seltene *types* reduziert werden, bis die Anzahl der betrachteten *n*-Gramm-Vektoren

14 Die Erzeugung versetzt überlappender *chronons* ermöglicht einen besseren Umgang mit *chronon*-„Rändern". Siehe dazu de Jong, Rode und Hiemstra 2005, S. 4.

15 Der Datensatz enthält zwar vereinzelt auch deutlich ältere Texte ab dem 8. Jh. Als Trainingsdaten reichen diese aber noch nicht aus. Erst ab der Ming-Zeit ist eine substanzielle Anzahl an *DFZ* erhalten geblieben. Vgl. auch Dennis 2015, S. 2.

16 Aus der Datenqualität des Korpus ergeben sich dabei Einschränkungen. Die Kritierien für die Auswahl der Texte werden in Kapitel 4.2 dargelegt (siehe S. 67).

17 Bei der verwendeten Auswahl werden etwa 10 % der Texte in zwei *chronons* verwendet.

18 Siehe dazu Kapitel 3.3. S. 53, S. 53.

19 Siehe S. 70; siehe auch Kapitel 4.2, S. 66.

20 Insgesamt ist im Betrachtungszeitraum ein eindeutiger Trend zu längeren Texten mit mehr *types* nachweisbar.

derjenigen des „kürzesten“ *chronons* entspricht. Durch eine derartige Reduktion der Dimensionen können andererseits allerdings auch Merkmale eliminiert werden, die eine temporale Diskriminierung der Dokumente überhaupt erst ermöglichen würden.[21]

Als Testdaten werden unter Ausschluss der Trainingsdaten 216 zufällig ausgewählte *Difangzhi* verwendet, deren Veröffentlichung gleichmäßig über den Betrachtungszeitraum von 1475–1925 verteilt ist. Die Häufigkeit $P(\hat{\theta} \mid c)$ von Wörtern aus einem zu datierenden Text d, die in einem Vergleichs-*chronon* c nicht auftreten, aber für die Berechnung von *NLLR* und *KLD* benötigt werden (sogenannte *unseen events*),[22] wird zunächst angenommen als Hälfte der niedrigsten Häufigkeit eines Wortes im längsten *chronon* c_{long}, also $P(\hat{\theta} \mid c) = \lambda \times P_{min}(w \mid c_{long})$ mit $\lambda = 0,5$. Dadurch soll vereinfacht modelliert werden, dass die Wahrscheinlichkeit dieses *unseen events* niedriger ist als die geringste tatsächlich in einem *chronon* auftretende Worthäufigkeit. Diese naive Glättungsmethode bezeichne ich im Folgenden als *Largest chronon minimum* (*LCM*) *smoothing*.[23]

Experimente mit *difangzhi* 地方誌-Sprachmodellen

Im Folgenden werden Experimente zur Beantwortung der eingangs formulierten Fragen durchgeführt. Tabelle 6.1 (S. 165) gibt einen Gesamtüberblick über die Ergebnisse dieser Untersuchungen.

1–2-Gramme.
Als Basismodell wird ein vollständiges (d. h. bei Verwendung aller verfügbaren *types* unabhängig von ihrer Häufigkeit und Sinnhaftigkeit) 1–2-Gramm Modell verwendet.[24] Während die Texte mit Jaccard *similarity* kaum besser datiert werden als mit einem Zufallsgenerator, stellen sich alle anderen verwendeten Metriken als aussagekräftig heraus. *Tf-idf* zur Gewichtung der *CS* verbessert die Performance im Vergleich zur einfachen *CS* erheblich, während die aufwändigere Gewichtung von *KLD* und *NLLR* mit temporaler Entropie zunächst eher einen negativen Effekt zu haben scheint (Abb. 6.1). Ein Grund dafür ist sicherlich die große Menge an *Hapaxen*, bzw. dass *types*, die ein einziges Mal in einem *chronon* auftreten tendenziell übergewichtet werden, da ihr seltenes Auftreten auch zufällig sein kann, ohne ein tatsächliches Indiz für die Entstehungszeit des Textes zu sein. *KLD* und *NLLR* sind tendenziell aussagekräftiger als CS_{tfidf}, *KLD* und *NLLR* quasi gleichwertig.[25]

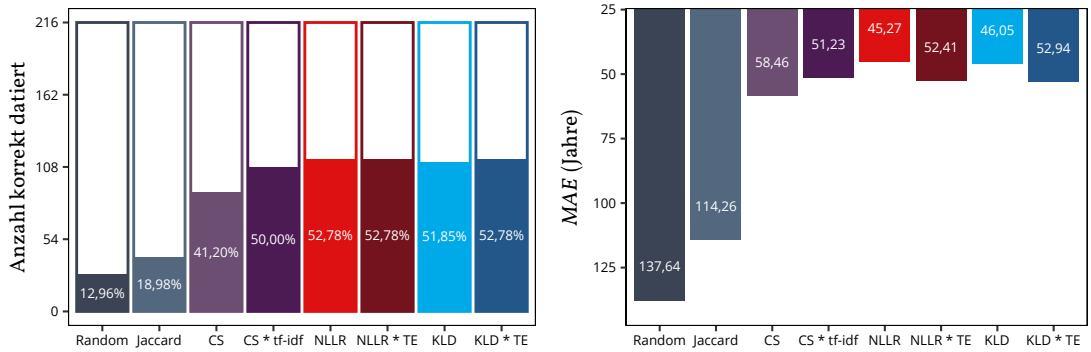
21 Alternativ könnte ein Ausgleich über eine unterschiedliche Anzahl von Texten oder eine unterschiedliche Länge der *chronons* erreicht werden.

22 Siehe Kapitel 3.3, ab S. 45.

23 Diese Herangehensweise ist prinzipiell ähnlich einem *add one* bzw. Laplace-*smoothing*, aber ohne Veränderung der Grundmasse bzw. ohne die tatsächliche Vorkommenshäufigkeit zu verändern. Die Verwendung anspruchsvollerer *smoothing*-Methoden wird in Abschnitt 6.1.1 (S. 164) beschrieben und untersucht.

24 Dies ist nicht direkt mit der Verwendung eines Bigramm-Modells für westliche Sprachen vergleichbar, da Bigramme hier sowohl zwei Wörter mit je einem Zeichen, ein Wort aus zwei Zeichen, oder die Kombination aus dem letzten Zeichen eines Wortes mit mehreren Schriftzeichen und dem ersten Zeichen des nächsten Wortes usw. repräsentieren können. Siehe auch Kapitel 3.3, S. 47.

25 Vgl. auch Kraaij 2004, S. 208.



(a) Anteil richtig datierter *difangzhi* **(b)** Durchschnittliche Abweichung in Jahren (*MAE*)

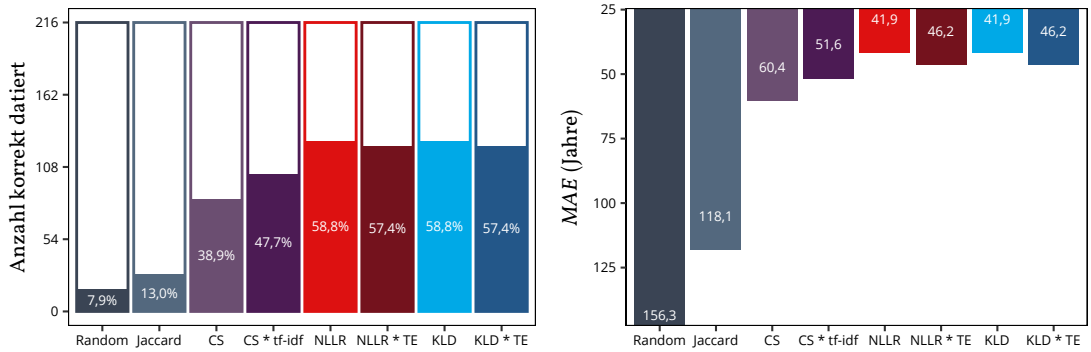
Abbildung 6.1 Performance mit 1–2-Gramm *difangzhi*-Sprachmodell

Reduktion auf Lexemdimensionen.
Durch eine Reduktion der verwendeten Dimensionen um etwa 96 % auf diejenigen 238.707 Uni- und Bigramme des Modells, die als Lexeme im *DHYDCD* aufgeführt sind, lässt sich die Dauer der Berechnung um etwa 95 % reduzieren. Die so erzeugten *SLM*s bezeichne ich im Folgenden als Lexem-Modelle. Im Unterschied zu einem klassischen *BoW*-Modell handelt es sich hierbei streng genommen um stark dimensionsreduzierte Zeichen-*n*-Gramm-Modelle.[26] Dabei werden zunächst auch Vorkommen von Einzelzeichen aus dem Modell eliminiert, wenn diese im *DHYDCD* nicht lexikalisiert sind und dadurch auch ca. 8.000 Unigramm-*types* bzw. chinesische Schriftzeichen weniger berücksichtigt, als das beim reinen *n*-Gramm Modell der Fall war.[27] Das Lexem-Modell enthält mit nur 21.750 (knapp 10 %) zudem deutlich weniger *Hapaxe* als das *n*-Gramm-Modell.[28] Die relativen Häufigkeiten und die Gewichte *tf-idf* und *TE* des so entstandenen Lexem-Sprachmodells werden sowohl für die *chronons* als auch das gesamte Modell auf Grundlage der Lexem-*tokens* neu berechnet. Trotz der Verwendung nur eines Bruchteils der *features* erhöht sich der Anteil der korrekt datierten Texte mit *NLLR* und *KLD* und der *MAE* sinkt um etwa 5 Jahre. Lediglich die Performance von *CS* bzw. CS_{tfidf} geht durch die Reduktion der Dimensionen leicht zurück (Abb. 6.2; Tabelle 6.1).

26 Siehe dazu auch Kapitel 4.5.3, S. 94.

27 Siehe auch Experiment Nr. 10 in Tabelle 6.1: Die Verwendung dieser zusätzlichen Zeichen für die Modellberechnung führt lediglich zu marginalen Unterschieden. Es handelt sich bei diesen Zeichen um im *DHYDCD* fehlende Einträge wie *bing* 丙 und *mei* 美 (siehe dazu Kapitel 5.4.1, S. 120), nicht automatisch normalisierte Zeichenvarianten, z. B. 吳 für *wu* 吳/吴, 出 für *chu* 出, 蒙 für *meng* 蒙 (siehe dazu Kapitel 4.3, ab S. 70), seltene Zeichen, die nicht im *HYDCD* lexikalisiert sind, wie *chun* 蓴 („Kräutermedizin"), *niang* 蘘 („Medizinkräuter"), einige Zeichen, die sehr wahrscheinlich durch Codierungsfehler entstanden sind, wie der javanische Buchstabe *ba* ꦧ, sowie Symbole aus Schriftsystemen ethnischer Minderheiten wie der *Yi* 彝.

28 In einer Studie zur Anwendbarkeit von Zipf's Gesetz auf das Chinesische stellt Xiao fest, dass in einem „großen", modernen Korpus über 40 % *Hapax legomena* auftreten. Siehe Xiao Hang 2008: „On the Applicability of Zipf's Law in Chinese Word Frequency Distribution". In: *Journal of Chinese Language and Computing* 18.1, S. 33–46, S. 44. Dies entspricht etwa dem hier (s. o.) für *n*-Gramme beobachteten Anteil, obwohl Xiao mit einem getaggten, segmentierten Korpus arbeitet. Es muss also vermutet werden, dass ein relevanter Teil der in den Texten enthaltenen Wörter nicht im *HYDCD* lexikalisiert ist.



(a) Anteil richtig datierter *difangzhi* **(b)** Durchschnittliche Abweichung in Jahren (*MAE*)

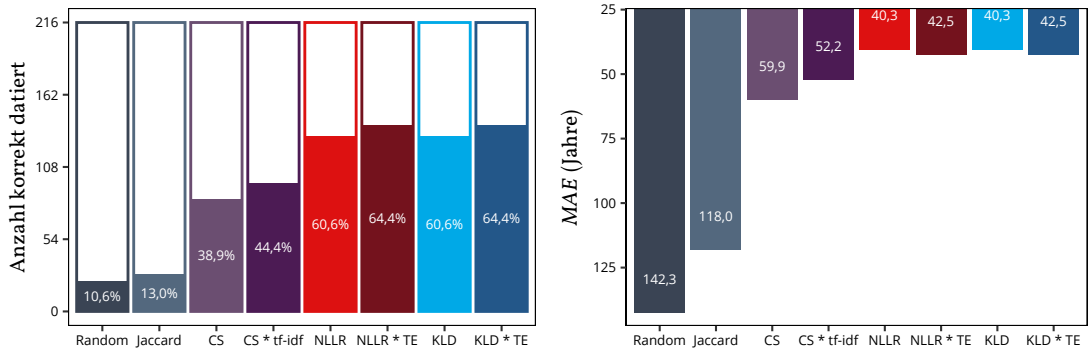
Abbildung 6.2 Performance von 1–2-Zeichen *difangzhi*-Lexem-Sprachmodellen

Verwendung gleichförmiger *chronons*.
Hierfür wird die Anzahl der *types* jedes *chronon*s durch Entfernen niedrigfrequenter *types* so lange reduziert, bis sie der Anzahl der *types* des „kürzesten" *chronons* (hier 112.281) entspricht, also eine weitere Dimensionsreduktion um je 0–27 %. Hierdurch ist nur eine entscheidende Veränderung in der Performance feststellbar: die Anzahl der mit Jaccard *similarity* richtig datierten Texte erhöht sich um fast 30 Prozentpunkte auf 41.7 % (ohne Abb.). Das Ergebnis reicht nicht an die komplexeren bzw. gewichteten Metriken heran, weist aber darauf hin, dass es für die Datierung schriftsprachlicher Texte – zumindest in diesem Textkorpus – vielleicht wichtiger ist, *welche types* in einem Text auftreten, als wie häufig sie auftreten und dass – wie auch die relative „Erfolglosigkeit" der Gewichtung mittels temporaler Entropie zeigt – der Sprachwandel im Hinblick auf die Veränderung der Häufigkeiten von Wörtern im Betrachtungszeitraum in diesem Korpus verhältnismäßig gering ist.

Festlegung einer Mindesthäufigkeit von *types* für die Berücksichtigung in Metriken. Durch die Festlegung einer Mindesthäufigkeit von *types* könnte der Effekt übergewichteter, seltener *types* minimiert werden. Erhöht man die Mindesthäufigkeit von *types* für die Erzeugung des Modells und die Berechnung der Ähnlichkeitsmaße auf 2, lässt sich allerdings keine nennenswerte Veränderung der Performance feststellen, da der positive Effekt der *feature reduction* vermutlich durch den Verlust seltener, diskriminativer *types* ausgeglichen wird.

Verwendung von Namen und Zeitausdrücken.
Während die Erweiterung der verwendbaren „sinnvollen" *types* um Personen- und Ortsnamen aus der *CBDB* keinen positiven Einfluss auf die verwendeten Metriken hat,[29] kann durch eine Verwendung von Zeitausdrücken[30] die Performance des bisher besten Modells (1–2 Gramm-Lexeme, keine Mindesthäufigkeit, ohne gleichförmige *chronons*), noch leicht gesteigert werden:

29 Die Ergebnisse der Experimente 7–14 sind in Tabelle 6.1, S. 165 zusammengetragen.
30 Zu diesem Zweck wird die *DHYDCD*-Lexemliste um Äranamen und *tiangan dizhi* 天干地支 Ausdrücke erweitert. Siehe dazu auch Kapitel 4.8, ab S. 103.



(a) Anteil richtig datierter *difangzhi* **(b)** Durchschnittliche Abweichung in Jahren (*MAE*)

Abbildung 6.3 Performance von 1–2-Zeichen *difangzhi*-Lexem-Sprachmodellen mit zusätzlichen temporalen Ausdrücken

Mit einer *Accuracy* von 0,64 zeigt sich erstmals auch ein spürbar positiver Effekt der Gewichtung mit TE. Diese funktioniert deutlich besser, wenn mehr häufigere Begriffe auch wirklich in verschiedenen *chronons* unterschiedlich stark auftreten, was bei „gewöhnlichen" Lexemen offensichtlich weniger der Fall ist. Ohne *TE* kann ein etwas geringerer *MAE* erzielt werden.

Abb. 6.4 zeigt die Detailergebnisse der Datierung von 216 *difangzhi* mit diesem *SLM* bei der Verwendung von *NLLR* und *TE*. Unterhalb der gestrichelten Linie ist ein etwas größerer Anteil als „zu alt" datierter Texte erkennbar. Insgesamt sind die Datierungen und Abweichungen über den gesamten Zeitraum etwa gleichmäßig verteilt.
Das Fehlen einer zeitlich diskriminativen Kraft von Ortsnamen ist wenig überraschend, da die Angaben zur ersten Nennung in der *CBDB* offensichtlich viele frühere Quellen unberücksichtigt lassen.[31] Der fehlende Effekt der Verwendung von Personennamen verdient jedoch weitere Aufmerksamkeit, insbesondere da in vielen *DFZ* mehrere hundert Übereinstimmungen mit Personennamen vorhanden sind[32] und erwartbar ist, dass in Chroniken zu unterschiedlichen Zeiten über unterschiedliche Akteure berichtet wird. Ursachen für eine ausbleibende Verbesserung der Datierungsergebnisse können in der bereits in Kapitel 4.7 erörterten Problematik liegen.[33] Zeichenfolgen, die in ihrer lexikalisierten Bedeutung vorkommen, können fälschlich als Name identifiziert werden. Hinzu kommt die Problematik mehrfacher Namensträger.[34] Möglich ist auch, dass unterschiedliche Erwähnungen von Personen in Lokalchroniken stärker eine geographische als eine temporale Zuordnung stützen.

In Tabelle 6.1 sind die Ergebnisse der oben beschriebenen Experimente zusammenfassend dargestellt. Die Spalte *DHYDCD* gibt an, ob die *types* auf die Lexeme des *DHYDCD* reduziert sind.[35] „Namen" bzw. „Orte" gibt an, ob der Lexem-Raum um die Personen- bzw. Ortsnamen aus der *CBDB* erweitert ist,[36] „Zeit" gibt an, ob die Zeitausdrücke aus der *DDBC*-Datenbank

31 Siehe auch Kapitel 4.7, ab S. 97.
32 Vgl. dazu Abb. 6.25 in Kapitel 6.2.
33 Siehe S. 97.
34 Siehe Kapitel 4.7, ab S. 97.
35 Die Angabe „+1-*grams*" (Experiment Nr. 10) bedeutet, dass hier nicht nur die lexikalisierten, sondern alle vorkommenden Einzelzeichen berücksichtigt wurden, also 7.912 zusätzliche *types*.
36 Siehe auch Kapitel 4.7, ab S. 97.

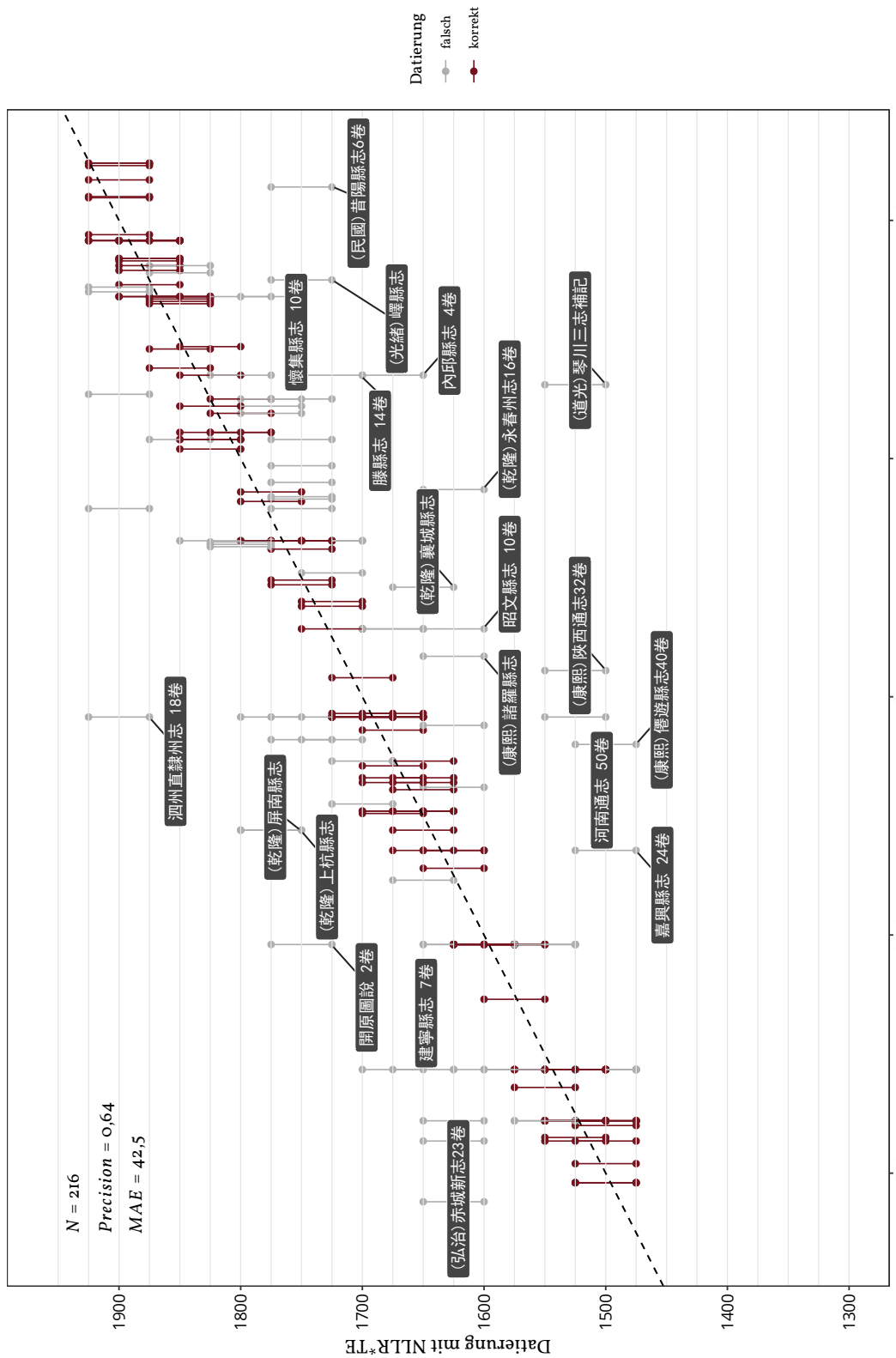


Abbildung 6.4 Performance von 1-2-Zeichen *difangzhi*-Lexem-Sprachmodellen mit zusätzlichen temporalen Ausdrücken

hinzugezogen wurden.[37] Experimente mit detaillierterer Darstellung sind mit Verweisen auf die entsprechenden Abbildungen versehen.

Smoothing von *DFZ*-Sprachmodellen

Wie beschrieben wurde bisher für alle *unseen events* in Vergleichs-*chronons* eine Wahrscheinlichkeit $P(\hat{\theta} \mid c) = \lambda P_{min}(w \mid c_{long})$ mit $\lambda = 0{,}5$ angenommen, also die Hälfte der niedrigsten Wahrscheinlichkeit des längsten *chronons* (*LCM smoothing*). Diese einfache Annahme, mit der *unseen events* zuverlässig eine niedrige Häufigkeit zugewiesen wird, scheint sich bisher gut zu bewähren. Im Gegensatz zu gewöhnlich im Kontext statistischer Sprachmodelle eingesetzter *Smoothing*-Techniken bleibt allerdings die Häufigkeit des jeweiligen Wortes bzw. *n*-Gramms im Korpus unberücksichtigt, d. h. allen *unseen events* wird dieselbe Häufigkeit zugeschrieben. Auch de Jong, Rode und Hiemstra verwenden eine niedrige Häufigkeit für *types*, die im Query-Dokument, aber nicht in einem *chronon*-Modell auftreten und stellen fest, dass die Auswirkung von ihnen getesteter *Smoothing*-Methoden (Dirichlet-*Smoothing*, *linear interpolation*) bei der Verwendung von *chronon*-Modellen gering ist.[38] Für die Verwendung ungewichteter *NLLR* trifft das auch für die *DFZ*-Modelle zu: Die Auswirkungen sowohl eines Weglassens der entsprechenden Dimensionen, als auch die Anpassung der angenommenen Häufigkeit in einem niedrigen Bereich zwischen 0,000.000.01 und 0,000.000.1 sind gering. Das mag damit zusammenhängen, dass der Anteil der *unseen events* aus Query-Dokumenten in *chronons* hier zwischen etwa 1 und 10 % liegt, also zur Berechnung von *KLD* oder *NLLR* nur 1–10 % der *types* überhaupt von diesem vereinfachten *smoothing* betroffen sind. Bei der Verwendung von Gewichten wie *TE* kann die Auswirkung der *Smoothing*-Methode und Parameter allerdings immens werden, da gerade diejenigen *types*, die nur in wenigen oder in einem *chronon*-Modell auftreten, eine hohe Entropie haben. Zhai Chengxiang, der sich intensiv mit *Smoothing*-Methoden im Kontext von *SLM*s beschäftigt, stellt fest, dass „nonoptimal smoothing can degrade retrieval performance significantly."[39]

Im Folgenden werden daher einige der im Rahmen von Textdatierungen mittels *SLMs* üblichen *Smoothing*-Methoden[40] auf ihre Eignung für die Datierung von *DFZ*-Texten überprüft.

1. Laplace / *add one smoothing*. Dabei wird angenommen, dass die Häufigkeit aller *events* um einen gewissen Wert λ höher ist. Dadurch erhalten *unseen events* in jedem *chronon* eine Häufigkeit von λ:

$$P(w, \theta) = \frac{c(w, D) + \lambda}{|D|}$$

   Mit $\lambda = 1$ wird von *add one smoothing* gesprochen. Relative Häufigkeiten und Gewichte werden entsprechend neu berechnet.

37 Siehe dazu auch Kapitel 4.8, ab S. 103.

38 Siehe de Jong, Rode und Hiemstra 2005, S. 3. Es wird eine „very small (non-zero) probability" verwendet, ohne auf die Anwendung bzw. Parameter-Bestimmung für das angeblich verwendete *linear interpolation smoothing* und *Dirichlet smoothing* einzugehen.

39 Siehe Zhai Chengxiang 2008: „Statistical Language Models for Information Retrieval: A Critical Review". In: *Foundations and Trends in Information Retrieval* 2.3, S. 137–213. doi: 10.1561/1500000008, S. 154.

40 Siehe dazu Kapitel 3.3, S. 54.

Tabelle 6.1 Ergebnisse der beschriebenen Experimente mit *difangzhi-SLM*

	Modell	*#types*	gleichförmige *chronons*	Beschränkung von *types* auf *DHYDCD*	Namen	Ortsnamen	Zeitausdrücke	◆ *NLLR* *Accuracy* (%)	◆ *NLLR* *MAE* (Jahre)	◆ *NLLR * TE* *Accuracy* (%)	◆ *NLLR * TE* *MAE* (Jahre)
1	1-*grams*	23.437	-	-	-	-	-	54,2	42,69	49,1	50,69
2	1–2 (Abb. 6.1)	7.400.012	-	-	-	-	-	52,8	45,27	52,8	52,41
3	1–2 (Abb. 6.2)	238.707	-	x	-	-	-	58,8	41,86	57,4	46,19
4	1–2	265.721	-	x	x	-	-	59,7	41,54	56,9	47,02
5	1–2	232.704	x	x	x	-	-	58,8	42,09	57,4	47,37
6	1–3	315.794	-	x	x	-	-	60,2	41,29	56,5	49,1
7	1–3	268.623	x	x	x	-	-	60,2	41,5	57,4	48,15
8	1–2	239.701	-	x	-	x	-	59,7	41,74	56,5	46,54
9	1–2 (Abb. 6.3)	238.978	-	x	-	-	x	60,6	**40,33**	**64,4**	42,54
10	1–2	246.890	-	+ 1-*grams*	-	-	x	59,7	**39,84**	60,6	42,65
11	1–2	266.954	-	x	x	x	x	59,7	41,36	61,1	45,4
12	1–2 (min. 2 *tokens*)	147.402	-	x	-	-	x	57,9	41,1	62	43,32
13	1–3	250.537	-	x	-	-	x	60,6	40,33	64,4	42,29
14	2-*grams*	223.452	-	x	-	-	x	62	42,18	61,1	43,94

2. Dirichlet *smoothing*:

$$P(w,\theta) = \frac{|Q|}{(|Q| + \mu)} \times P(w,Q) + \frac{\mu}{(\mu + |Q|)} \times P(w,C)$$

3. Jelinek-Mercer, auch *linear interpolation smoothing*[41] genannt:

$$P(w,\theta) = (1 - \beta) \times P(w,Q) + \beta \times P(w,C)$$

4. *Nearest neighbour smoothing*. Kanhabua und Nørvåg schlagen vor, eine Interpolation der Trainingsdaten dahingehend vorzunehmen, dass *unseen events* innerhalb eines *chronon* durch Daten aus benachbarten *chronons* ergänzt werden, je nach Datenlage durch Verwenden der niedrigsten Häufigkeit aus benachbarten *chronons* oder durch Verwendung des Durchschnittswertes aus Häufigkeiten der benachbarten *chronons*.[42] Diese Art der Interpolation bezeichne ich als *nearest neighbour smoothing*. Die Verwendung des Mittelwertes der benachbarten *chronons* scheint die Verwendung der Gewichtung durch *TE* absolut unbrauchbar zu machen (Experiment Nr. 5 in Tabelle 6.2). Ein Zusammenhang besteht sicherlich mit der stark unterschiedlichen Anzahl *types* in den *chronons*, wodurch die Häufigkeiten aus Nachbar-*chronons* über- oder untergewichtet werden können. Eine bessere Annahme scheint daher zu sein, dass die Häufigkeit eines *unseen events* die jeweils niedrigste aus den vorangehenden und nachfolgenden *chronons* ist (Experiment Nr. 4 in Tabelle 6.2). Allerdings kann auch so die *Accuracy* mit *NLLR * TE* und *LCM smoothing* nicht übertroffen werden.

Tabelle 6.2 gibt einen Überblick über die Ergebnisse.

Tabelle 6.2 Test von *Smoothing*-Methoden mit unterschiedlichen Parametern

			◆ *NLLR*		◆ *NLLR * TE*	
	Smoothing-Methode	Parameter	*Accuracy* (%)	*MAE* (Jahre)	*Accuracy* (%)	*MAE* (Jahre)
1	*Baseline*: Largest *chronon* minimum (Abb. 6.3)	$\lambda = 0{,}5$	60,6	**40,33**	**64,4**	42,54
2	Laplace (»add one«)	$\lambda = 1$	59,7	40,93	62	42,19
3	Laplace	$\lambda = 0{,}5$	60,6	41	12	200
4	Neighbour minimum & LCM	$\beta = 0{,}5$	**61,1**	40,63	55,6	48,36
5	Neighbour mean & LCM	$\beta = 0{,}5$	61,1	40,72	12,5	199,79
6	Jelinek-Mercer	$\beta = 0{,}01$	57,9	41,57	56,9	45,93
7	Jelinek-Mercer	$\beta = 0{,}1$	60,2	40,65	60,2	44,52
8	Jelinek-Mercer	$\beta = 0{,}4$	57,9	41,57	56,9	45,93
9	Jelinek-Mercer	$\beta = 0{,}9$	27,3	55,99	36,1	67,47
10	Dirichlet	$\mu = 0{,}1$	52,8	44,6	50,5	48,8
11	Dirichlet	$\mu = 0{,}4$	54,6	43,17	51,4	48,21
12	Dirichlet	$\mu = 0{,}9$	54,6	43,02	51,9	47,6

41 Siehe Kraaij 2004, S. 209.

42 Siehe Kanhabua und Nørvåg 2008, S. 362; Die Konsultation benachbarter *chronons* wird auch von A. Kumar 2013, S. 52, vorgeschlagen. Weder Kumar noch Kanhabua u. Nørvåg führen allerdings weitergehende Experimente mit dieser Art der Interpolation durch. Erst Bamman et al. 2017, verwenden erfolgreich ein *moving average* mit Daten der benachbarten *chronons*, wobei allerdings mit einem deutlich umfassenderen Korpus gearbeitet werden kann und alle Häufigkeiten durch diese Glättung angepasst werden. (Siehe S. 4.)

Fazit. Gegenüber der einfachen Annahme, die für das *LCM Smoothing* getroffen wird, lassen sich innerhalb dieses Korpus durch die Anwendung aufwändigerer *Smoothing*-Methoden keine Performanceverbesserungen bei der Verwendung statistischer Sprachmodelle zur Datierung erzielen. Die Häufigkeit von im Vergleichs-*chronon* nicht vorkommenden Lexemen sollte leicht unter der minimalen Häufigkeit liegen, die in einem *chronon* möglich ist. Dies bestätigt die ursprüngliche Annahme von de Jong, Rode und Hiemstra, dass die *Smoothing*-Effekte gering sind,[43] zeigt aber in einigen der durchgeführten Experimente auch, dass ungeeignete Glättungsmaßnahmen die Aussagekraft der Modelle massiv verschlechtern können.[44] In ihrem viel beachteten Aufsatz „An Empirical Study of Smoothing Techniques for Language Modeling" schreiben Chen und Goodman:

> Whenever data sparsity is an issue, smoothing can help performance, and data sparsity is almost always an issue in statistical modelling. In the extreme case where there is so much training data that all parameters can be accurately trained without smoothing, one can almost always expand the model, such as by moving to a higher-order *n*-gram model, to achieve improved performance. With more parameters, data sparsity becomes an issue again, but with proper smoothing the models are usually more accurate than the original models. […][45]

Durch die Verwendung der über einen Zeitraum von je 50 Jahren aggregierten *chronon*-Modelle aus jeweils 50 Texten sind die Trainingsdaten hier tatsächlich relativ umfangreich. Da *DFZ* nur als 1–3-Gramm Zählungen vorliegen, ist eine Erweiterung des „echten" *n*-Gramm Raumes nur sehr bedingt möglich, da keine Daten über die Reihenfolge des Auftretens der 2–3-Gramme vorhanden sind. Dass hier mit keiner der getesteten *Smoothing*-Techniken eine Verbesserung der Ergebnisse zu erzielen war, deutet an, dass durch das *Smoothing* regelmäßig in höherem Maß temporale Diskriminatoren „weggeglättet" werden, als durch die Glättung eine Verbesserung des Modells stattfindet. Es bleibt allerdings zu erforschen, wie eine Glättung sich bei der Verwendung deutlich kürzerer bzw. kleinerer *chronons*, oder sogar bei der Verwendung des „ähnlichsten" Dokuments zur Datierung (*document co-dating*) auswirken würden.

Notizen zu Berechnungen bei der Verwendung statistischer Sprachmodelle

— 1. ***Bag of words*** (*BoW*). Wie bereits angedeutet handelt es sich bei den hier als „Lexem-Modelle" bezeichneten Textrepräsentationen letztlich nicht um eine klassische *BoW*,[46] da bei Betrachtung von Wörtern mit einer Länge von 1–*n* Zeichen fast alle Vorkommen von Wörtern mit 2 oder mehr Zeichen Länge zusätzlich als Vorkommen der enthaltenen Einzelzeichen gezählt werden. Versuche, dies z. B. durch entsprechende Abzüge von Vorkommen längerer Wort-*n*-Gramme von den Unigramm-Zählungen auszugleichen, um eine stärkere Annäherung an eine tatsächliche *bag of **words*** zu erreichen, wirken sich aber tendenziell negativ auf die Leistungsfähigkeit der oben beschriebenen Modelle aus. Durch die Betrachtung von 2–*n*-Grammen wird der angesprochene Effekt allerdings ohnehin minimiert.

— 2. **Logarithmen**. Bei der Berechnung sowohl von *idf* und *TE* als auch von Ähnlichkeitsmaßen wie der *NLLR* werden zur Normalisierung Logarithmen verwendet. In der wissenschaftlichen Fachliteratur wird das Symbol *log* austauschbar sowohl für den *logarithmus naturalis* (*ln*) mit Basis

43 Siehe de Jong, Rode und Hiemstra 2005, S. 3.
44 Vgl. auch Chen und Goodman 1998, S. 59.
45 Ebd., S. 58.
46 Siehe auch Kapitel 4.5.3, S. 94.

e, als auch den Logarithmus mit Basis 2 (log_2) oder sogar 10 (log_{10}) verwendet. Obwohl „the precise base of the logarithm is not material to ranking“,[47] konnten oben durch Verwendung von log_2 für *NLLR* und *KLD* die eindeutig besten Ergebnisse erzielt werden.

— 3. ***Term frequency.*** Die *natürliche* bzw. rohe Worthäufigkeit $f_{w,d}$ ist definiert als Anzahl der Vorkommen von w in einem Dokument d.[48] Es ist offensichtlich, dass bei der Arbeit mit einem Korpus aus Dokumenten unterschiedlicher Länge eine Normalisierung erfolgen muss. In der Fachliteratur sind unterschiedliche Varianten einer solchen normalisierten *term frequency* verbreitet.

— 3.1 Normalisierung auf das häufigste Wort des Dokuments. $tf_{cmax}(w,d) = a + (1-a)\frac{f_{w,d}}{f_{max,d}}$, wobei a zwischen 0 und 1 gewählt werden kann und gewöhnlich auf 0,4 oder 0,5 gesetzt wird.[49] Mit $a = 1$ gilt dann $tf_{cmax}(w,d) = \frac{f_{w,d}}{c_{max,d}}$.

— 3.2 Normalisierung auf die Länge des Dokuments. Hierbei ist $|d|$ definiert als die Anzahl aller *tokens* in d.[50]

$$tf(w,d) = \frac{f_{w,d}}{|\,d\,|}$$

— 3.3 Zusätzlich kann die Größe des verfügbaren Wortschatzes $|\,V\,|$ betrachtet werden, also die Anzahl der unterschiedlichen *types* in d oder eines Korpus C,[51] z. B.:

$$tf_{vocsize}(w,d) = \frac{f_{w,d}}{|\,d\,| + |\,V\,|}$$

In einer experimentellen Überprüfung der Wirkung unterschiedlicher Interpretationen der *term frequency* auf die *Accuracy* des besten Modells (1–2 Zeichen Lexeme + temporale Ausdrücke) stellt sich der Unterschied zwischen den letzten beiden Varianten $tf_{vocsize}$ und tf erwartungsgemäß als marginal heraus, Die tf_{cmax} hingegen eignet sich nicht für Berechnungen mit *NLLR* oder *KLD*.[52]

— 4. ***Inverse document frequency (idf).*** Wie bei der tf gibt es auch für die zur Gewichtung verwendete inverse Dokumentenhäufigkeit idf zahlreiche Möglichkeiten der Berechnung. Buck und Koehn geben ohne Anspruch auf Vollständigkeit sechs unterschiedliche Definitionen an und bemerken, dass überdies „in der freien Wildbahn noch geringfügige Variationen dieser Definitionen zu finden sind.“[53] Aus Gründen der Einfachheit wird hier lediglich die in Kapitel 3.3 implizierte Variante verwendet:

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

47 Christopher D Manning, Prabhakar Raghavan und Hinrich Schütze 2008: *Introduction to Information Retrieval*. Cambridge & New York: Cambridge University Press, S. 109.
48 Siehe z. B. ebd., S. 107.
49 Siehe z. B. ebd., S. 117.
50 Vgl. z. B. Zhai Chengxiang 2008, S. 145.
51 Vgl. z. B. Chen und Goodman 1998, S. 8.
52 Die *Accuracy* der *NLLR-/KLD*-Datierung sinkt von über 60 auf 21.8, während sie bei Verwendung von *CS* gleich mittelmäßig bei 38,9 bzw. 44,4 mit tf_{cmax} – idf bleibt.
53 Christian Buck und Philipp Koehn 2016: „Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance“. In: *Proceedings of the First Conference on Machine Translation, Berlin, Germany, August 11-12, 2016*. Bd. 2. Berlin: Association for Computational Linguistics, S. 672–678, S. 674, übersetzt durch den Verfasser.

— 5. ***Cosine similarity (CS)* und *tf-idf*.**[54] Zur Gewichtung der *tf* für die Berechnung der *CS* liefert bei der Verwendung von *DFZ-chronon*-Sprachmodellen der Vergleich der Vektoren mit *idf*-gewichteten *tf* des *chronon* mit denen der ungewichteten *tf* des zu datierenden Dokuments eine höhere *Accuracy* und einen niedrigeren *MAE*.[55] Für Modelle auf Dokumentebene ist dies nicht der Fall. Hier ist es erforderlich, die Worthäufigkeiten des zu datierenden Dokuments gleichermaßen zu gewichten.[56]

— 6. **Temporale Entropie (*TE*).** Für die Berechnung der *TE* werden hier die relativen Häufigkeiten *tf*, also $P(w, C) = \frac{f(w,c)}{|C|}$ usw. und nicht die absoluten Vorkommen verwendet. Die so berechneten Gewichte liegen zwar nicht zwischen 0 und 1, die Ergebnisse sind aber deutlich besser. Dies hängt ebenfalls mit der stärkeren Überbewertung von *Hapaxen* zusammen, die mit absoluten Häufigkeiten immer das maximale Gewicht von 1 erhalten.

— 7. ***NLLR* vs. *KLD*.** Ein nennenswerter Unterschied in der Performance ist nicht feststellbar. Kraaij bemerkt: „for ad hoc search, KL(Q|D) is essentially equivalent to the length normalized query likelihood [...] since the query entropy [...] is a constant which does not influence document ranking."[57]

6.1.2 Co-Datierung von Dokumenten

Wie unter anderem von de Jong, Rode und Hiemstra (2005) und Bamman et al. (2017) vorgeschlagen,[58] kann alternativ zur Verwendung aggregierter *chronons* dem zu datierenden Dokument der Zeitstempel des ähnlichsten Dokuments aus den Trainingsdaten zugewiesen werden.

Hierfür wird ein neues Korpus-*SLM* aus den einzelnen Trainingsdokumenten des *DFZ*-Datensatzes erzeugt. Aus Gründen der Vergleichbarkeit werden hierfür dieselben Texte aus dem Zeitraum 1475–1925 verwendet, wie für die Erzeugung der *chronon* Sprachmodelle in Abschnitt 6.1.1.[59] Das Modell enthält die relativen Häufigkeiten der 1–2 Zeichen-Lexeme und temporalen Ausdrücke von 772 Einzeltexten aus dem Zeitraum 1475–1925 (ursprünglich 17 *chronons*).

Anstatt wie oben den Zeitstempel des ähnlichsten *chronons* zu vergeben, wobei mit *NLLR* eine *Accuracy* von 60,6 % bei einem *MAE* von 40,3 Jahren erreicht wurde, wird hier der Zeitstempel des ähnlichsten Dokuments vergeben und zur besseren Vergleichbarkeit der Zeitraum des entsprechenden *chronons* genutzt. Die Datierung der 216 Texte des Testdatensatzes mit Jaccard *similarity*, *CS* mit und ohne *idf*-Gewichtung und *NLLR* mit einem *LCM-smoothing* mit $\lambda = 0,5$ wird mit den entsprechenden Ergebnissen der bereits durchgeführten Datierung mit

54 Siehe dazu auch Kapitel 3.3, v. a. S. 51 u. S. 53.

55 Bei Betrachtung von 1–2 Gramm Lexemen wird so eine *Accuracy* von 47,7 % und ein *MAE* von 51,9 Jahren erzielt (S. 161). Werden die *term frequencies* des zu datierenden Dokuments ebenfalls *idf*-gewichtet, reduziert sich die *Accuracy* auf 45,4 %, bei einem *MAE* von 58,4 Jahren. Beide Varianten liefern bessere Ergebnisse als die ungewichtete *CS*.

56 Siehe Abschnitt 6.1.2, S. 170. Bei Betrachtung von 1–2-Gramm Lexemen wird so eine *Accuracy* von 42,1 bei einem *MAE* von 60,9 Jahren erzielt. Die Verwendung von *CS* als Vergleich der ungewichteten *tf* des zu datierenden Dokuments mit den gewichteten *tf* der Vergleichsdokumente liefert eine *Accuracy* von 3,2 %, bei einem *MAE* von 224 Jahren.

57 Kraaij 2004, S. 208.

58 Siehe Kapitel 3.3, S. 50; siehe auch de Jong, Rode und Hiemstra 2005, S. 7; Bamman et al. 2017, S. 4.

59 Siehe S. 158.

chronon-Sprachmodellen verglichen (Tabelle 6.3).[60] Die Laufzeit der Berechnung beträgt etwa das 45-fache derjenigen bei Verwendung aggregierter *chronons*.

Tabelle 6.3 Ergebnisse mit *DFZ* Co-Datierung vs. *chronon-SLM*

				beschränkt auf		◆ *NLLR*		◆ *CS*		◆ *CS*tf-idf*	
	Modell	*#types*	*Smoothing*	*DHYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 / *chronons*	238.978	$\lambda = 0{,}5$	x	x	**60,6**	**40,33**	38,9	59,94	44,4	52,21
2	1–2 / *documents*	238.978	$\lambda = 0{,}5$	x	x	46,3	59,29	40,7	62,2	42,1	60,93

Tabelle 6.3 (Fortsetzung)

				beschränkt auf		◆ *Jaccard*		◆ *Random*	
	Modell	*#types*	*Smoothing*	*DHYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*
1	1–2 / *chronons*	238.978	$\lambda = 0{,}5$	x	x	13	117,97	10,6	142,32
2	1–2 / *documents*	238.978	$\lambda = 0{,}5$	x	x	39,4	66,78	15,7	139,43

Auch bei Verwendung eines dokumentenbasierten Modells lässt sich mit *NLLR* die höchste *Accuracy* (46,3 %) und der niedrigste *MAE* (59,3 Jahre) erzielen. Die Jaccard *similarity* gewinnt an Aussagekraft, übertrifft die Performance der komplexeren Metriken, die auch Worthäufigkeiten berücksichtigen, jedoch nicht. Insgesamt bleiben die Ergebnisse jedoch deutlich hinter denen der *chronon*-Datierung zurück.

Für die Entscheidung zwischen aggregierten *chronon*-Modellen und einem Direktvergleich von Dokumenten können neben der gewünschten Granularität der zu vergebenden Zeitstempel auch praktische Erwägungen eine Rolle spielen: die Verfügbarkeit geeigneter Trainingsdaten, sowie die Art der zu datierenden Texte.

Eine Wiederholung des Experiments mit dem *XXSKQS*-Datensatz mit 1–3-Grammen schriftsprachlicher chinesischer Texte,[61] deutet an, dass mit Co-Datierung unter Umständen ähnlich gute Ergebnisse erzielt werden können wie mit *chronons*. Im Fall des *XXSKQS* hängt dies sicherlich mit der heterogenen Natur des Korpus zusammen.[62]

Aus dem Datensatz wird je ein *chronon*-Sprachmodell und ein Co-Datierungs-Sprachmodell mit 1–2 Zeichen-Lexemen, Namen und Zeitausdrücken erzeugt. Für beide Modelle werden dieselben 717 Texte als Trainingsdaten verwendet und ein Testdatensatz mit 176 Texten zufällig ausgewählt.[63] Wie bereits in Abschnitt 6.1.1 werden Texte aus dem Zeitraum von 1475–1925 berücksichtigt. Um Defizite in den zur Verfügung stehenden Metadaten auszugleichen, werden bei der Auswahl der Trainings- und Testdaten nur Texte ausgewählt, bei denen die Namen der Verfasser:innen (mit *zhuan* 撰, „zusammenstellen, verfassen, komponieren“) in den Metadaten

60 Die Berechnung von *TE* zur Gewichtung der *features* für die Berechnung von *NLLR* oder *KLD* erscheint bei der Betrachtung einzelner Dokumente wenig intuitiv. Denkbar wäre aber ein hybrider Ansatz, der auf die *TE* der entsprechenden *chronons* zurückgreift.

61 *XXSKQS*, siehe auch Kapitel 4.2, S. 68, siehe auch Abschnitt 6.1.3, v. a. S. 174.

62 Siehe Kapitel 4.2, S. 68.

63 Es wurde eine ausgewogene Aufteilung der Testdaten auf die *chronons* angestrebt. Für die Zeiträume 1475–1525, sowie 1675–1725 stehen aber wg. der Auswahlkriterien und Trainingsdaten nur noch einzelne Texte zur Verfügung. Daher muss der Testdatensatz kleiner ausfallen als bei den *DFZ*.

angegeben sind und die Dynastie der Veröffentlichung zu deren biographischen Daten passt.[64] Tabelle 6.4 zeigt die Ergebnisse bei der Datierung der 176 Texte aus den Testdaten mit *chronon*- und Co-Datierungs-Modellen im direkten Vergleich. Es wurde jeweils ein *LCM Smoothing* mit $\lambda = 0,5$ angewendet und *DHYDCD*-Lexeme, Namen und Zeitausdrücke berücksichtigt.

Tabelle 6.4 Ergebnisse mit *XXSKQS* Co-Datierung vs. *chronon-SLM*

			◆ *NLLR*		◆ *CS*		◆ *CS*tf-idf*		◆ *Jaccard*		◆ *Random*	
	Modell	*#types*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 / *chronons*	267.349	23.3	98.11	19,9	108,52	**33,5**	**80,76**	11,9	103,81	11,9	141,54
2	1–2 / *documents*	267.349	21.6	101.32	25	94	27,3	94,26	**29,5**	**86,17**	11,4	142,84

Insgesamt bleiben die Ergebnisse erwartungsgemäß weit hinter denen der *DFZ*-Experimente zurück, da die *XXSQKS* als diachrones Korpus deutlich problematischer sind.[65] Erneut lassen sich mit einem aggregierten *chronon*-Modell bessere Ergebnisse erzielen. Die Diskrepanz zwischen *chronon*- und dokumentenbasiertem Modell ist aber deutlich kleiner.

Die höchste *Accuracy* (33,5 %) und der niedrigste *MAE* (80,8 Jahre) werden mit ersterem bei Verwendung von *CS* mit *tf-idf*-Gewichtung erreicht. Ein vergleichsweise niedriger *MAE* (86,2) kann jedoch auch mit Co-Datierung und Jaccard *similarity* erreicht werden. Die Zuordnung erfolgt dabei unabhängig von Worthäufigkeiten allein aufgrund einer Überschneidung der vorkommenden Lexeme. Die Erfolgschancen einer dokumentenbasierten Datierung dürften stärker als bei der *chronon*-Datierung davon abhängen, dass dem zu datierenden Dokument inhaltlich und damit im Wortschatz ähnliche Texte in den Trainingsdaten vorhanden sind. Insgesamt sollte es damit wenig Anwendungsfälle geben, in denen eine dokumentenbasierte Vorgehensweise vorzuziehen ist, die überdies deutlich höhere Laufzeiten mit sich bringt.

6.1.3 Datierung mit *DHYDCD*-Sprachmodellen

Mit aus dem *DFZ*-Korpus generierten Sprachmodellen lassen sich mit *NLLR* und *TE*, einer vereinfachten *Smoothing*-Methode und der Einschränkung auf Lexeme und temporale Ausdrücke bereits eine *Accuracy* von über 60 % bei einem *MAE* von etwa 40 Jahren erreichen. Die Datierung ist dabei jedoch durch die Trainingsdaten auf den Zeitraum vom 15. bis zum Anfang des 20. Jahrhunderts beschränkt. Auch für die Datierung von Texten außerhalb des vorgegebenen Genres sind die Erfolgsaussichten als gering einzustufen. Um die ungefähre zeitliche Einordnung von Texten für die gesamte chinesische Schrifttradition und über Genregrenzen hinweg zu ermöglichen, verwende ich als Trainingskorpus die datierbaren Textzitate aus dem *DHYDCD*.[66] Aus den 53 Segmenten dieses Behelfskorpus mit repräsentativem Textmaterial unterschiedlicher Genres aus dem Zeitraum von ca. 700 v. u. Z. bis zum 20. Jh. können temporale Sprachmodelle mit einer Auflösung von 100 Jahren und einer Überlappung von 50 Jahren berechnet werden.[67] 700–

64 Das *XXSKQS* enthält zahlreiche spätere Ausgaben, bei denen der Name des Herausgebers bzw. eines Kommentators angegeben ist. Solche Werke aufgrund ihrer *n*-Gramm-Häufigkeiten zu datieren ist aussichtslos, da Textmaterial aus unterschiedlichen Zeiträumen in unterschiedlichen Anteilen untrennbar vermischt ist. Texte mit fragwürdigen Angaben (siehe dazu auch Kapitel 4.2, ab S. 67.) werden ebenfalls ausgeschlossen.

65 Datierungsergebnisse mit dem *XXSQKS*-Korpus werden in Abschnitt 6.1.3, ab S. 175, ausführlicher diskutiert.

66 Siehe Kapitel 5.6, ab S. 137.

67 Für eine kürzere *chronon*-Dauer reicht die Genauigkeit der Datierung der zugrunde liegenden Daten leider nicht aus. Vgl. auch Kapitel 5.7, S. 139.

600 v. u. Z. ist also das erste, 650–550 v. u. Z. das zweite usw. und 1900–2000 das letzte *chronon* solcher Modelle.

Die aus den in Abschnitt 6.1.1 (ab S. 158) beschriebenen Experimenten als am effektivsten hervorgegangenen Herangehensweisen werden mit diesen *DHYDCD*-Sprachmodellen erprobt. Hierfür werden — 1. Die 25 offiziellen Dynastiegeschichten[68] gegen die Textbeispiele aus dem *DHYDCD* datiert. — 2. dieselben zufälligen 216 Texte aus dem *difangzhi* 地方誌-Korpus datiert. — 3. Da keine Trainingsdaten aus dem *Difangzhi*-Datensatz benötigt werden, kann dieser über einen größeren Zeitraum von 1300–1925 genutzt werden. — 4. Um eine Datierung von Texten mit dem *DHYDCD*-Sprachmodell über Genregrenzen hinweg zu testen, werden überdies zufällige Texte aus dem *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書 datiert.

Als *Baseline* dient erneut ein Zufallsgenerator, der – bei 53 überlappenden *chronons* – Texte mit einer Wahrscheinlichkeit von etwa 4 % korrekt zuordnet. Eine Co-Datierung von Texten wie in 6.1.2 ist so weder sinnvoll noch möglich, da die aggregierten Trainingsdaten keine konkreten Texte mehr repäsentieren. Ebensowenig erfolgt eine Unterteilung des Korpus in Test- und Trainingsdaten.[69]

Experiment 1: *zhengshi* 正史

Im ersten Test der aus dem Zitatmaterial des *DHYDCD* erzeugten *SLM*s werden die 25 offiziellen Dynastiegeschichten zugeordnet. Die Sprachmodelle können so mit einem Referenzkorpus getestet werden, dessen Texte aus einem sehr großen Zeitraum von 2.019 Jahren, von ca. 91 v. u. Z. bis 1928, stammen. Dabei darf nicht vergessen werden, dass ein großer Teil der *zhengshi* als Belegstellen im *DHYDCD* häufig bis sehr häufig vertreten ist[70] und dadurch bedingt eine hohe Übereinstimmung an *types* zwischen den zu datierenden Texten und den jeweils korrekten *chronons* besteht. Die Testreihe ist also teilweise inzestuös. Sie kann aber Aufschluss über die Eignung der unterschiedlichen Metriken und die geeignete Größe des *Smoothing*-Parameters λ für das *LCM Smoothing* geben. Da die *zhengshi* als Volltext vorliegen, besteht keine Beschränkung des *n*-Gramm-Raums auf eine Länge von 3 Zeichen mehr, so dass auch die Verwendung von Lexemen von 1–4 Zeichen Länge geprüft werden kann.[71]

Tabelle 6.5 Ergebnisse mit *zhengshi* und mit *DHYDCD-SLM*s

				beschränkt auf		◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Jaccard*	
	Modell	*#types*	λ	*HYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 *grams*	1.992.349	0,90	-	-	84	**88,68**	88	65,68	68	67,28	64	95,64
2	1–2 *grams*	1.992.349	0,50	-	-	80	77,04	80	**59,16**	68	67,28	64	95,64
3	1–2 *grams*	1.992.349	0,10	-	-	68	72,64	76	59,52	68	67,28	64	95,64
4	1–2 *grams*	1.992.349	0,01	-	-	60	87,28	76	59,52	68	67,28	64	95,64
5	1–2	228.191	0,90	x	-	84	88,68	80	72,4	60	77,96	64	90,48
6	1–2	228.191	0,50	x	-	80	87,04	76	62,76	60	77,96	64	90,48
7	1–2	228.191	0,10	x	-	80	78,24	76	60,76	60	77,96	64	90,48
8	1–2	228.191	0,01	x	-	68	73,84	72	62,32	60	77,96	64	90,48
9	1–2	228.438	0,90	x	x	84	86,68	76	73,96	60	77,96	64	90,48

68 Siehe dazu Kapitel 2.3, ab S. 20.

69 Theoretisch möglich wäre die Erzeugung von Pseudotext aus den Belegstellen jedes einzelnen zitierten Texts – was allerdings in den wenigsten Fällen eine ausreichende Menge an Sprachmaterial ergäbe.

70 Siehe dazu auch Kapitel 5.7.4, ab S. 150

71 Aufgrund ihrer geringen Zahl ist die Berücksichtigung von Wörtern mit einer Länge von 5+ Zeichen wenig aussichtsreich. Siehe dazu Kapitel 5.7.3, ab S. 146 und Kapitel 4.5.2, S. 92.

Tabelle 6.5 (Fortsetzung)

				beschränkt auf		◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Jaccard*	
	Modell	*#types*	λ	*HYDCD*	+ Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
10	1–2	228.438	0,50	x	x	80	87,04	76	62,76	60	77,96	64	90,48
11	1–2	228.438	0,01	x	x	68	73,84	72	61,12	60	77,96	60	97,8
12	1–3	240.601	0,90	x	x	84	86,68	76	73,96	60	77,96	68	82,68
13	1–4	256.800	0,90	x	x	84	86,68	76	73,96	60	77,96	68	82,68

1. Die beste *Accuracy* von 88 wird bei Verwendung aller 1–2-Gramme mit *NLLR*TE* und einem *Smoothing*-Parameter von $\lambda = 0,9$ erreicht, der geringste *MAE* von 59,5 Jahren mit einem niedrigeren *Smoothing*-Faktor – auf Kosten einer gesunkenen *Accuracy*.
2. Auch wenn aus den Ergebnissen für die nur 25 hier datierten Texte keine voreiligen Schlüsse gezogen werden sollten, scheint ein hoher Wert von λ eine tendenziell bessere *Accuracy* und ein niedriger Wert einen geringeren *MAE* zu ermöglichen. Mit $\lambda = 0,9$ werden also mehr Texte korrekt datiert, mit $\lambda = 0,01$ weniger starke Abweichungen erzielt.
3. Durch Eingrenzung der betrachteten *n*-Gramme auf Lexeme ergibt sich erwartungsgemäß eine drastische Steigerung der Verarbeitungsgeschwindigkeit, da nur etwas mehr als 10 % der *types* genutzt werden. Die Ergebnisse der Datierung verschlechtern sich dabei marginal.
4. Eine Erweiterung der Lexeme um Zeitausdrücke hat hier kaum Auswirkungen, da die Trainingsdaten nur sehr wenige Vorkommen aufweisen.
5. Eine Vergrößerung des *n*-Gramm-Raums auf Lexeme mit 1–3 bzw. 1–4 Zeichen Länge bringt ebenfalls keine deutliche Verbesserung mit sich, da Wörter mit einer Länge von mehr als 2 Zeichen einen zu geringen Anteil haben. Aufgrund der großen Anzahl an *zhengshi*-Zitaten im Korpus wird bei der Verwendung von Jaccard *similarity* ein geringer Effekt sichtbar.
6. Wieder ist *NLLR* den einfacheren Ähnlichkeitsmaßen überlegen, durch *TE*-Gewichtung wird eine etwas geringere *Accuracy* bei besserem *MAE* erzielt.

Experiment 2: 216 *Difangzhi* 地方誌
Im zweiten Versuch werden die für Abschnitt 6.1.1 zufällig für die Datierung mit *DFZ-SLM*s ausgewählten Texte erneut unter Verwendung der *DHYDCD SLM*s datiert. Es bestätigt sich, dass mit einem weniger spezialisierten Trainingskorpus bzw. einem längeren Betrachtungszeitraum zunächst deutlich schlechtere Ergebnisse erzielt werden können, als dies in Abschnitt 6.1.1 bzw. in Experiment 1 der Fall war. Die beste erreichte *Accuracy* von 16,2 % liegt deutlich unter der Performance bei Verwendung der *DFZ* selbst zur Erzeugung des temporalen Sprachmodells, ebenso wie der *MAE* von 140,6 Jahren.[72] Es sollte jedoch bedacht werden, dass nun längere *chronons* und ein sechsmal längerer Betrachtungszeitraum von 2.700 Jahren verwendet werden. In dieser Relation ist eine Datierung mit einer durchschnittlichen Genauigkeit von 140 Jahren als durchaus aussagekräftig anzusehen. Die genauen Ergebnisse von Experiment 2 sind in Tabelle 6.6 wiedergegeben.[73]

72 Dieselben Texte konnten in Abschnitt 6.1.1 zu 64 % korrekt bei einem *MAE* von etwa 40 Jahren datiert werden. Siehe S. 162.

73 Um einen möglichst geringen *MAE* zu erzielen, wurde für *unseen events* hier ein *LCM smoothing* mit $\lambda = 0,01$ angewandt.

Tabelle 6.6 Ergebnisse mit *Difangzhi* und *HYDCD-SLM*s.

				◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Random*	
	Modell	*#types*	Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 Gramme	1.992.349	-	16,2	152,92	15,7	141,92	**19,9**	**136,04**	5,1	1.072,06
2	1–2 字 Lexeme	228.191	-	15,3	165,97	16,2	141,45	6	322,31	4,6	1.134,55
3	1–3 字 Lexeme	240.601	x	16,2	158,53	16,2	140,79	17,1	202,69	4,6	1.058,95
4	1–2 字 Lexeme	228.438	x	15,7	157,45	**16,2**	**140,56**	17,1	210,7	4,2	1.165,05

Zudem deutet sich an, dass bei kargerer Datenlage *CS* mit *tf-idf* bessere Ergebnisse liefern kann als die komplexeren Metriken, allerdings nur bei Verwendung aller 1–2-Gramme – bei Reduktion der Dimensionen auf Lexeme und Zeitausdrücke sind die *Accuracy*-Unterschiede zur *NLLR* marginal. Zudem wird bei Verwendung von *NLLR* ein geringerer *MAE* erzielt, der durch Gewichtung mittels *TE* auf 140,56 Jahre reduziert werden kann.

Experiment 3: *Difangzhi*, 1300–1925

Der *DFZ*-Datensatz enthält Texte aus dem Zeitraum vom 8. bis zum Anfang des 20. Jhs. Erst ab dem 15. Jh. sind jedoch ausreichend Texte im Korpus, um temporale *SLM*s zu erzeugen. Als reine Testdaten können jedoch auch ältere Texte ab etwa 1300 eingesetzt werden. In Experiment 2 soll festgestellt werden, ob sich eine neue, zufällige Textauswahl aus dem längeren Zeitraum 1300–1925 ebenso gut datieren lassen wie die Texte aus Experiment 2 bzw. Abschnitt 6.1.1. Hierfür werden 1–2 Zeichen-Lexeme und temporale Ausdrücke verwendet und 119 Texte zufällig zur Datierung ausgewählt. Das Experiment wird viermal mit identischen Parametern wiederholt, so dass jeweils teilweise unterschiedliche Texte datiert werden.

Tabelle 6.7 Ergebnisse mit *Difangzhi* 1300–1925 und *HYDCD-SLM*s.

	◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Random*	
	Accuracy (%)	*MAE* (Jahre)	*A* (%)	*MAE* (Jahre)	*A* (%)	*MAE* (Jahre)	*A* (%)	*MAE* (Jahre)
1	14,3	162,98	20,2	133,9	21	214,95	4,2	1.071,46
2	14,3	167,12	17,6	135,76	22,7	218,22	5,9	1.044,2
3	10,1	174,71	16,8	137,5	18,5	227,52	4,2	1.002,95
4	9,2	163,97	15,1	136,32	15,1	198,79	4,2	990,39

Die Ergebnisse zeigen ein sehr ähnliches Bild wie in Experiment 2, die Wiederholungen geben zudem Aufschluss über die Schwankungsbreite bzw. Stabilität der Datierungsergebnisse. *NLLR* liefert einen konstant niedrigen *MAE*, der sich durch *TE*-Gewichtung auf ca. 135 Jahre reduzieren lässt. Mit *CS* und *tf-idf* lassen sich erneut tendenziell etwas mehr Texte einem korrekten *chronon* zuordnen, der *MAE* bleibt jedoch deutlich schlechter, als das mit *NLLR* der Fall ist. Die Erweiterung des Datierungszeitraums wirkt sich nicht nachteilhaft aus.

Experiment 4: *Xu xiu si ku quan shu* 續修四庫全書, 14. Jh.–1927

In Experiment 4 soll festgestellt werden, ob eine ungefähre Datierung mit den *DHYDCD-SLM*s auch über Genregrenzen hinweg möglich ist. Aus dem *N-gram dataset of Xu xiu si ku quan shu* 續修四

庫全書 wird ein Testdatensatz von 105 Texten aus dem Zeitraum vom 14. Jh. bis 1927 gleichmäßig verteilt zufällig ausgewählt.[74]

Die Texte werden mit *NLLR*, *NLLR*TE*, *CS*tf-idf* und einem Zufallsgenerator als Baseline datiert, für *NLLR* wird ein *LCM Smoothing* mit $\lambda = 0,01$ verwendet. Die Ergebnisse sind in Tabelle 6.8 aufgelistet.

Tabelle 6.8 Ergebnisse mit *Xu xiu si ku quan shu* und *HYDCD-SLM*s.

			◆ *NLLR*		◆ *NLLR * TE*		◆ *CS * tf-idf*		◆ *Random*	
	Modell	Zeit	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*	*A*	*MAE*
1	1–2 Gramme	-	21,9	323,79	21,9	298,92	21	346,67	3,8	1.135,09
2	1–2 字 Lexeme	-	21	358,49	23,8	284,37	13,3	390,55	5,7	1.140,45
3	1–2 字 Lexeme	x	21	354,81	23,8	284,37	13,3	386,74	3,8	1.009,33
4	1–3 字 Lexeme	x	21	354,81	22,9	286,45	13,3	384,36	2,9	1.104,76

Mit einer *Accuracy* von 23,8 % bei Verwendung von *NLLR* mit *TE*-Gewichtung ist hier sogar ein etwas größerer Anteil an Texten korrekt datiert als in den Experimenten 2 und 3 mit Texten aus dem *DFZ*-Datensatz. Der *MAE* ist allerdings mit fast 300 Jahren deutlich größer. Weder die Erweiterung des *n*-Gramm-Raums auf Lexeme mit bis zu drei Zeichen, noch die Berücksichtigung von Zeitausdrücken verursachen einen spürbaren Effekt. Werfen wir einen Blick auf die Detaildarstellung der 105 mit *NLLR*TE* datierten Texte, lassen sich schnell zwei wesentliche Probleme erkennen (Abb. 6.5).

— 1. Texte werden übermäßig häufig auf das *chronon* 1550–1650 datiert. Diese Präferenz ist kein Zufall, da für dieses *chronon* Daten zu mehr *types* zur Verfügung stehen, als dies bei den benachbarten *chronons* der Fall ist. Durch Verwendung „ausgewogener" Modelle kann diese Problematik zwar aufgelöst werden, der Datenverlust führt aber insgesamt zu einer deutlichen Verschlechterung der Ergebnisse.[75] — 2. Die Titel einiger stark zu früh datierter Texte (Abb. 6.5), z. B. *Sanguo zhi zhu bu* 三國志注補, (etwa: „Ergänzung zur kommentierten Ausgabe der Chroniken der Drei Reiche", 1644), *Shiji kaozheng* 史記考證 (etwa: „Untersuchungen über die Authentizität des *Shiji*", 1788), *Chuci yi* 楚詞譯 (etwa: „Interpretation der Elegien von Chu", 1901) usw., deuten darauf hin, dass es sich dabei um neue, kommentierte Ausgaben der jeweils im Titel genannten Texte (*Sanguo zhi*, Ende des 3. Jahrhunderts, *Shiji*, 1. Jh. v. u. Z., *Chuci*, ca. 3. Jh. v. u. Z. usw.) handelt, oder zumindest um Werke, die diese Texte in hohem Maße zitieren und damit einen sehr hohen Anteil an früherem Sprachmaterial enthalten. Texte wie das lt. Metadaten 1409 entstandene *Sheng xue xin fa* 聖學心法, datiert auf das *chronon* 400–300 v. u. Z, zeigen aber, dass auch stilistisch „alte" Texte für die linguistische Datierung ein großes Problem darstellen, vor allem, wenn sie kaum oder kein zeitgenössisches Vokabular enthalten.

In Experiment 4 werden zwar nur rund ein Viertel der Texte korrekt datiert und der *MAE* ist mit 284 Jahren sehr hoch, dennoch datiert ein Großteil der Texte ungefähr in den korrekten Zeitraum. Extremwerte falscher Datierungen sind teilweise auf die Natur des *XXSKQS* mit einem hohen Anteil an Texten, die frühere Texte ganz oder teilweise enthalten – z. B. kommentierte Ausgaben – zurückzuführen. Diese Art der Intertextualität, die bei traditionellen Herangehensweisen für die Datierung eines Textes hilfreich sein kann, denn „if text A cited text B, then text

74 Da hier ein größerer Zeitraum betrachtet werden kann als in Abschnitt 6.1.2 (ab S. 169), wird ein neuer Testdatensatz generiert.

75 Ohne Abb. Mit *NLLR * TE* wird noch eine *Accuracy* von $A = 11,4$ bei einem *MAE* von 433 Jahren erreicht. Detaillierte Versuche zur Verwendung gleichförmiger *chronons* in Abschnitt 6.1.1, S. 161 zeigen sehr ähnliche Verschlechterungen.

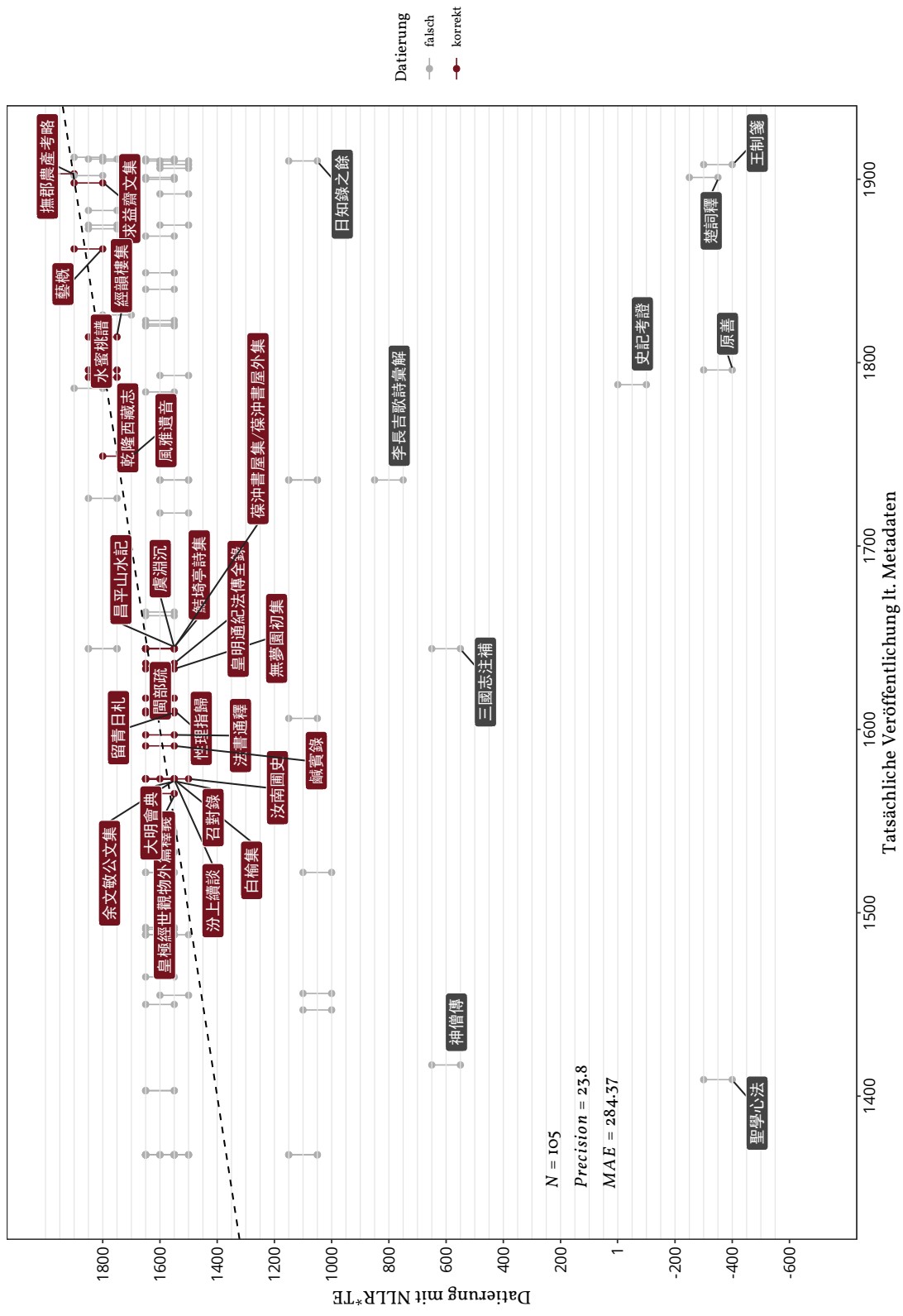


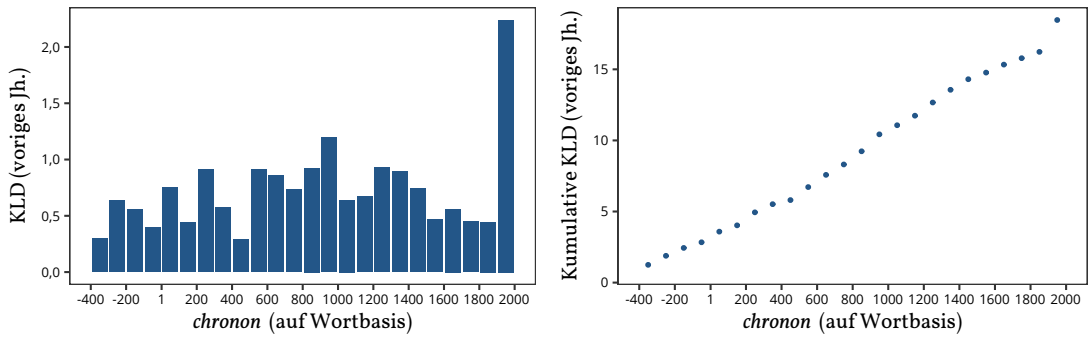
Abbildung 6.5 Experiment 4.4: 105 zufällige Texte aus dem *XXSKQS*

B must be older than text A"[76] ist für eine quantitative bzw. statistische Betrachtung, bei der Kommentar und Haupttext nicht unterschieden werden können, problematisch. Das Beispiel des Qing-zeitlichen *Yuan shan* 原善 (1796), hier dem *chronon* 300–200 v. u. Z. zugeordnet, zeigt, dass einige Texte resistent gegenüber einer statistischen Analyse sein dürften.[77]

Dass es bei Verwendung ausgewogener Modelle zu einer spürbaren Verschlechterung der *MAE*-Performance kommt, deutet auf einen für diese Zwecke wohl zu geringen Umfang der *DHYDCD*-Trainingsdaten hin. Die Datierungen sind dabei nicht abwegig, sondern schwanken um einen Zeitraum von mehreren hundert Jahren um die tatsächliche Entstehung der jeweiligen Texte. Diese Ungenauigkeit spiegelt die Problematik eines nur langsamen Sprachwandels des schriftsprachlichen Stils für Datierungsversuche auf Basis statistischer Sprachmodelle wider.

6.1.4 Sprachwandel im Sprachmodell

Wie die Experimente in 6.1.1 bis 6.1.3 gezeigt haben, können mit *SLM*s schriftsprachliche chinesische Texte zwar nicht exakt datiert werdem, mit *KLD* oder *NLLR* aber – abhängig von untersuchtem Material und Trainingsdaten – ungefähr zeitlich eingestuft werden. Durch Betrachtung des Unterschieds der einzelnen *chronon* Sprachmodelle voneinander, kann mit derselben Methodik auch die Veränderung des Wortgebrauchs als Aspekt des Sprachwandels quantifiziert werden.[78] Die Beobachtungen sollten dabei allerdings nur als grobe Richtung bzw. linguistische Trends betrachtet werden, da die Belegstellen aus dem *DHYDCD* ein stark begrenztes Textmaterial darstellen, das zudem ein *Bias* mit Präferenzen für bestimmte Texte und Textgattungen aufweist.[79]



(a) KLD, -100 Jahre

(b) KLD (kumulativ), -100 Jahre

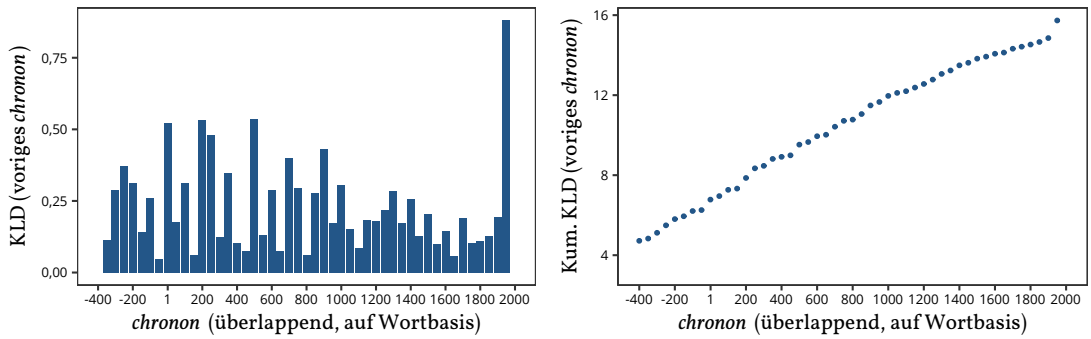
Abbildung 6.6 Kullback-Leibler-Divergenz zum *chronon* des jeweils vorangegangenen Jahrhunderts

Abbildung 6.6a stellt die Veränderung eines jeden *chronon* zum Modell des jeweils vorangegangenen Jahrhunderts dar, d. h. es wird z. B. die *KLD* des *chronon* 400–500 zum *chronon* 300–400 gemessen.[80] Eine außergewöhnlich starke Veränderung ist dabei vom 19. hin zum 20. Jh. sicht-

76 Toner und Han Xiwu 2019, S. 17–18. Siehe auch Kapitel 3, S. 39.
77 Der Text wird in Abschnitt 6.2.5 ausführlicher diskutiert, siehe S. 209.
78 Vgl. auch de Jong, Rode und Hiemstra 2005, S. 1.
79 Siehe dazu Kapitel 5.7, ab S. 138.
80 Die Kullback-Leibler-Divergenz bietet sich hier als sinnvolleres Maß an, da sie – im Gegensatz zur *NLLR* – den Unterschied zwischen zwei Sprachmodellen misst.

bar, in der sich wahrscheinlich unter anderem die Auswirkungen der Bewegung des 4. Mai (*wusi yundong* 五四運動, ab 1919) mit einer stärkeren Verschriftlichung der Umgangssprache (*baihua* 白話) widerspiegeln. Die kumulative Darstellung (Abb. 6.6b) lässt zudem eine langfristig weitgehend lineare Veränderung erahnen, die etwa ab dem 13. Jh. leicht abzuflachen scheint.

Werfen wir noch einen Blick auf die Veränderung ohne Auslassung der überlappenden *chronons* (Abb. 6.7). Dies suggeriert in der kumulativen Darstellung ebenfalls eine abflachende Kurve, innerhalb derer allerdings eine gewisse Zyklik erkennbar wird.[81] Zudem sind kleinere *s*-Formen zu sehen, die an das Piotrowski-Gesetz erinnern[82] – für eine belastbare Interpretation in diese Richtung wäre allerdings umfassenderes Datenmaterial erforderlich.



(a) KLD, -50 Jahre

(b) KLD (kumulativ), -50 Jahre

Abbildung 6.7 Kullback-Leibler-Divergenz zum vorigen *chronon*

Mit einer umgekehrten Jaccard *similarity* kann zudem die reine Übereinstimmung der in den Textbeispielen verwendeten *types* zum jeweils vorangegangenen Jahrhundert unabhängig von ihrer Häufigkeit gemessen werden.[83]

Auch in Abb. 6.8 bleibt eine stärkere Veränderung des verwendeten Wortschatzes im 20. Jh. sichtbar – es werden also nicht nur die vorhandenen Lexeme stark unterschiedlich häufig verwendet, sondern auch ein größerer Anteil *anderer* Lexeme als zuvor. In der kumulativen Darstellung deutet sich ebenfalls wieder eine schwache *s*-Form an.

Mit Ausnahme eines großen Sprungs im 20. Jh. lassen die Beobachtungen aus dem Zitatmaterial des *DHYDCD* auf eine langfristig betrachtet verhältnismäßig konstante Veränderung des Wortschatzes schließen; die Veränderungsrate der Wortnutzung scheint dabei einer gewissen Schwankung zu unterliegen. Da die verwendeten Trainingsdaten sehr begrenzt sind und die Auswahl der zugrunde liegenden Texte starke Unausgewogenheiten aufweist,[84] muss aber von

81 Bereits der Sinologe Hans Georg Conon von der Gabelentz hatte ein Modell des zyklischen Sprachwandels bzw. „Spirallaufs der Sprachgeschichte" eingeführt, das natürlich nicht belegt ist. Ihm geht es dabei primär um einen syntaktischen Sprachwandel. Hans Georg Conon von der Gabelentz 1901 [1891]: *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Hrsg. von Albrecht Graf von der Schulenburg. 2. vermehrte und verbesserte Auflage. Leipzig: Tauchnitz, S. 255.

82 Siehe dazu Kapitel 2.1, S. 14 und v. a. Kapitel 5.7, S. 146

83 Die Jaccard *similarity J* ist hier als Schnittmengenanteil der *types* zweier Wortlisten definiert (siehe S. 52), gibt also den Grad ihrer Übereinstimmung an. Um die Abweichung vom *chronon* des vorangegangenen Jahrhunderts zu messen, wird 1 – *J* verwendet. Dieses Maß bezeichne ich im Folgenden als |Jaccard-*Divergenz*.

84 Siehe dazu auch Kapitel 5.7.4, ab S. 150.

allgemeingültigen Aussagen über eine Zyklik oder Geschwindigkeit des Sprachwandels im Chinesischen Abstand genommen werden.[85]

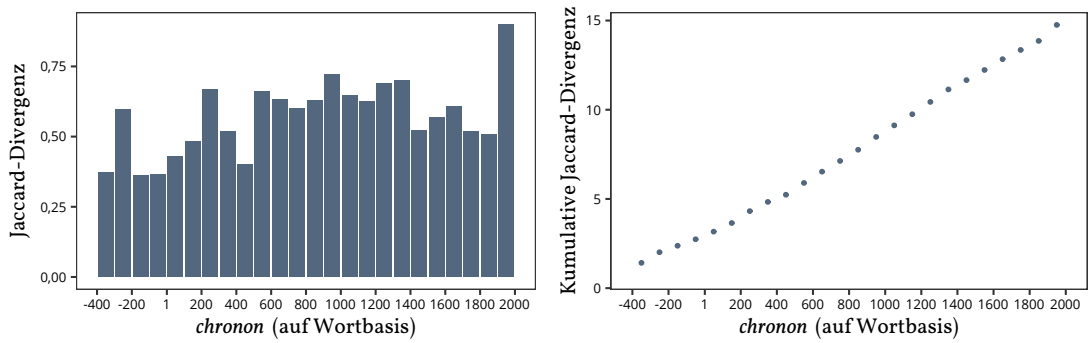


Abbildung 6.8 Jaccard-„Divergenz" zum *chronon* des vorigen Jahrhunderts

6.2 Datierung mit Neologismusprofilen

> *„[…] we foresee a role for parsed entries from historical dictionaries in this context as well."*[86]
>
> Franciska de Jong, Henning Rode and Djoerd Hiemstra

Wie in Kapitel 6.1 dargestellt, ist eine ungefähre Datierung schriftsprachlicher chinesischer Texte mit statistischen Sprachmodellen grundsätzlich möglich, wenn einige Voraussetzungen erfüllt sind: Das Genre muss bekannt sein und der Text sollte nicht in einem antiken Stil verfasst sein, der die Verwendung zeitgenössischer Lexeme und Satzkonstruktionen vermeidet. Um gute Ergebnisse zu erzielen, wird überdies ein passendes Trainingskorpus benötigt.

Die im Folgenden vorgestellte Datierungsmethodik basiert auf der aus dem *DHYDCD* erzeugten diachronen Lexemdatenbank[87] und ermöglicht eine ungefähre Altersschätzung von Texten, die genreunabhängiger ist und ohne spezifisches Trainingskorpus auskommt. Sie basiert auf der einfachen Idee, dass ein Text *mindestens* so neu ist, wie das neueste darin enthaltene Lexem (*Newest Word in Text*). Diese Herangehensweise stellt quasi eine Digitalisierung des *Lexical Dating* dar, „a method of establishing the chronology of a text through an examination of its vocabulary."[88] Würden wir den *Locus classicus jedes* Lexems bzw. jeder Kombination von Schriftzeichen und deren früheste Verwendung kennen, könnte durch einen Abgleich der *types* eines Texts mit einer entsprechenden Datenbank sein „neuestes Wort" ermittelt werden. Bei Vollständigkeit der verwendeten Daten ließe sich so implizit das *maximale* Alter des Textes ermitteln.

85 Vgl. aber Arapov und Cherc 1983 [1974], S. 88: „Die linguistische Erfahrung lehrt, daß einige Epochen der Sprachentwicklung durch stürmische Veränderungen gekennzeichnet sind, andere dagegen durch relative Stagnation. Diese Erfahrung beruht aber eher auf der Erforschung der historischen Phonetik und Morphologie und ist dann auf die schwer überschaubare Lexik übertragen worden."

86 de Jong, Rode und Hiemstra 2005, S. 1.

87 Siehe Kapitel 5.5, ab S. 120.

88 Toner und Han Xiwu 2019, S. 33–34.

In der Praxis ist der Wortschöpfungsprozess selten so genau dokumentiert. Mit den historischen Lexikalisierungsdaten aus dem *DHYDCD* stehen uns dennoch umfangreiche Daten zur Verfügung, die ernsthafte Experimente mit diesem Gedankenspiel zulassen. Einschränkungen, die sich aus der Unvollständigkeit dieser Daten ergeben, lassen sich durch Ergänzung früherer Belegstellen aus datierten Texten reduzieren.[89]

Die in einem zu datierenden Text erkannten Lexeme können auf dieser Datenbasis chronologisch zugeordnet und diese Zuordnung entsprechend visualisiert werden. Die für diese Methode verwendete Darstellung der Lexikalisierung bezeichne ich im Folgenden als *Neologismusprofil* bzw. als *temporales Profil* eines Textes. Durch die chronologische Einordnung der enthaltenen *types* können die Profile zum einen als philologisches Werkzeug genutzt werden, um Rückschlüsse über die stilistische, inhaltliche und temporale Einordnung des Textes zu ziehen. Mithilfe statistischer Überlegungen können sie zudem durch Software interpretiert werden, was auch einen groben Performancevergleich mit den Datierungsmethoden aus Kapitel 6.1 ermöglicht. Betrachten wir zur Veranschaulichung ein Neologismusprofil des *Meng xi bi tan* 夢溪筆談 (*MXBT*) von Shen Kuo 沈括 (1031–1095):[90]

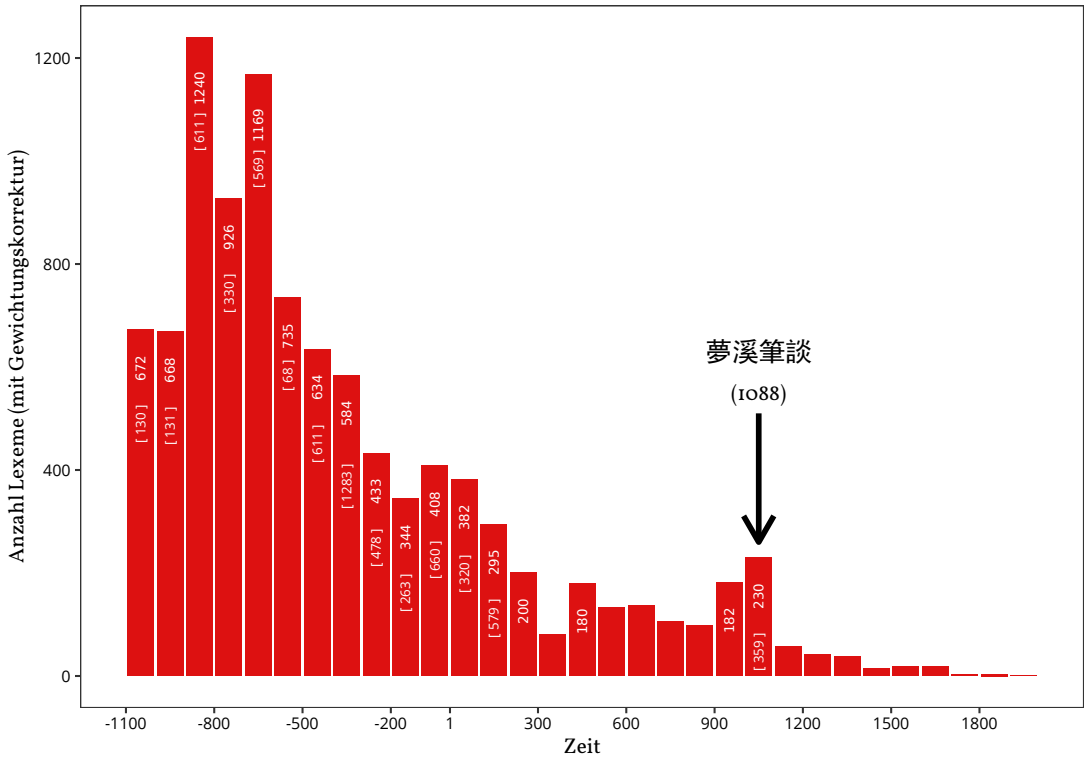


Abbildung 6.9 Neologismusprofil mit Gewichtungskorrektur für das *MXBT*, 2–4 Zeichen Lexeme

89 Siehe Kapitel 5.5.4, S. 134. Eine *vollständige* Lexikalisierungsgeschichte, die für eine naive *Newest Word in Text*-Datierung benötigt würde, könnte nur durch Betrachtung *sämtlicher* chinesischsprachiger Texte nachgezeichnet werden und bleibt damit eine theoretische Überlegung.

90 Shen Kuo 沈括 2008 [1088]: *Meng xi bi tan* 夢溪筆談 *(Pinselunterhaltungen am Traumbach)*. Project Gutenberg eBook. URL: http://www.gutenberg.net (besucht am 10. 09. 2018).

Das Balkendiagramm (Abb. 6.9) stellt die Anzahl der im *MXBT* enthaltenen Lexem-*types* pro Jahrhundert dar,[91] indem sie dem Jahrhundert ihrer frühesten bekannten Belegstelle zugeordnet werden.[92] Der Text enthält einen großen Anteil an Lexemen, die bereits sehr früh belegt sind. Zur Gegenwart hin nimmt die Anzahl der pro Jahrhundert nachgewiesenen Lexeme ab. Der Verlauf dieser Abnahme erinnert an die „arith-logarithmic equation" mit der George Zipf den Zusammenhang zwischen Häufigkeit und Alter von Wörtern herstellt.[93] Mikhail Arapov und Maja Cherc führen Zipfs Entdeckung aus:

> Es existiert ein Zusammenhang zwischen der Häufigkeit eines Wortes und der Zeit seiner Entstehung in der Sprache. Es zeigt sich, daß die Mehrheit der Wörter mit großer Auftretenshäufigkeit von den sehr alten Wörtern gebildet wird; umgekehrt ist die Chance dafür, daß es sich bei einem Wort um einen Neologismus handelt, umso größer, je geringer seine Häufigkeit ist.[94]

Obwohl die Lexeme des *MXBT* unabhängig von ihrer Häufigkeit im untersuchten Text dargestellt sind, scheint ein vergleichbarer Zusammenhang zu bestehen.[95]

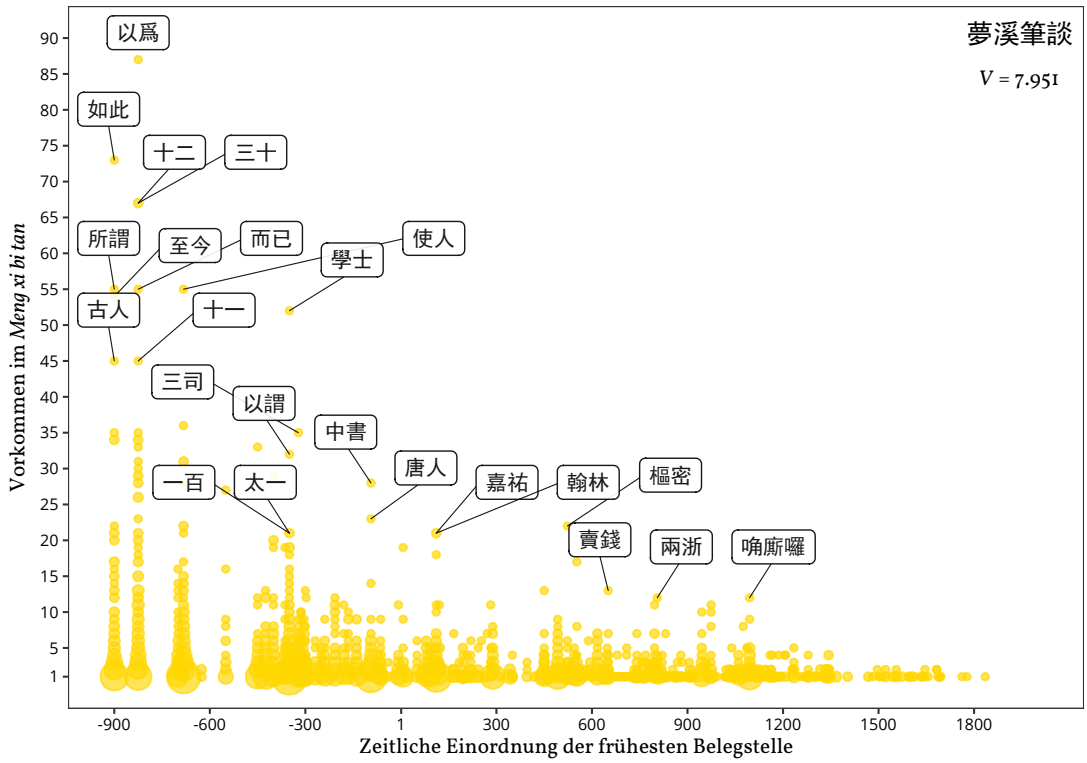


Abbildung 6.10 2–4 Zeichen Lexeme im *MXBT* chronologisch vs. Häufigkeit im Text

91 Vgl. auch Kapitel 5.7 (ab S. 138).

92 Die Erzeugung dieser Darstellung wird in Abschnitt 6.2.1, S. 182 erläutert, die gewählte Gewichtungskorrektur ab S. 185. In eckigen Klammern ist die ungewichtete für das jeweilige Jahrhundert festgestellte Anzahl *types* angegeben. Zusätzlich zu den Belegstellen aus dem *DHYDCD* werden weitere Daten herangezogen, wie in Kapitel 5.5.4 (ab S. 134) beschrieben.

93 Siehe Zipf 1947, S. 527; zitiert in Arapov und Cherc 1983 [1974], S. 29.

94 Arapov und Cherc 1983 [1974], S. 1; vgl. auch Zipf 1947, S. 523.

95 Vergleichbare logarithmische Verläufe der diachronen Verteilung des Vokabulars zeigen sich auch für andere Texte. Siehe z. B. Abb. 6.19b, S. 191, Abb. 6.20b, S. 191 und Abb. 6.25, S. 201.

Auch bei diachroner Betrachtung der Häufigkeiten von Lexemen im *MXBT* zeigt sich, dass die häufigsten *types* bereits sehr früh nachgewiesen sind (Abb. 6.10). In Richtung Gegenwart nimmt die Wahrscheinlichkeit für sehr häufige Lexeme immer weiter ab.[96]

Der überwiegende Anteil der Lexeme datiert vor die Entstehung des Textes im Jahr 1088. Zudem weist ihre chronologische Zuordnung im 11. Jh. eine Spitze auf (Abb. 6.9).[97] In den Zeitraum nach der tatsächlichen Entstehung des Textes (1100–2000) werden insgesamt noch 188 Lexeme datiert. Ihre Anzahl nimmt zwar kontinuierlich ab, zeigt aber auch, wie viele *types* allein in diesem einen Text enthalten sind, die zwar im *DHYDCD* lexikalisiert sind, deren älteste Belegstelle aber bis zu neun Jahrhunderte später datiert ist – obwohl das *MXBT* selbst im *HYDCD* zitiert wird und den Kompilator:innen offensichtlich vorlag. Einige Beispiele für Zeichenkombinationen, die deutlich früher belegbar sind als der *Locus classicus* im *HYDCD*, werden in Kapitel 5.5.4 besprochen.[98]

Während ein Großteil der als „zu neu" eingestuften Lexeme auf die „Nachlässigkeit" der *HYDCD*-Herausgeber zurückzuführen sein mag,[99] sind auch *false positives* vorhanden, die durch die naive *n*-Gramm-Segmentierung des Textes entstehen. So wird z. B. die Zeichenkombination *nancao* 南漕, im *MXBT* enthalten in *Huainan caoqu* 淮南漕渠 als Binomen dem 20. Jh. zugeordnet. Unabhängig von ihrer Ursache sind solche falschen Zuordnungen bei der Verwendung diachroner Lexikalisierungsdaten mit Text-*n*-Grammen gleichermaßen problematisch wie unvermeidbar.

6.2.1 Erzeugung von Neologismusprofilen

Die zur Erzeugung von Abb. 6.9 durchgeführten Schritte werden am Beispiel des *MXBT* erläutert:

1. Nicht verwendbare Zeichen (hier die englischsprachigen Angaben des *Project Gutenberg* zum *MXBT*) werden entfernt. Inklusive Interpunktion verbleiben 99.188 chinesische Zeichen.
2. Zur Normalisierung werden vorkommende Kurzzeichen durch Langzeichen ersetzt.[100] So werden u. a. alle Instanzen von *lu* 栌 durch 櫨 und *sha* 铩 durch 鎩 ersetzt.
3. Sonstige Varianten werden mit dem im *DHYDCD* vorgegebenen Standard normalisiert.[101] Dabei werden z. B. alle Instanzen von *wei* 為 durch 爲 (1.111 Vorkommen) und *xu* 敘 durch 敍 (52 Vorkommen) ersetzt. Insgesamt werden im *MXBT* so 49 Zeichen-*types* mit 1.594 Vorkommen ersetzt.[102]
4. Alle im Text enthaltenen 2–4-Gramme[103] werden ermittelt und gezählt.[104] Er enthält 226.029 2–4-Gramm *types*.

96 Vgl. auch Arapov und Cherc 1983 [1974], S. 51–87.
97 239 der 359 dem 11. Jh. zugeordneten Lexeme sind mit dem *MXBT* selbst belegt – die Spitze wäre also weniger markant, würde das *MXBT* nicht als Primärquelle für das *DHYDCD* dienen.
98 Siehe S. 135.
99 Siehe dazu auch Kapitel 5.3, ab S. 113.
100 Für diesen Schritt wird, wenn nötig, die Funktion `tradify` aus dem Paket `mafan` (Schaaf 2017, siehe auch Kapitel 4.3, S. 70) eingesetzt.
101 Siehe Kapitel 4.3, ab S. 69.
102 Sowohl durch `tradify` als auch `hydcd_standardize` können auch *types* erzeugt werden, die eigentlich nicht im Text enthalten sind, z. B. wenn *...li jian...* 里間 zu *lijian* 裏間 wird. Der Einsatz von Normalisierungswerkzeugen ist daher eine Abwägungsfrage, die je nach Art und Qualität der untersuchten Textdaten entschieden werden kann.
103 Einzelzeichen sind – zumindest in ihrer normalisierten Form – temporal kaum diskriminativ, wie auch in Kapitel 5.7 und 6.3 (ab S. 210) diskutiert.
104 Siehe dazu Kapitel 4.5.2, S. 91.

5. Aus der entstandenen Liste mit *n*-Gramm-*types* wird die Schnittmenge mit Einträgen im *DHYDCD* gebildet. 8.679 von den 226.029 im Text unterschiedenen 2–4-Grammen sind darin lexikalisiert.
6. Zu 8.210 (94,6 %) dieser Lexeme stehen chronologische Daten zur Verfügung, die aus der Datenbank geladen werden.
7. Jedes Lexem wird dem Jahrhundert zugeordnet, in das die älteste Quelle mit einer entsprechenden Belegstelle datiert ist. Bei ungenau datierten Quellen wird die Zuordnung – wie in Abschnitt 6.2.1 (S. 184) erläutert – ggf. aufgeteilt (*Slicing*).
8. Für die Darstellung in Abb. 6.9 wurde zudem eine Gewichtungskorrektur vorgenommen, welche die im *HYDCD* vorhandene Gewichtung ausgleichen soll.[105]

Verwendung zusätzlicher Belegstellen

Um die Problematik zu später Belegstellen im *DHYDCD* abzumildern, wurden für die zeitliche Zuordnung der Lexeme bzw. Zeichenkombinationen zusätzliche Korpus-Belegstellen herangezogen.[106] Abb. 6.11 zeigt die Neologismusprofile des *MXBT* mit Gewichtungskorrektur ohne diese zusätzlichen Daten (links), ergänzt um die Belege aus den *zhengshi*- und Loewe-Korpora (Mitte), sowie mit den *difangzhi* 地方誌-Belegen (rechts) im direkten Vergleich. Letztere stellen gerade im Bereich vom 17. bis Anfang des 20. Jhs. eine wichtige Ergänzung dar.

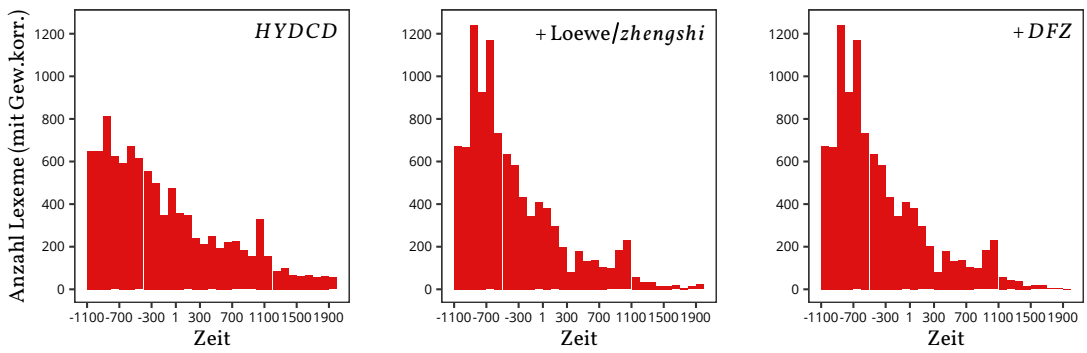


Abbildung 6.11 Profile des *MXBT* (nur *DHYDCD*-Belege; + Loewe-/*zhengshi* Belege; + *DFZ*-Belege)

Würden allein die Lexikalisierungsdaten aus dem *DHYDCD* vewendet, enthielten die meisten Texte also zahlreiche „zu neue" Wörter. Betrachtet man auf dieser Basis den Anteil verspätet belegter Zeichenkombinationen am Beispiel des *MXBT*, kann eine Fehlerquote von etwas weniger als 10 % konstatiert werden: Von den 8.136 Lexemen, zu denen chronologische Daten vorliegen, sind 7.393 dem Zeitraum zwischen 1100 v. u. Z. und 1100 zugeordnet, die restlichen sind rezenter. Mit den zusätzlichen Belegstellen sinkt diese Fehlerquote auf 2,3 %.

105 Die Gewichtungskorrektur wird ab S. 185 erläutert. Zur zeitlichen Gewichtung der *HYDCD*-Einträge siehe Kapitel 5.7.2, ab S. 142.
106 Siehe Kapitel 5.5.4, ab S. 134.

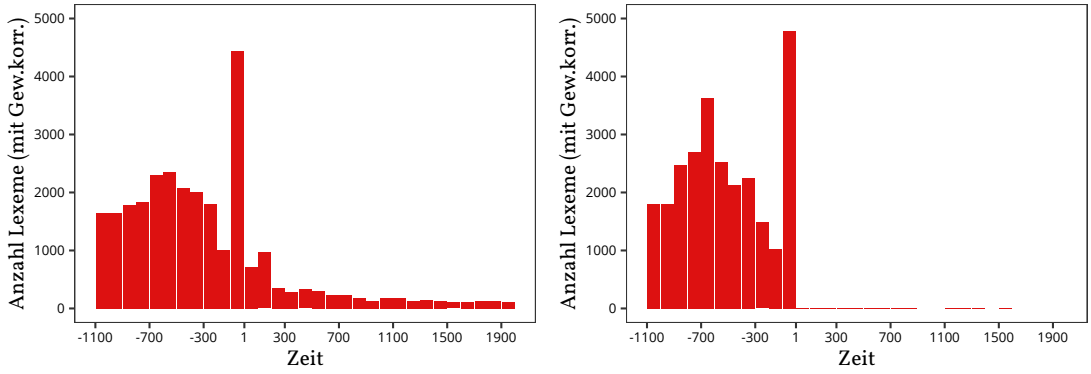


Abbildung 6.12 Neologismusprofile des *Shiji* ohne und mit zusätzlichen Korpus-Belegen

Bemerkenswert ist, dass auch im *DHYDCD* häufig zitierte Texte wie das *Shiji* 史記[107] eine hohe Anzahl an im *DHYDCD später* datierten Zeichenkombinationen enthalten können: 28,4 % der festgestellten Lexeme sind ohne die Erweiterungen der Datenbank später datiert als der Text selbst (Abb. 6.12).

Slicing – Zu welchem *chronon* gehört ein Lexem?

Um alle Lexeme eines Textes einem bestimmten Zeitraum – hier ein Jahrhundert – zuzuordnen, sollte der *Locus classicus* eindeutig auf ein bestimmtes Jahr datiert sein. Vor allem ältere Quellen sind aber schlechtestenfalls nur auf den Zeitraum einer ganzen Dynastie, oder zumindest der Lebensspanne des Autors genau datierbar.[108] Die Zuordnung muss gegebenenfalls also entsprechend aufgeteilt werden. Der Satz „鄜延境内有石油，舊説高奴縣出脂水，即此也。“[109] enthält z. B. 57 unterschiedliche 2–4-Gramm-*types*, von denen drei als Lexeme chronologisch zugeordnet werden können:

1. Die früheste Belegstelle zu *zhishui* 脂水 im *DHYDCD* stammt aus einem *fu* 賦 von Du Mu 杜牧 (803–852),[110] dessen Lebensdaten aus der *CBDB* mit 803–853 übernommen wurden. Der Datierungszeitraum fällt damit vollständig ins 9. Jh., was bei *chronons* von 100 Jahren unproblematisch ist.
2. *Shiyou* 石油 wird dem 11. Jh. zugerechnet, da das *MXBT* selbst als *Locus classicus* angegeben ist und in der *CBDB* auf das Jahr 1095 datiert ist.
3. *Jici* 即此 belegt das *DHYDCD* mit dem tangzeitlichen Gedicht *Qiu huai shi* 秋懷詩 von Han Yu 韓愈 (768–824[111]). Da der Zeitraum ungleich auf das 8. und 9. Jh. aufgeteilt ist, erfolgt eine anteilige Zurechnung zu beiden *chronons*, die ich als *Slicing* bezeichne. Von 57 Jahren entfallen 32 auf das 8., 25 auf das 9. Jh. Demnach werden dem 8. Jh. $\frac{32}{57}$ = 0,561 *types*, dem 9. Jh. $\frac{25}{57}$ = 0,439 *types* zugerechnet.

107 Siehe dazu Kapitel 5.7, ab S. 138.
108 Siehe v. a. Kapitel 5.5.2, ab S. 127 bzw. 5.5.3, ab S. 132.
109 *DHYDCD*, 石油. Die Belegstelle für *shiyou* stammt aus dem *MXBT*.
110 Siehe Emmerich 2004, S. 160.
111 Siehe ebd., S. 155, 768–825 lt. *CBDB*.

Der Beispielsatz aus dem *MXBT* hätte damit folgendes Profil, das mit nur drei Beobachtungen natürlich wenig Aussagekraft hat – es dient lediglich zur Veranschaulichung.

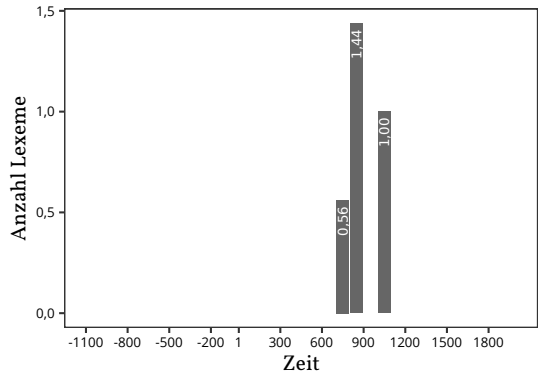


Abbildung 6.13 Neologismusprofil für „鄜延境内有石油，舊説高奴縣出脂水，即此也。“

Gewichtungskorrektur – Entfernung des *HYDCD-Bias*

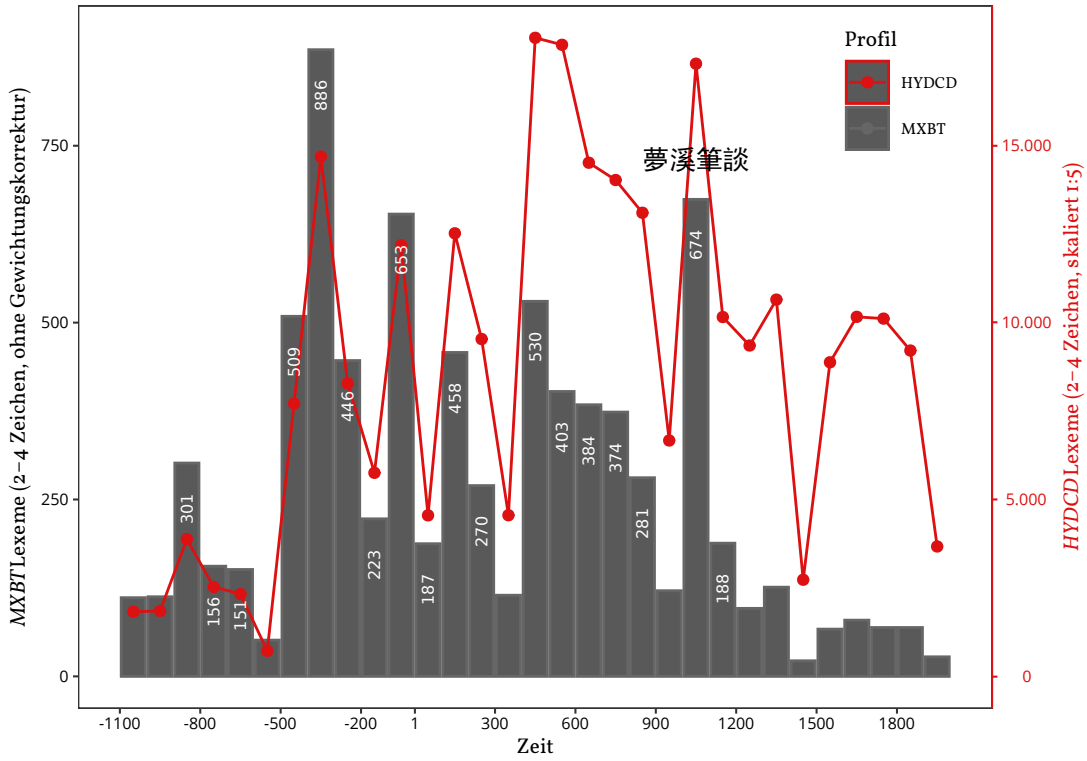


Abbildung 6.14 Neologismusprofil für das *Meng xi bi tan* vs. Lexikalisierung im *HYDCD*

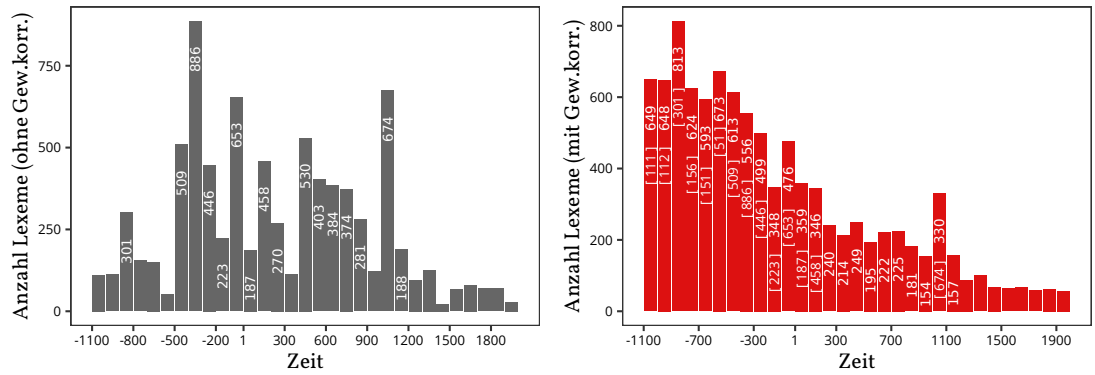
Wird – wie im gerade gezeigten Beispiel – die „rohe“ Anzahl *types* pro Jahrhundert zur Erzeugung von Neologismusprofilen verwendet, sind diese oft schwierig zu interpretieren, da die

dem Text inhärente Lexikalisierungsgeschichte vom *Bias* des *HYDCD* überdeckt wird, bzw. eine ungewollte Gewichtung stattfindet.[112] Abb. 6.14 zeigt das Profil des *MXBT* mit Rohdaten. Zum direkten Vergleich sind alle in das jeweilige Jahrhundert datierten 2–4-Zeichen-Lexeme aus dem *DHYDCD* im Maßstab 1:5 darüber gelegt (rot). Vor allem bis zur Entstehung des Textes im 11. Jh. folgen die für das *MXBT* beobachteten Schwankungen sehr deutlich der Lexikalisierung des *DHYDCD*.

Dieser Effekt lässt sich durch eine Gewichtungskorrektur gemäß der Lexikalisierungsdaten nahezu vollständig eliminieren. Zu diesem Zweck können unterschiedliche Überlegungen getroffen werden. Trifft man die stark vereinfachende Annahme einer „tendency of vocabulary to change at a uniform rate",[113] dass also die Aufnahme neuer Wörter in den Wortschatz ein insgesamt kontinuierlicher Prozess ist und in jedem Jahrhundert damit etwa gleich viele Wörter hinzukommen würden,[114] kann für jedes Jahrhundert im Beobachtungszeitraum ein Gewicht w_c berechnet werden, mit dem multipliziert sich für alle Jahrhunderte dieselbe Anzahl an *types* ergibt. Hierzu wird die Gesamtmenge aller über Belegstellen aus dem *DHYDCD* datierten Lexeme $|\ V\ |$ durch die Anzahl der Jahrhunderte im Betrachtungszeitraum $|\ C\ |$ und die Menge der auf das jeweilige Jahrhundert datierten Lexeme $|\ V_c\ |$ dividiert.

$$w_c = \frac{|\ V\ |}{\frac{|C|}{|V_c|}}$$

Wendet man w_c als Korrekturfaktor auf die y-Werte in Abbildung 6.15a an, ergibt sich ein homogeneres, einfacher zu interpretierenendes *Neologismusprofil* des *MXBT* (Abb. 6.15b, siehe auch Abb. 6.9):



(a) Ohne Gewichtungskorrektur **(b)** Mit Gewichtungskorrektur

Abbildung 6.15 Neologismusprofile des *MXBT* (ohne Korpus-Belegstellen)

112 Siehe auch Kapitel 5.7.2, Abb. 5.8, S. 143

113 Swadesh 1955, S. 122; vgl. auch Arapov und Cherc 1983 [1974], S. 89: „[I]n einigen Fällen [scheint] der lexikalische Wandel wirklich mit nahezu konstanter Geschwindigkeit zu erfolgen."

114 Diese veraltete Annahme aus der Lexikostatistik ist bestenfalls als stark vereinfachend anzusehen. Ebenfalls mögliche sprunghafte Erweiterungen des Wortschatzes durch wichtige gesellschaftliche Ereignisse, Umbrüche, Sprachkontakt, Internationalisierung, Technologiewandel etc. werden von solchen Modellen nicht berücksichtigt. Zudem konnte verschiedentlich gezeigt werden, dass der Wortschatzzuwachs, wie auch andere Manifestationen von Sprachwandel, eher einer *s*-Kurve folgt.

Gegen die Annahme einer konstanten Veränderung des Vokabulars spricht, dass auch die Lexikalisierungsdaten des *DHYDCD* suggerieren, dass seine Zunahme einer *s*-Kurve folgt. Für eine entsprechende *s*-Gewichtungskorrektur der Neologismusprofile (Abb. 6.16) kann – wie bereits in Kapitel 5.7.2 gezeigt – die Funktion der idealisierten Lexikalisierung mit *R* geschätzt und die resultierenden Werte der theoretischen Neulexikalisierung für jedes Jahrhundert berechnet werden.[115] Der *s*-Korrekturfaktor s_c ergibt sich aus dieser idealisierten und der tatsächlich gemessenen Lexikalisierung des jeweiligen Jahrhunderts.

$$s_c = \frac{| V_{c_{ideal}} |}{| V_c |}$$

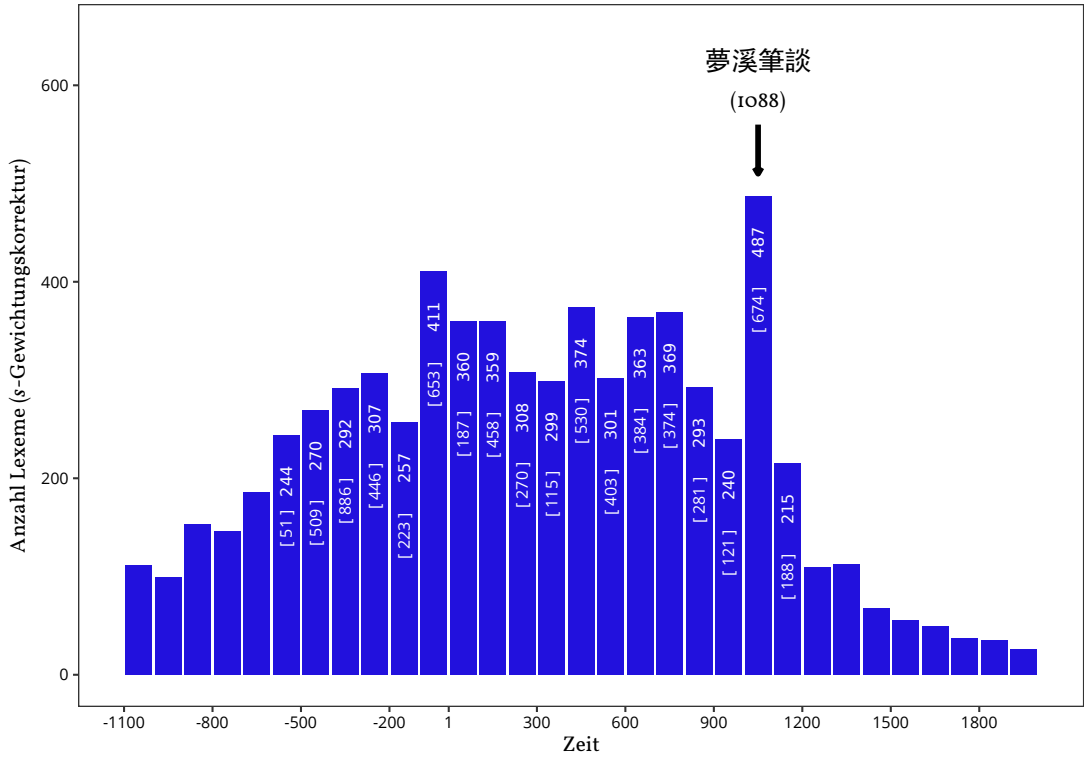


Abbildung 6.16 Neologismusprofil für das *MXBT* (*s*-Gewichtungskorrektur, ohne Korpusbelegstellen)

Welche Darstellung als Grundlage für eine philologische Interpretation vorzuziehen ist, mag von subjektiven Präferenzen bestimmt sein – für eine automatisierte Datierung funktioniert die einfache Annahme des konstanten Wortschatzwachstums jedoch minimal besser, wie in Abschnitt 6.2.5 (ab S. 197) gezeigt wird. Bei optischer Analyse des Profils sind eventuelle Peaks mit linearer Gewichtungskorrektur zudem gegebenenfalls leichter erkennbar.

Die Werte für den Faktor w_c werden gemäß der im *DHYDCD* aufgezeichneten Lexikalisierung pro Jahrhundert berechnet. Damit sind sie abhängig von der Zeichenlänge der Lexeme und der Veränderung ihrer Datierung durch die oben beschriebene Ergänzung früherer Belegstel-

115 Wie bereits in Kapitel 5.7.2, v. a. S. 146, verwende ich hierfür die Funktion `drm` aus dem *R*-Paket `drc`. Siehe Ritz 2016.

len. Tabelle 6.9 gibt die entsprechenden Werte von w_c für den gesamten Betrachtungszeitraum wieder. Die ersten vier Wertespalten beziehen sich auf 2–4 Zeichen-Lexeme ohne Berücksichtigung von zusätzlichen Korpusbelegstellen. Bei einer durchschnittlichen Lexikalisierung von 8.305 Wörtern pro Jahrhundert ergibt sich aus der naiven Annahme einer linearen Lexikalisierung der jeweilige Faktor aus der Anzahl der *types* V_c. Die Spalte $V_{c_{ideal}}$ gibt die mit der Annahme eines *s*-förmigen Wortschatzwachstums modellierte Lexikalisierung an, die nächste Spalte den daraus errechneten *s*-Faktor. Da dem Modell eine kumulative Betrachtung zugrunde liegt, wird der Wortschatzzuwachs als Differenz zum jeweils vorherigen Jahrhundert berechnet. Der erste Faktor s_{-1100} ist daher mit 1 angegeben. Für die aufgezeichnete Lexikalisierung nach Berücksichtigung der zusätzlichen Korpusbelegstellen sind ebenfalls die Werte von V_c und w_c, jeweils für die Betrachtung von 2–4 und 2–3-Zeichen Lexemen angegeben.[116]

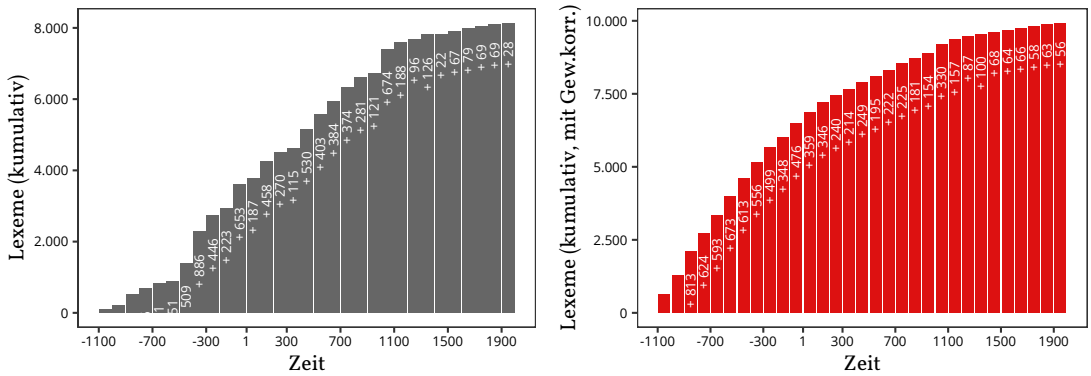
Tabelle 6.9 Lexikalisierung und Gewichtungskorrekturfaktoren nach Jahrhundert

Zeitraum	2–4 Z., ohne Korpusbelege				2–4 Z., m. Belegen		2–3 Z., m. Belegen	
Jh.	# *types* V_c	w_c	$V_{c_{ideal}}$	s_c	V_c	w_c	V_c	w_c
1100-1000 v. u. Z.	1.423	5,836	1.423	1	1.612	5,165	1.509	5,100
1000-900 v. u. Z.	1.442	5,761	1.624	1,126	1.627	5,115	1.523	5,053
900–800 v. u. Z.	3.082	2,695	1.968	0,639	4.099	2,030	3.777	2,038
800–700 v. u. Z.	2.071	4,010	2.378	1,148	2.967	2,806	2.842	2,708
700–600 v. u. Z.	2.114	3,929	2.862	1,354	4.055	2,053	3.974	1,937
600–500 v. u. Z.	632	13,133	3.430	5,424	766	10,861	708	10,871
500–400 v. u. Z.	6.889	1,206	4.089	0,594	8.023	1,037	7.449	1,033
400–300 v. u. Z.	13.224	0,628	4.844	0,366	18.293	0,455	17.178	0,448
300–200 v. u. Z.	7.425	1,119	5.697	0,767	9.197	0,905	8.678	0,887
200–100 v. u. Z.	5.312	1,563	6.641	1,250	6.363	1,308	6.084	1,265
100–1 v. u. Z.	11.399	0,729	7.663	0,672	13.457	0,618	12.680	0,607
1–100	4.332	1,917	8.739	2,017	6.972	1,194	6.691	1,150
100–200	10.993	0,756	9.834	0,895	16.335	0,510	15.640	0,492
200–300	9.336	0,890	10.902	1,168	11.954	0,696	11.401	0,675
300–400	4.453	1,865	11.886	2,669	3.382	2,461	3.209	2,399
400–500	17.693	0,469	12.725	0,719	18.510	0,450	17.632	0,436
500–600	17.193	0,483	13.362	0,777	16.836	0,494	16.116	0,478
600–700	14.350	0,579	13.749	0,958	11.995	0,694	11.109	0,693
700–800	13.773	0,603	13.853	1,006	10.149	0,820	9.514	0,809
800–900	12.854	0,646	13.666	1,063	10.094	0,825	9.521	0,808
900–1000	6.513	1,275	13.205	2,027	8.191	1,016	7.554	1,019
1000–1100	16.960	0,490	12.504	0,737	13.028	0,639	12.098	0,636
1100–1200	9.947	0,835	11.617	1,168	7.531	1,105	6.729	1,144
1200–1300	9.161	0,907	10.604	1,158	7.976	1,044	7.036	1,094
1300–1400	10.466	0,794	9.523	0,910	9.999	0,832	9.192	0,837
1400–1500	2.682	3,097	8.429	3,143	1.999	4,163	2.656	2,897
1500–1600	8.706	0,954	7.365	0,846	6.631	1,255	8.455	0,910
1600–1700	9.938	0,836	6.363	0,640	8.055	1,033	7.062	1,090
1700–1800	9.939	0,836	5.444	0,548	7.435	1,119	5.605	1,373
1800–1900	9.076	0,915	4.619	0,509	6.561	1,269	3.552	2,167
1900–2000	4.081	2,035	3.892	0,953	3.925	2,121	1.413	5,446
Ø V_c	8.305				8.323		7.696	

116 Letztere ist für die Verwendung von *n*-Gramm Datensätzen, deren *n*-Gramm Raum auf 3 limitiert ist, unverzichtbar. Siehe Abschnitt 6.2.5, ab S. 197, bzw. Kapitel 4.2, S. 66.

Kumulative Darstellung

In einer kumulativen Darstellung (Abb. 6.17) ließe sich unter optimalen Bedingungen ein zunächst logarithmischer Anstieg beobachten, der ab dem Jahrhundert, aus dem der untersuchte Text stammt, stark abflacht. Da die Lexikalisierung pro Jahrhundert allerdings deutlich schlechter erkennbar ist und für ein echtes Abflachen zu viele *false positives* vorhanden sind, ist die distinktive Darstellung vorzuziehen.



(a) Ohne Gewichtungskorrektur

(b) Mit Gewichtungskorrektur

Abbildung 6.17 Kumulative Neologismusprofile des *Meng xi bi tan* (ohne Korpusbelegstellen)

6.2.2 Temporale Textprofile: Erweiterung um Namen und Zeitausdrücke

Neben Wörtern bzw. Lexemen enthalten Texte oft zusätzliche zeitlich konnotierte Zeichenfolgen. Dazu zählen *temporal expressions* – Erwähnungen bestimmter Jahre oder Monate, sowie Personennamen ab einer Länge von drei Zeichen,[117] die mit der Lebensspanne der genannten Person eine temporale Dimension gewinnen.[118] Zu diesem Zweck werden Informationen zu Personen aus der *CBDB* geladen,[119] sowie auf Basis der *DDBC* nach *temporal expressions* gesucht. Wie in Kapitel 4.8 diskutiert, werden Zeitangaben in schriftsprachlichen chinesischen Texten gewöhnlich in Form von Regierungsdevisen und Jahresangaben gemacht, gegebenenfalls gefolgt von Monats- und Datumsangaben im lunisolaren Kalender.[120]

Die Neologismusprofile lassen sich so zu einem *temporalen Textprofil* erweitern, das die Informationen zusammenfassend darstellt. Theoretisch denkbar ist zudem die Nutzung von Ortsnamen, die ebenfalls in der *CBDB* auf die früheste den Herausgebern vorliegende Nennung datiert sind. Dass diese Daten mit Vorsicht zu genießen sind, zeigt Abb. 6.18. Gut ein Drittel der insgesamt 102 Übereinstimmungen mit unterschiedlichen Ortsnamen sind dem 12.–16. Jh. zugeordnet.

Auch unter den Personennamen finden sich – trotz des Ausschlusses uneindeutiger Namen – einzelne *false positives*, die ebenfalls dem 12.–16. Jh. zugeordnet sind. Eine ein-eindeutige Zuordnung von Namen zu bestimmten Personen ist nicht zuverlässig möglich, da Texte andere

117 Namen mit zwei Zeichen weisen ein sehr hohes Ambiguitätspotenzial auf. Siehe Kapitel 4.7, ab S. 97.
118 Siehe auch Abschnitt 6.1.1, S. 159. Es werden dabei nur Namen berücksichtigt, die in der *CBDB* nur einer einzigen Person zugeordnet sind.
119 Siehe Kapitel 4.7, ab S. 97.
120 Siehe Kapitel 4.8, ab S. 103.

Personen gleichen Namens erwähnen können. Außerdem können Zeichenfolgen, die für Namen verwendet werden, in derselben Kombination auch in lexikalisierten Bedeutungen auftreten.[121]

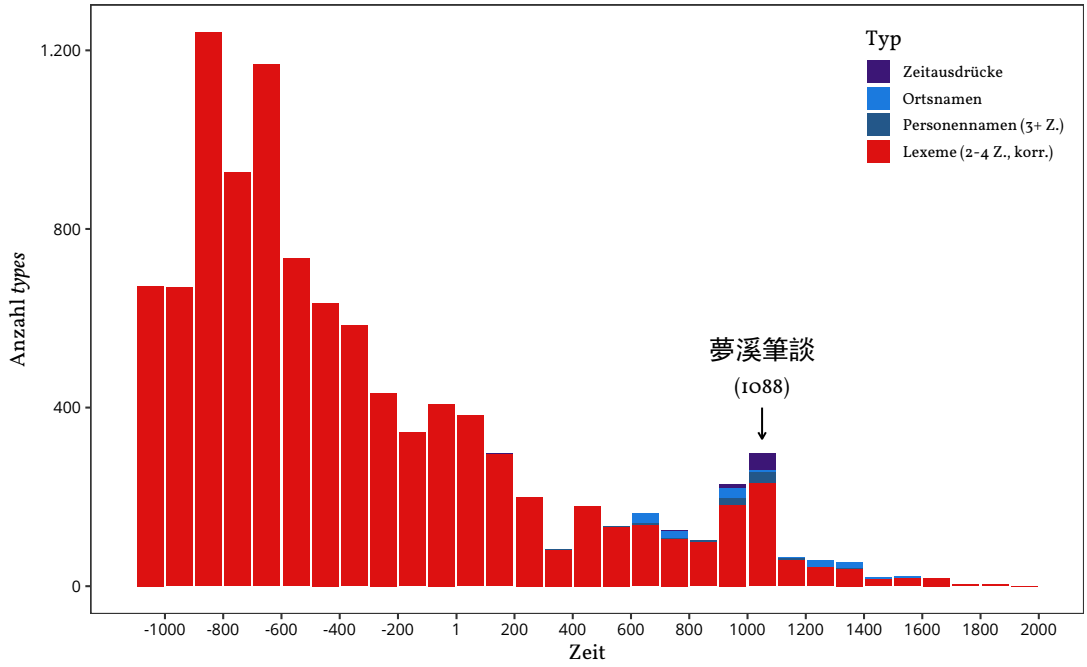


Abbildung 6.18 Temporales Textprofil für das *Meng xi bi tan*

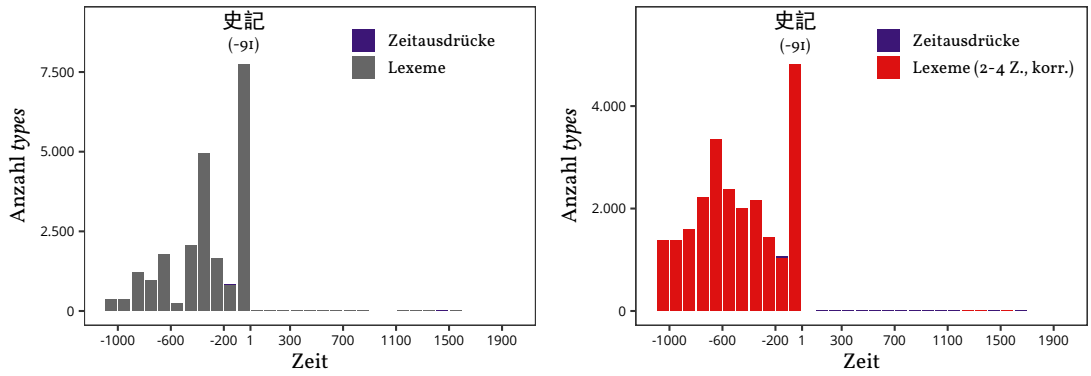
6.2.3 Interpretation temporaler Textprofile

Für den bereits gezeigten Fall des *MXBT* lässt sich aus dem temporalen Textprofil mit dem 11. Jh. bereits graphisch ein wahrscheinlicher Zeitraum der Entstehung des Textes ablesen. Er enthält einen großen Anteil älteres Vokabular, der dann zum Jahrhundert der Textgenese hin abnimmt. Zudem ist wieder ein größerer Anteil an Vokabular aus der Zeit der Textentstehung enthalten, sowie Namen von Zeitgenoss:innen (Abb. 6.18). Es kommt uns zugute, dass das *MXBT* selbst von den Kompilator:innen des *DHYDCD* gerne als *Locus classicus* herangezogen wurde.[122] Zudem wird im Text auf rezente historische Ereignisse Bezug genommen, so dass die chronologische Zuordnung von *temporal expressions* bekräftigt wird.[123]

Einfacher wäre die Interpretation bei Vollständigkeit der historischen Lexemdatenbank: Der untersuchte Text könnte in der Regel dem spätesten Jahrhundert zugeordnet werden, aus dem Lexeme nachgewiesen sind. Durch Betrachtung von zwei *zhengshi*-Texten aus dem Trainingskorpus[124] lassen sich diese Optimalbedingungen simulieren. *Shiji* 史記 ist auf die Lebensspanne von

121 Siehe dazu die Erläuterungen und Beispiele in Kapitel 4.7, ab S. 97.
122 Von 8.129 betrachteten Lexemen werden 359 dem 11. Jh. zugeordnet, bei 239 davon ist das *MXBT* selbst als *Locus classicus* angegeben.
123 Insgesamt 113 unterschiedliche Angaben werden erkannt, davon 54 mit Angaben zum 11. Jh, z. B. *Yuanfeng wu nian* 元豐五年 („das fünfte Jahr [i. e. 1082] der Regierungsdevise *Yuanfeng* [1078–1085] von Kaiser Shenzong 神宗 der Song-Dynastie (reg. 1067–1085“)
124 Siehe dazu Kapitel 5.5.4, ab S. 134.

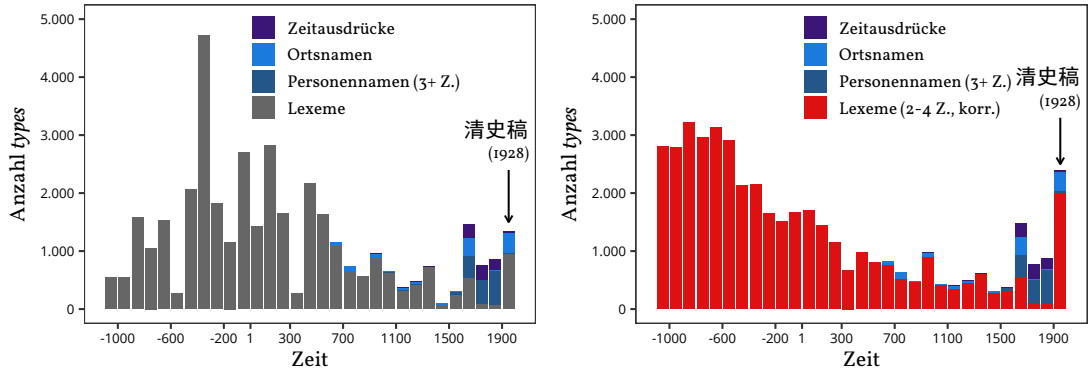
Sima Qian 司馬遷 (ca. 145–90 v. u. Z.) datierbar,[125] das 1928 veröffentlichte *Qingshi Gao* 清史稿 behandelt die Geschichte der Qing-Dynastie (清, 1644–1912) und ist in einem an das *Shiji* angelehnten Stil verfasst.



(a) *Shiji* **(b)** *Shiji*, mit Gewichtungskorrektur

Abbildung 6.19 Temporale Profile des *Shiji* 史記

Bei Verwendung der zusätzlichen Belegstellen sind im Profil des *Shiji* kaum *types* sichtbar, die erst nach dem 1. Jh. v. u. Z. belegt sind. Eine Spitze, die noch deutlicher auf dieses Jahrhundert hinweist, ist wenig überraschend, da der Text selbst häufig als Belegstelle herangezogen wird.[126]



(a) *Qingshi gao* **(b)** *Qingshi gao*, mit Gewichtungskorrektur

Abbildung 6.20 Temporale Profile des *Qingshi gao* 清史稿

Für das *Qingshi gao* sind etliche Lexeme sichtbar, die über den gesamten Beobachtungszeitraum, bis ins 20. Jh., datiert werden. Im Profil ohne Gewichtungskorrektur (Abb. 6.20a) macht sich an den starken Schwankungen in der nachgewiesenen Anzahl an Lexemen erneut das bereits an-

125 Das *Shiji* wurde von Sima Tan 司馬談 (gest. 110 v. u. Z.) begonnen und dann von Sima Qian fertiggestellt. Auch wenn begründete Zweifel an der Authentizität einzelner Kapitel bestehen, bei denen es sich um spätere Rekonstruktionen handeln könnte, sollte der Text im Wesentlichen zu Lebzeiten Sima Qians entstanden sein. Siehe Anthony François Paulus Hulsewé 1993b: „Shih chi 史記". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 405–406.

126 Siehe Kapitel 5.7, ab S. 138.

gesprochene *Bias* bemerkbar. Der hohe Anteil an Vorkommen von Personennamen und *temporal expressions* für den Zeitraum zwischen 1600–1900 spiegelt zudem den Inhalt des Texts wider.

Selbst mit kontinuierlicher Ergänzung der Lexemdatenbank durch frühere Belegstellen ist ein Zustand, in dem für *alle* Lexeme die früheste Belegstelle korrekt und genau datiert ist, quasi unerreichbar. In der Praxis kann die Interpretation zudem auch dann komplexer und aufwändiger sein, wenn Texte in einem „altertümelnden Stil" abgefasst sind, der nicht nur syntaktisch die Entstehungszeit des Textes verschleiert, sondern auch zeitgenössisches Vokabular bewusst vermeidet. Dies kann entweder durch bewusste Fälschung geschehen,[127] oder ist den Anforderungen oder Gepflogenheiten bestimmter Textgattungen, stilistischen Trends oder Vorlieben von Autor:innen geschuldet.

6.2.4 Das *Zhongjing* 忠經 als Anwendungsbeispiel

Ein schriftsprachlicher Text, dessen Entstehungszeit von Sinolog:innen diskutiert wird, ist das *Zhongjing* 忠經 („Klassiker der Loyalität").[128] Er wird traditionell Ma Rong 馬融 (79–166) zugeschrieben, was eine Datierung in die Han 漢-Zeit impliziert.[129] Die Autorschaft gilt jedoch als widerlegt, da Zitate aus Alttext-Teilen des *Shangshu* 尚書 (*Guwen Shangshu* 古文尚書) enthalten sind. Diese Textteile entstanden wahrscheinlich erst Anfang des 4. Jahrhunderts.[130]

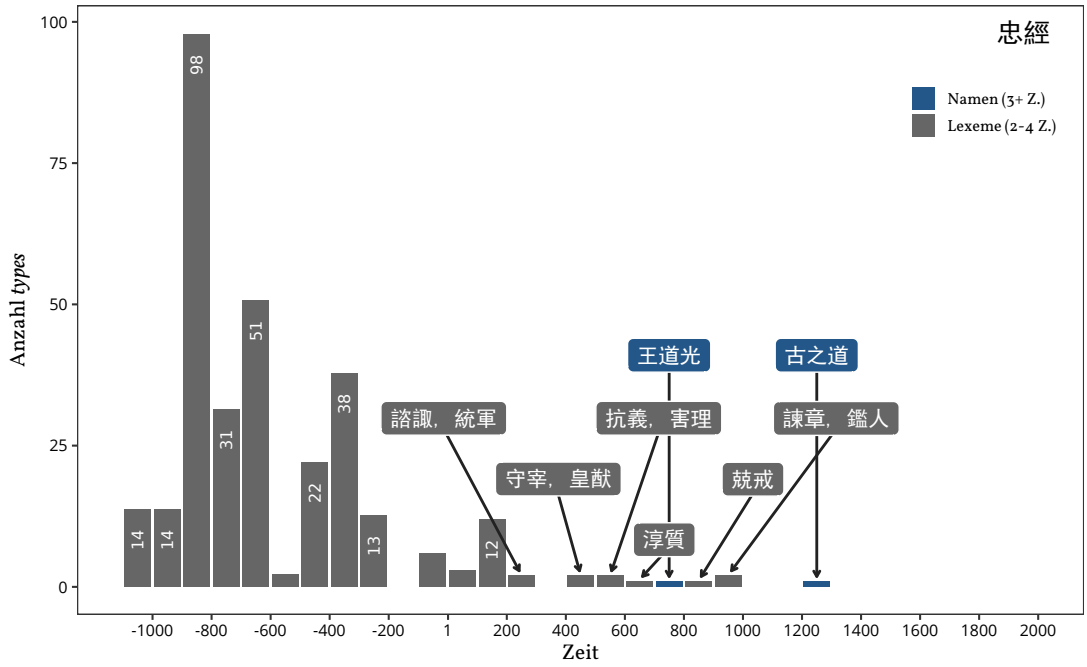


Abbildung 6.21 Temporales Profil des *Zhongjing* 忠經, ohne Gewichtungskorrektur

127 Siehe dazu auch Kapitel 3, S. 37.
128 Siehe auch Kapitel 3, ab S. 35. Den Hinweis, das *Zhongjing* in diesen Kontext zu stellen, verdanke ich Kai Vogelsang.
129 Siehe Suwald 2008, S. 7.
130 Siehe Nylan 2001, S. 134; zitiert in Suwald 2008, S. 70, siehe auch Kapitel 3, S. 39.

Die tatsächliche Entstehungszeit des *Zhongjing* bleibt Gegenstand von Spekulationen, die sich auf einen Zeitraum zwischen etwa dem Jahr 320 und dem Beginn der Song 宋-Dynastie (960–1279) erstrecken Aus jener Zeit stammen die ältesten schriftlich überlieferten Belege für die Existenz des *Zhongjing*.[131] Bedenkt man, dass die entlarvenden Zitate theoretisch auch nachträglich zu der heute vorliegenden Textfassung hinzugefügt worden sein können, bleibt sogar die traditionelle Zuschreibung im Bereich des Möglichen.

Betrachten wir das temporale Textprofil einer digitalen Ausgabe (Abb. 6.21).[132] Acht Lexeme und zwei „Namen" datieren später als das 3. Jh. Letztere sind mit einem Blick auf den Kontext schnell als *false positives* entlarvt: Wang Daoguang 王道光[133] (gest. 751) und Gu Zhidao 古之道[134] (früheste Belegstelle von 1294) kommen beide nicht als Namen, sondern als Folgen von Zeichen mit lexikalisierten Bedeutungen vor. Zeichenkombinationen, die erst nach der Song-Zeit nachgewiesen sind, enthält der Text nicht, was Suwalds Annahme, dass eine Entstehung des Textes nach 1040 sehr unwahrscheinlich ist,[135] bekräftigt.

Die acht Lexeme, die in den Zeitraum zwischen dem 4. und 10. Jh. datiert sind (Tabelle 6.10), können nun mit überschaubarem Aufwand geprüft werden, etwa durch Suche nach weiteren Belegstellen. Der Verdacht auf eine spätere Datierung lässt sich somit erhärten oder abschwächen. Eine derartige Vorgehensweise entspricht im Wesentlichen auch derjenigen Suwalds, die ebenfalls – allerdings ohne entsprechende Datenbank – mithilfe der vorkommenden Zeichenfolgen der Entstehungszeit des *Zhongjing* auf den Grund geht. Dabei kann sie zusätzlich Ausdrücke untersuchen, die im *DHYDCD* nicht lexikalisiert sind, z. B. *Zhou Kong zhi cai* 周孔之才.[136]

Tabelle 6.10 2–4-Zeichen-Kombinationen im *Zhongjing* mit Nachweisen *nach* dem 3. Jh.

#	Lexem	belegt in:[137]	datiert auf:	Vork.	Kontext
1	*jiànzhāng* 諫章	*Jiu Tang shu* 舊唐書	945	1	…書》云「旌別淑忒」，其是謂乎忠諫章第十五忠臣之事君也，莫先於諫，…
2	*jiànrén* 鑑人	*Jiu Tang shu* 舊唐書	945	1	…後從諫則聖。」證應章第十六惟天鑑人，善惡必應。善莫大於作忠，惡莫…
3	*jīngjiè* 兢戒	*Shang Jiang shilang qi* 上蔣侍郎啟	*ca.* 860–866	1	…之，天下盡忠，以奉上也。是以兢兢戒慎，日增其明，祿賢官能，式歉大…
4	*chúnzhì* 淳質	*Sui shu* 隋書	617	1	…，則人不爭。故得人心和平，天下淳質，樂其生，保其壽，優遊聖德，以…
5	*kàngyì* 抗義	*Wei shu* 魏書	*ca.* 551–554	1	…則非忠臣。夫諫，始於順辭，中於抗義，終於死節，以成君休，以寧社稷…
6	*hàilǐ* 害理	*Nanqi shu* 南齊書	*ca.* 509–537	1	…審則分。君子去其私，正其色，不害理以傷物，不憚勢以舉任。惟善是與…
7	*shòuzǎi* 守宰	*Hou Han shu* 後漢書	*ca.* 445	2	…詩》云「靖共爾位，好事正直。」守宰章第五在官惟明，蒞事惟平，立身… …觀乎子，則人愛之，如愛其親，蓋守宰之忠也。《詩》云「愷悌君子，民…
8	*huángyóu* 皇猷	*Song shu* 宋書	*ca.* 492–493	1	…大化，惠澤長久，萬民咸懷。故得皇猷丕丕，行於四方，揚於後代，以保…

131 Eine ausführliche Diskussion findet sich in Suwald 2008, S. 71–77.

132 *Zhongjing* 忠經. Unkommentierte Ausgabe auf *WikiSource*. URL: https://zh.wikisource.org/zh-hant/%E5%BF%A0%E7%B6%93 (besucht am 30. 03. 2019), Diese Version des Textes hat eine Länge von 2.499 Zeichen mit insgesamt 6.287 2–4-Gramm-*types*, von denen 326 im *DHYDCD* lexikalisiert sind. Zu 313 davon liegen chronologische Daten vor.

133 „忠臣之事君也，莫先於諫，下能言之，上能聽之，則王道光矣。" ebd., 辨忠章十四.

134 揚聖章第十三 ebd., „不足則補之，聖明則揚之，古之道也。".

135 Siehe Suwald 2008, S. 80, S. 82.

136 Siehe ebd., v. a. S. 78–80.

1. *jianzhang* 諫章 ist ein *false positive*, das durch die fehlende Segmentierung entsteht: *zhong jian* 忠諫 ist der Titel des 15. Kapitels.
2. *jianren* 鑑人 ist im Kontext als Verb-Objekt-Konstruktion zu verstehen.[138]
3. Bei *jingjie* 兢戒 handelt es sich ebenfalls um ein *false positive*.[139] *jingjing* 兢兢 ist bereits im *Shijing* 詩經 belegt,[140] *jie shen* 戒慎 im *Liji* 禮記.[141]
4. *chunzhi* 淳質 („rein und natürlich")[142] ist erst im *Sui shu* 隋書 nachgewiesen und kann als Indiz für eine Entstehung des *Zhongjing* deutlich nach der Han-Zeit gewertet werden.[143]
5. Dasselbe gilt für *kangyi* 抗義 („Einspruch")[144], das erst mit dem *Wei shu* 魏書 belegt ist.
6. Für die Kombination *hai li* 害理 (etwa: „Ordnungen zerstören")[145] ist das *Nanqi shu* 南齊書 als früheste Textstelle angegeben, was als weiterer Hinweis für eine spätere Entstehung gewertet werden kann.
7. *shouzai* 守宰 (etwa: „Gebietsverwalter")[146] ist der Titel des fünften *Zhongjing*-Kapitels. Da der Begriff mit dem *Hou Han shu* 後漢書 belegt ist, kann davon ausgegangen werden, dass er während der Han-Zeit bereits Verwendung fand.
8. *huangyou* 皇猷 („kaiserliche Vorhaben")[147] ist – mit Ausnahme des *Zhongjing* selbst – ebenfalls erst im *Song shu* 宋書 belegt.

Da auch das *Zhongjing* selbst als Belegquelle für ein Lexem angegeben sein kann, oder es aufgrund der traditionellen Datierung von den Herausgeber:innen des *DHYDCD* fälschlich als *Locus classicus* angenommen worden sein kann, müssen auch diejenigen Lexeme beleuchtet werden, die mit dem *Zhongjing* belegt sind (Tabelle 6.11).[148]

1. Für *bingzhi* 秉職 „an der Pflicht festhalten"[149] lässt sich mithilfe des *Chinese Text Project* eine Han-zeitliche Belegstelle im Text *San lüe* 三略 finden.[150]
2. *jingzhi* 敬職 („die Ämter mit Ehrerbietigkeit versehen")[151] lässt sich mit dem ebenfalls Han-zeitlichen *Qian fu lun* 潛夫論 belegen.[152]
3. *qianyun* 潛運 ist erst in einem Text von 魏收 Wei Shou (506–572) mit dem Titel *Xi Liang wen* 檄梁文 belegt. Die Textstelle wird im Tang-zeitlichen *Yiwen leiju* 藝文類聚 (erschienen 624) wiedergegeben.[153] Suwald bemerkt, dass der gesamte Ausdruck *chen mou qian yun* 沉謀潛

137 Angaben aus der diachronen Lexemdatenbank.
138 „惟天鑑人，善惡必應。" – „Nur der Himmel überblickt den Menschen, gut und böse werden gewiß beantwortet." Übs. aus Suwald 2008, S. 249.
139 Die Textstelle im *Zhongjing* lautet „是以兢兢戒慎，日增其明。" 2019, Kapitel 2 聖君章; „Wer deshalb auf vorsichtige Weise achtsam und aufmerksam ist, dessen Klarheit wird sich täglich steigern." Suwald 2008, S. 200.
140 *DHYDCD*, 兢兢.
141 *DHYDCD*, 戒慎; Siehe auch Suwald 2008, S. 200.
142 Suwald 2008, S. 78, siehe auch S. 238.
143 Vgl. auch ebd., S. 78.
144 Ebd.
145 Ebd., S. 126.
146 Ebd., S. 88.
147 Ebd., S. 78.
148 Im *DHYDCD* wird der Text Ma Rong 馬融 zugeschrieben. Sechs Lexeme werden aufgrunddessen in der Graphik dem 2. Jh. zugeordnet.
149 Suwald 2008, S. 78.
150 Huang Shigong 黃石公 ca. 100–9 v. Chr. *San lüe* 三略. Hrsg. von Donald Sturgeon. ctext.org.
151 Suwald 2008, S. 226.
152 Wang Fu 王符 ca. 102–167 v. Chr. *Qian Fu Lun* 潛夫論. Hrsg. von Donald Sturgeon. ctext.org.
153 Ouyang Xun 歐陽詢 et. al. 0624: *Yiwen leiju* 藝文類聚. Hrsg. von Donald Sturgeon. ctext.org.

運 („tiefliegende Pläne und verborgene Wendungen“) erst wieder in einem mingzeitlichen Text zu finden ist.[154]

4. Ebenfalls im *Yiwen leiju* ist die Phrase *zhigong wusi* 至公無私 („höchstes Allgemeinwohl und keine Eigensucht“[155]) zu finden. Zitiert wird aus dem deutlich älteren Text *[Shengxian] gaoshi zhuan [zan]* [聖賢] 高士傳 [贊] von Ji Kang 嵇康 (223–262)[156].
5. Der Begriff *zhongchen* 冢臣, „herausragender Untertan“[157] ist Titel des dritten *Zhongjing*-Kapitels. Die Herausgeber des *DHYDCD* finden erst im Qing-zeitlichen *Diaoqiao zhuang ge* 雕橋莊歌 eine weitere Belegstelle.[158]
6. Während die erste Textstelle mit *zhongneng* 忠能 eine Subjekt-Verb-Konstruktion („Die Loyalität vermag es…“) und damit ein *false positive* ist, kann die zweite Textstelle als Lexem (etwa: „[seine] Loyalitätskompetenz“)[159] gelesen werden. Eine weitere Belegstelle findet sich im *Yin Wen zi*, für dessen überlieferte Fassung aber eine spätere Datierung als die Han-Zeit angenommen wird.[160]

Tabelle 6.11 2–4-Zeichen-Kombinationen im *Zhongjing* mit inzestuösen Belegstellen

#	Lexem	belegt in:	datiert auf:[162]	Vork.	Kontext
1	*bǐngzhí* 秉職	*Zhongjing* 忠經	166	1	…行其政，居則思其道，動則有儀。秉職不回，言事無憚，苟利社稷，則不…
2	*jìngzhí* 敬職	~ ~	~	1	…以之而克則無怨，夫如是，則天下敬職，萬邦以寧。《詩》云「載馳載驅…
3	*qiányùn* 潛運	~ ~	~	1	…色直辭，臨難死節而已矣在乎潘謀潛運，正己安人，任賢以爲理，端委而…
4	*zhìgōngwúsī* 至公無私	~ ~	~	1	…所覆，莫大乎忠。忠者，中也，至公無私。天無私，四時行地無私，萬物…
5	*zhǒngchén* 冢臣	~ ~	~	2	…詩》云「昭事上帝，聿懷多福。」冢臣章第三爲臣事君，忠之本也，本立… …臣事君，忠之本也，本立而化成。冢臣於君，可謂一體，下行而上信，故…
6	*zhōngnéng* 忠能	~ ~	~	2	…心之謂矣。爲國之本，何莫由忠。忠能固君臣，安社稷，感天地，動神明… …者備矣，然後可以理人。君子盡其忠能，以行其政令，而不理者，未之聞…

Mit der obigen Analyse kann weder die Han-zeitliche Entstehung des Textes wider-, noch eine spätere Textgenese belegt werden. Die im Text gefundenen Lexeme liefern uns aber Hinweise, die aus lexikographischer Sicht für eine Entstehung des Textes nach der Han- und spätestens wäh-

154 Siehe Suwald 2008, S. 78.
155 Siehe auch ebd., S. 195. *gong* 公 und *si* werden bereits bei Han Fei 韓非 als Antonyme präsentiert.
156 Siehe Zhu Jinxiong 朱錦雄 2013: „*Lun* Ji *Kang ‚Shengxian gaoshi zhuanzan‘ zhong de ‚gao shi‘ fanxing* 論嵇康《聖賢高士傳贊》中的「高士」範型 (‚Diskussion des Begriffs *gaoshi* in Ji Kangs *Biographien heiliger gaoshi (etwa: untadeliger Menschen)*‘)“. In: 國立臺北教育大學語文集刊 *Guo li Taibei daxue yuwen jikan (Journal of Language and Literature Studies)* 24, S. 233–260, S. 236.
157 Suwald 2008, S. 202. Suwald vermutet eine Anspielung auf den *Zhou*-zeitlichen Begriff *zhongzai* 冢宰.
158 Im *DHYDCD* wird der Begriff schlicht als *dachen* 大臣 („Minister / hoher Beamter“) übersetzt. Siehe *DHYDCD*, 冢臣.
159 Suwald übersetzt hier mit „Loyalität und Fähigkeiten“ Suwald 2008, S. 209.
160 Yin Wen 尹文: *Yin Wen Zi* 尹文子. Hrsg. von Donald Sturgeon. URL: https://ctext.org/yin-wen-zi (besucht am 17. 02. 2020).
162 Vgl. *CBDB*, `text_data`.

rend der Tang-Zeit sprechen.[163] Um solche Schlussfolgerungen zu stützen bzw. zu entkräften, sind zusätzliche philologische, bzw. „sinologischere" Methoden erforderlich. Suwald betrachtet daher ausführlich den Inhalt des *Zhongjing*, die Verwendung tabuisierter Zeichen,[164] sowie Sui- (隨, 581–618) und Tang-zeitliche (唐, 618–907) Literaturkataloge – in denen das *Zhongjing* allerdings nicht aufgeführt wird.[165]

Durch die chronologische Darstellung der in einem Text enthaltenen Lexeme können temporale Textprofile Hinweise auf sprachliche Anachronismen liefern und so helfen, Fälschungen aufzudecken, die mit dem Ziel geschaffen wurden, „älter" zu erscheinen. Sowohl die Erwähnung von Ereignissen, die sich erst nach dem angeblichen Verfassen eines Werkes abgespielt haben, sowie die Verwendung von Wörtern, die erst später gebraucht wurden, sind Hinweise auf eine Fälschung, da ein Text in der Regel keine Lexeme enthalten kann, die neuer sind als der Text selbst. Die Tatsache, dass offensichtliche Anachronismen für Fälscher:innen relativ leicht zu vermeiden sind, spricht allerdings gegen ein erfolgreiches Entlarven von Fälschungen mit dieser Methodik. Es kann lohnender sein, auf die „most trivial details", die „involuntary signs"[166] zu achten. Dazu zählen bei Texten die Worthäufigkeiten bzw. die am häufigsten verwendeten Wörter, sowie die Häufigkeit der Verwendung bestimmter Worttypen.[167]

Je vollständiger und genauer die historische Lexemdatenbank ist, die den temporalen Textprofilen zugrunde liegt, desto besser können von der Software Anachronismen aufgespürt werden, während dies für menschliche Leser:innen nur mit Expert:innenwissen und akribischer Recherche machbar ist. Einzelne Wörter können so „clear dating implications"[168] mit sich bringen, jedoch sollte auch die Gesamtheit der Sprache eines Texts kritisch betrachtet werden, denn „viewed in isolation, an individual word generally cannot yield decisive dating implications."[169] Die umfassende Ergänzung der Zitate aus dem *DHYDCD* um frühere Belegstellen aus einem relativ kleinen Textkorpus verdeutlicht zudem, dass eine diachrone Lexemdatenbank niemals vollständig sein kann.[170]

Temporale Textprofile können beim Anfangsverdacht einer Fälschung helfen, weitere Indizien zu finden, die aber einzeln geprüft werden müssen. In vielen Fällen wird es sich schlicht um eine frühere Belegstelle für das gefundene Lexem, oder um eine semantisch abweichende Zeichenfolge handeln, die zuvor nicht in der Datenbank erfasst war.

Limitationen ergeben sich weiterhin durch die geringe Detailtiefe von 100 Jahren und vor allem durch die vor und während der Han-Zeit noch ungenaueren historischen Lexikalisierungsdaten. Für Probleme wie die Klärung der Autorschaft von Textteilen, z. B. der Kapitel 81 bis 120 des *Hong lou meng* 紅樓夢,[171] oder die Unterscheidung von Alttext- und Neutext-Versionen

163 Tatsächliche Song-zeitliche Lexeme konnten in dem Text nicht nachgewiesen werden. Begrenzt man die Analyse auf die Lexikologie, könnte es sich beim *Zhongjing* theoretisch immer noch um einen Han-zeitlichen Text handeln, der *Locus classicus* für die oben untersuchten Lexeme ist.

164 Siehe dazu auch Kapitel 4.3, S. 69.

165 Siehe Suwald 2008, S. 72–77. Auch das ist allenfalls ein Indiz, kein Beweis, dass der Text erst in der Song-Zeit entstanden ist.

166 Carlo Ginzburg 1989: *Clues, Myths, and the Historical Method*. Baltimore & London: Johns Hopkins University Press, S. 97, 118; zitiert in Allison et al. 2011, S. 24.

167 Siehe Allison et al. 2011, S. 2, S. 24. Allison et al. sprechen von *language action types* wie z. B. *FirstPerson*, für die von ihnen verwendete Software *DocuScape* definierte linguistische Kategorien.

168 Leonard Neidorf 2014: „Lexical Evidence for the Relative Chronology of Old English Poetry". In: *SELIM* 20, S. 7–48, S. 11.

169 Ebd., S. 35.

170 Siehe dazu Kapitel 5.5.4, ab S. 134.

171 Siehe z. B. Hu Xianfeng, Wang Yang und Wu Qiang 2014: „Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber". In: *Advances in Adaptive Data Analysis* 6. doi: 10.1142/S1793536914500125, S. 17–18. Die

des *Shangshu*, reichen die zur Erzeugung der Profile verfügbaren Daten nicht aus. Methoden aus der Stilo(chrono)metrie oder Verfahren wie die *PCA* sind zur Bearbeitung solcher Fragen besser geeignet – solange passende Vergleichstexte bzw. Trainingsdaten vorliegen.[172]

6.2.5 Automatisierte Datierung mit temporalen Textprofilen

In den Abschnitten 6.2.3 und 6.2.4 wurden einzelne temporale Textprofile mit sinologischem Vorwissen und unter Heranziehen zusätzlicher Quellen gedeutet. Sie können jedoch auch für eine automatisierte Schätzung des Zeitraums der Textgenese eingesetzt werden. In diesem Abschnitt erarbeite ich anhand eines Trainingsdatensatzes von 432 Texten aus dem *difangzhi* 地方誌 Korpus (*DFZ*)[173] eine Herangehensweise für die automatisierte Interpretation temporaler Textprofile. Der bereits in Kapitel 6.1 verwendete Testdatensatz aus 216 *DFZ* wird erneut zur Evaluierung verwendet. Die beschriebene Methodik wird anschließend zusätzlich mit Texten aus dem *XXSKQS*-Datensatz und den *zhengshi* 正史 erprobt.[174] In Tabelle 6.12 sind die Ergebnisse der durchgeführten Experimente zusammenfassend dargestellt.[175] Zur Einschätzung der Ergebnisse dienen auch hier die *Accuracy*, also der Anteil der dem korrekten Jahrhundert zugeordneten Texte und der *mean average error* (*MAE*) in Jahren. Der *mean error* D_{mean} für die Zuordnung eines Texts ist als Mittelwert der Differenz zwischen Anfang und Ende des datierten *chronon c* und dem in den Metadaten angegebenen Jahr der Veröffentlichung definiert. Bei einer Granularität der Datierung von 100 Jahren ergibt sich daraus ein Mindestwert von $D_{mean} = 50$ für einen korrekt datierten Text.[176]

Betrachtung von Lexemen

Im Optimalszenario einer vollständigen diachronen Lexemdatenbank und einer fehlerfreien Segmentierung des zu datierenden Texts ließe dieser sich in der Regel dem spätesten Jahrhundert zuordnen, aus welchem noch Lexem-*types* vorkommen. Im Hinblick auf die tatsächlichen frühesten Belegstellen aller Lexem-*types* eines unbekannten Texts muss aber von einer unvollständigen Datenbank ausgegangen werden. Zudem führt die vereinfachte *n*-Gramm-Segmentierung zu weiteren *false positives*. Texte können daher einen variablen Anteil an Lexemen enthalten, die später eingeordnet sind als der zu datierende Text. Dieser nimmt tendenziell mit dem Alter des zu datierenden Texts zu (Abb. 6.22).

Autoren argumentieren anhand einer *Support Vector Machine* (*SVM*)-Klassifizierung der einzelnen Kapitel, dass nicht nur die Kapitel 81–120, sondern wahrscheinlich auch Kapitel 67 nicht von Cao Xueqin 曹雪芹 (gest. 1763/4) stammen.

172 Siehe Kapitel 3.1, S. 40.
173 *DFZ*, siehe auch Kapitel 4.2, S. 66.
174 Siehe auch Kapitel 6.1, S. 175 u. 172.
175 Siehe S. 207.
176 Siehe Kapitel 6.1, S. 157.

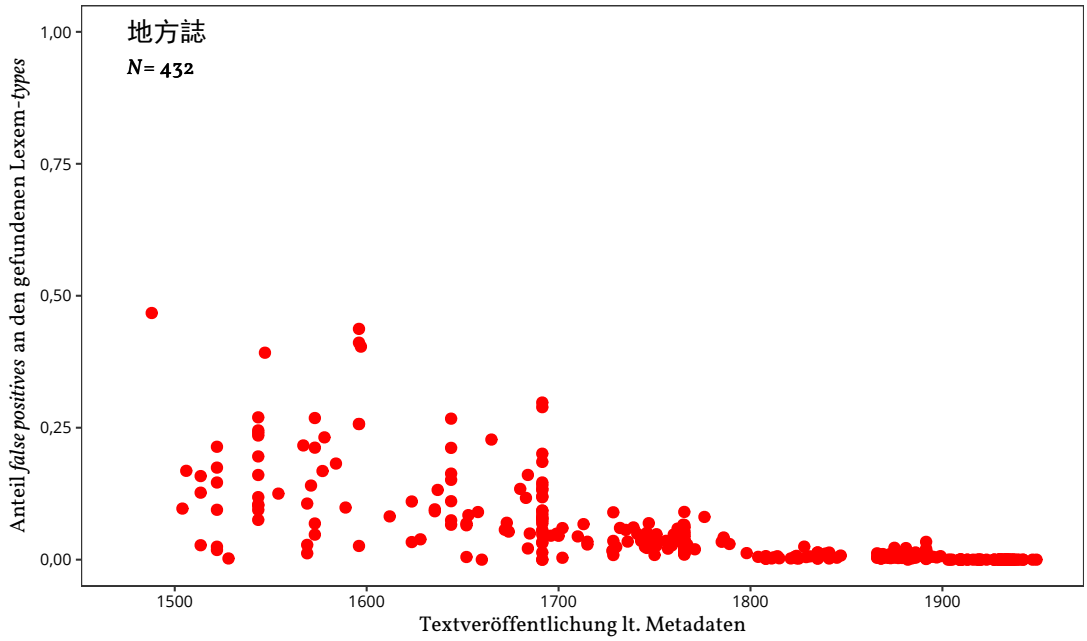


Abbildung 6.22 Anteil *false positives* (zu neu datierte Lexeme) nach Textveröffentlichung

Eine automatisierte Analyse von Profilen undatierter Texte kann sich daher nicht auf den erwarteten Anteil an *false positives* stützen. Zielführender ist es, den Anteil der Lexeme zu kalibrieren, der üblicherweise z. B. noch jeweils aus den Jahrhunderten vor und zur tatsächlichen Datierung festgestellt werden kann. Die Matrix in Abb. 6.23 zeigt anhand derselben Texte die Korrelationen zwischen der Gesamtzahl festgestellter Lexem-*types* und den *types*, die dem Jahrhundert der jeweiligen Textentstehung bzw. dem vorangegangenen Jahrhundert lexikographisch zugeordnet sind. Diese Berechnungen werden ohne, mit und mit *s*-Gewichtungskorrektur durchgeführt.

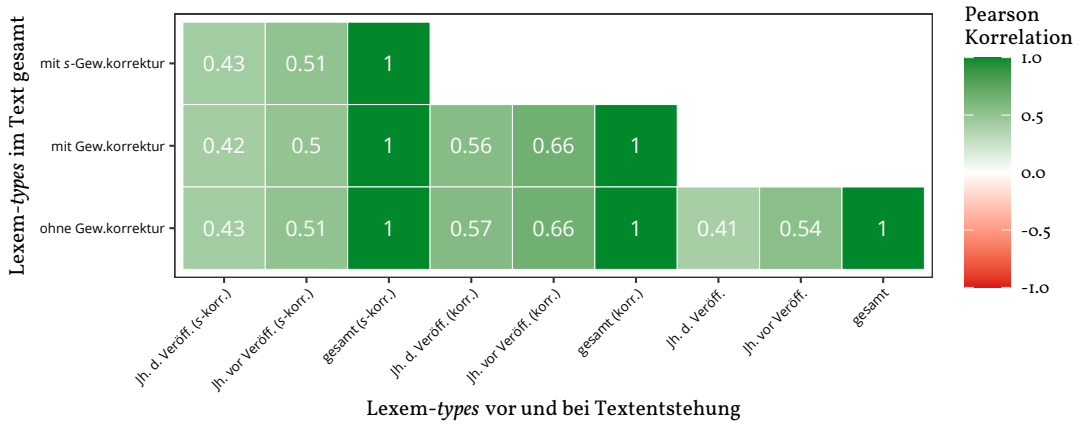


Abbildung 6.23 Korrelationsmatrix: *types* vor und zur Veröffentlichung und Gesamtanzahl *types*

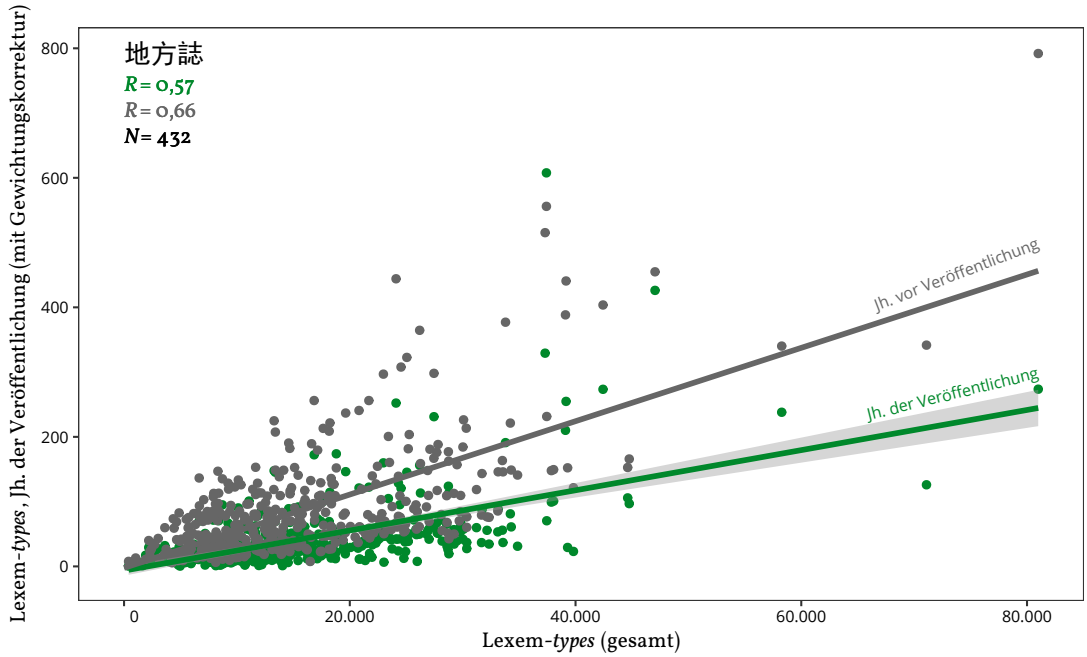


Abbildung 6.24 Korrelation Lexem-*types* zur und vor Veröffentlichung und Gesamtanzahl *types*

Die stärkste Korrelation R[177] von 0,66 ergibt sich dabei zunächst zwischen der Anzahl *types*, die bei Anwendung der linearen Korrektur dem Jahrhundert vor der Entstehung des Textes zugeordnet sind mit der Gesamtzahl der festgestellten *types*. Ohne bzw. mit *s*-Gewichtungskorrektur ist die Korrelation schwächer. Abb. 6.24 zeigt die Korrelation zwischen der Gesamtzahl der in den Texten enthaltenen 2–3-Zeichen Lexem-*types* T (*x*-Achse) und der Anzahl der *types*, die den Jahrhunderten vor und zur Veröffentlichung des jeweiligen Textes zugeordnet sind (*y*-Achse). An den unterschiedlichen Steigungen der Regressionsgeraden zeigt sich erneut, dass die pro Jahrhundert zugeordneten *types* in Richtung des Jahrhunderts der Textentstehung (grün) typischerweise abnehmend verlaufen.[178] Mittels der Funktion dieser Regressionsgeraden kann nun für einen Text mit T Lexem-*types* die für das Jahrhundert der Textentstehung erwartete Anzahl Lexem-*types* projiziert werden. Experimente mit den genannten Regressionsmodellen ergeben, dass ungeachtet der schwächeren Korrelation diese Herangehensweise zielführender ist, als die für das Jahrhundert vor der Entstehung des Textes erwarteten Lexeme zu projizieren.
Daraus resultiert folgende Herangehensweise zur automatisierten Datierung.

1. Die Anzahl der mit Gewichtungskorrektur für das Jahrhundert der Textentstehung c erwarteten *types* $t_{proj.}$ wird abhängig von der Gesamtzahl der gefundenen *types* T berechnet. Bei Verwendung der Lexemdatenbank mit zusätzlichen Korpusbelegen aus *zhengshi*,

177 Der Pearson-Korrelationskoeffizient R misst, wie gut die Messwerte zweier Merkmale in einem linearen Modell miteinander korrelieren.Dabei steht der Wert 1 oder –1 für eine perfekte Abhängigkeit zwischen den Merkmalen, ist der Wert 0, sind sie unkorreliert. Der Korrelationskoeffizient wird aus der Summe der quadrierten Standardabweichungen berechnet. Siehe z. B. Ludwig Fahrmeir et al. 2013: *Regression – Models, Methods and Applications*. Berlin & Heidelberg: Springer, S. 287.

178 Vgl. auch die Einzeldarstellungen in Abb. 6.18, S. 190 bzw. 6.25, S. 201.

Loewe und *DFZ*,[179] einer linearen Gewichtungskorrektur und bei Berücksichtigung von 2–3-Zeichen Lexem *types* ergibt sich:

$$t_{proj.} = 0,003 \times T - 6,513$$

Änderungen an der Betrachtung, also z. B. eine Erweiterung des *n*-Gramm-Raums auf 1–3- oder 2–4-Gramm-*types*,[180] sowie Anpassungen an oder Erweiterungen der Datenbank, erfordern eine Neuberechnung der Regression. Ohne Gewichtungskorrektur bzw. mit *s*-Gewichtungskorrektur ergeben sich ebenfalls eigene Werte für Steigung und Achsenabschnitt.

2. Aus dem Neologismusprofil mit Gewichtungskorrektur wird zunächst das Jahrhundert *c* mit der geringsten Differenz in der Menge zugeordneter *types* t_c zum errechneten Wert $t_{proj.}$ ausgewählt. Da wegen des negativen Achsenabschnitts bei Texten mit weniger als 2.000 *types* der Wert von $t_{proj.}$ unter 0 sinkt, wird als zusätzliche Bedingung ein Mindestwert von 5 festgesetzt. Andernfalls könnten einzelne *false* positives einen starken Einfluss auf das Datierungsergebnis nehmen.
3. Da die Profile mit (linearer) Korrektur der Gewichtung üblicherweise abnehmend verlaufen, werden anschließend die „benachbarten" Jahrhunderte betrachtet. Falls für die beiden vorangegangenen und nachfolgenden Jahrhunderte eine gegenläufige Tendenz zu beobachten ist, wird das Ergebnis *c* auf das späteste Jahrhundert korrigiert, für das $t_{proj.}$ bzw. 5 überschritten wird. Dies gilt also, wenn für eines der beiden vorherigen Jahrhunderte weniger, oder für eines der beiden späteren Jahrhunderte mehr Lexeme gemessen wurden als t_c.
4. Unter Berücksichtigung der bisherigen Erkenntnisse über die Möglichkeit von *Peaks* im Jahrhundert der Textentstehung, wird der Text überdies älter datiert, wenn für das vorangehende Jahrhundert ($c - 100$) deutlich mehr *types* nachgewiesen sind.[181]

Zur Veranschaulichung sei das temporale Profil eines Texts aus dem *DFZ*-Korpus gezeigt (Abb. 6.25). Die Veröffentlichung dieser *Guide fu zhi* 歸德府志 (*Chronik der Präfektur Guide*, heute Shangqiu 商丘, Henan 河南) wird mit 1754 angegeben.[182] Gemäß der Projektionsfunktion werden bei 17.643 im Datensatz festgestellten 2–3-Zeichen Lexem-*types* für das Jahrhundert der Entstehung des Textes 48,2 *types* erwartet. Der Wert mit der geringsten Differenz davon, 20,7, ist dem 18. Jh. zugeordnet, in dem der Text tatsächlich entstanden ist. Eine Umdatierung wegen eines gegen die Intuition verlaufenden Profils findet nicht statt. 82 *types* sind dem 17. Jh. zugeordnet, dem 19. Jh. nur 11,9. Ein *Peak* im 17. im Vergleich zum 18. Jh. besteht ebenfalls nicht.

Mit dem beschriebenen Algorithmus wird mit linearer Gewichtungskorrektur und 2–3-Gramm Lexem-*types*, der Testdatensatz aus 216 *Difangzhi* datiert. Für 47,2 % der Texte kann das Jahrhundert der Veröffentlichung bestimmt werden – bei einer durchschnittlichen Abweichung

179 Siehe Kapitel 5.5.4, ab S. 134. Verwendet man die zusätzlichen Belegstellen nicht, wird eine nahezu perfekte Korrelation mit $R^2 = 0,92$ zwischen Lexem-*types* (mit Gewichtungskorrektur) zum Jahrhundert der Veröffentlichung des Textes und der Gesamtzahl *types* erzielt – für Datierungszwecke ist dies allerdings dennoch nicht zuträglich, da kaum Diskrepanz zu „benachbarten" Jahrhunderten besteht.

180 Da nur 1–3-Gramme der *DFZ* vorliegen sind keine Experimente mit 2–4-Gramm-Daten möglich.

181 Bei den gewählten Parametern wird für „deutlich mehr" hier die vierfache Anzahl angenommen. Der beschriebene Algorithmus zum automatisierten „Lesen" der Neologismusprofile basiert auf der Betrachtung zahlreicher Profilverläufe und stellt lediglich eine von zahlreichen Möglichkeiten dar.

182 *DFZ*, # 0c87f43d7392c0589fbfb491e5165af9.

von 83,9 Jahren. Ohne Gewichtungskorrektur kann eine minimal bessere *Accuracy* von 48,6 % erreicht werden, der *MAE* erhöht sich jedoch auf 99,5 Jahre.[183]

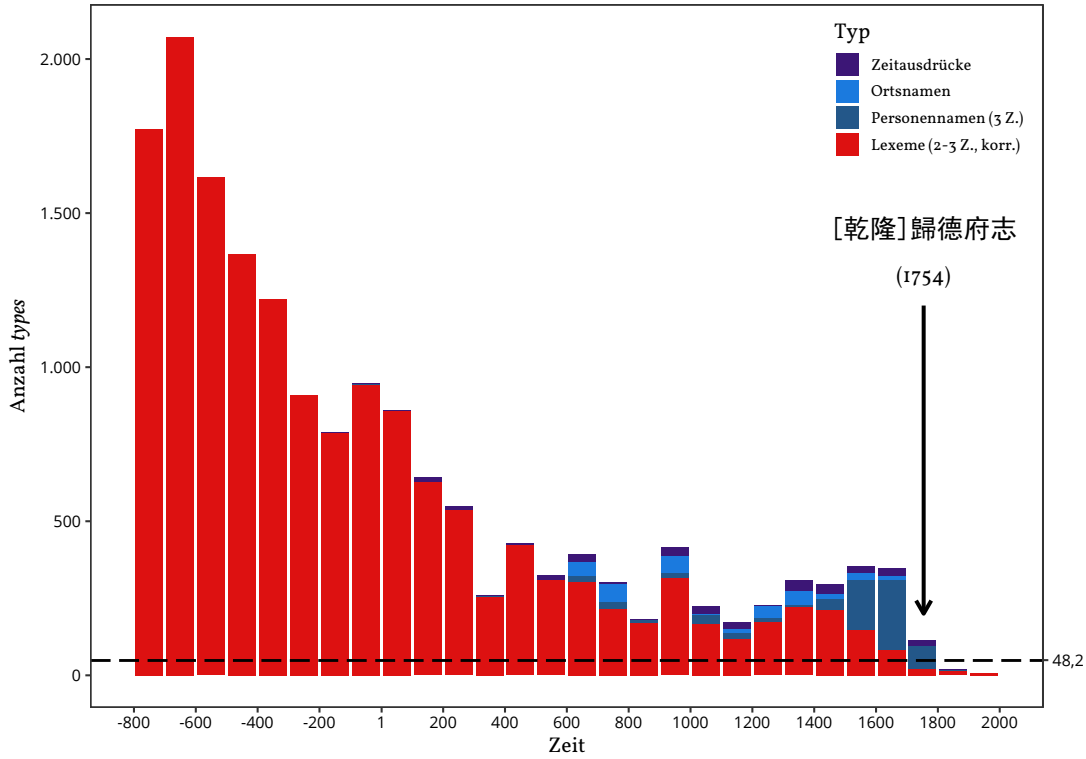


Abbildung 6.25 Temporales Profil des *Guide fu zhi* 歸德府志 von 1754

Berücksichtigung von Personennamen

Gerade in stilistisch alten Texten, die kaum zeitgenössisches Vokabular enthalten, können Personennamen wichtige Indizien für die Datierung liefern. Eine nur aufgrund der festgestellten Lexeme „zu alte" Datierung kann gegebenenfalls korrigiert werden, wenn Namen von Personen mit späteren biographischen Daten im Text genannt werden.[184] Zur Reduzierung von Ambiguitäten eignen sich dafür Namen mit einer Länge von mindestens drei Zeichen. Außerdem sollten diese zumindest in der *CBDB* eindeutig zugeordnet werden können, also nur eine einzige Person dieses Namens verzeichnet sein.[185] Trotz dieser Einschränkungen sind zwei Arten von *false positives* typisch. Personen gleichen Namens, über die kein Eintrag in der *CBDB* besteht, sowie Zeichenfolgen, die zufällig mit einem Namen übereinstimmen.[186] Ein Beispiel für ersteres aus dem *Guide fu zhi* ist Yang Zongji 楊宗稷, dessen biographische Daten in der *CBDB* mit 1865–1933

183 Siehe Abb. 6.27, S. 204; Tabelle 6.12, S. 207.
184 Aus der Nennung von Ortsnamen lassen sich bei der gegebenen Datenlage der *CBDB* kaum zuverlässige Erkenntnisse gewinnen.
185 Siehe Kapitel 4.7, ab S. 97.
186 Siehe auch Abschnitt 6.2.2, ab S. 189.

angegeben sind.[187] Im *Guide fu zhi* wird eine frühere Person desselben Namens als Teilnehmer an der Beamtenprüfung aufgelistet.[188]

Anders als bei Lexemen kann nicht davon ausgegangen, dass die Anzahl der jedem Jahrhundert zugeordneten Namen zum Jahrhundert der Textentstehung hin abnimmt. Texte können die Namen zahlreicher Zeitgenoss:innen nennen – gerade historiographische Texte können aber auch ausschließlich Personen aus früheren Jahrhunderten erwähnen.

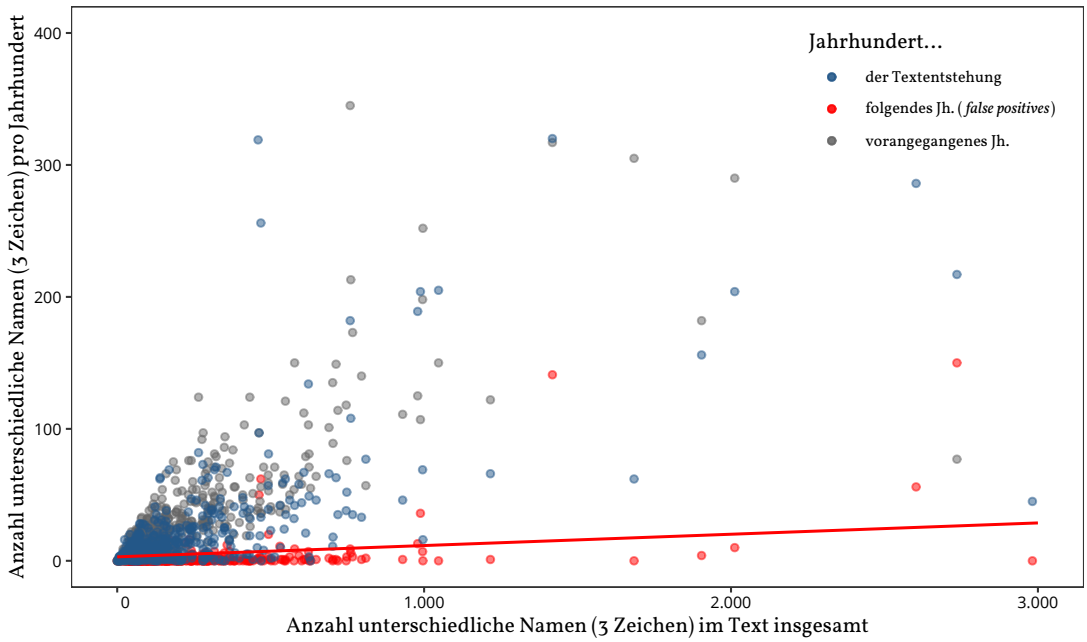


Abbildung 6.26 Namen in den Trainingsdaten

Abb. 6.26 zeigt die Anzahl unterschiedlicher 3-Zeichen Namen in den 432 Texten des Trainingsdatensatzes, die darin dem Jahrhundert der Veröffentlichung und den jeweils benachbarten Jahrhunderten zugeordnet sind in Relation zur Gesamtzahl der unterschiedlichen Namen *N* im Text. Insgesamt ist ein geringerer Anteil der Namen dem Jahrhundert der Textentstehung (blau) zugeordnet als dem vorangegangenen (grau). Dennoch lässt sich keine sinnvolle Abgrenzung vornehmen, da die Datenpunkte stark gestreut sind. Andererseits sind die Werte der *false positives* aus dem folgenden Jahrhundert (rot) deutlich niedriger. Im Gegensatz zu den Lexem-*types* sollte hier zudem kein Zusammenhang zwischen *false positives* und der Datierung des Textes (Abb. 6.22) bestehen.[189] Da unabhängig davon längere Texte potenziell eine größere Anzahl an *false positives* enthalten können, sollte ein Schwellenwert für die Anzahl von Namen, der eine Späterdatierung rechtfertigt, von der Textlänge abhängig gemacht werden. Eine Linearregression der für das Jahrhundert nach der Textentstehung festgestellten *false positives* auf *N* liefert zwar

187 *CBDB, ID* 76819. Die in der *CBDB* gemeinte Person war ein erfolgreicher *guqin* 古琴-Musiker.

188 Siehe Chen Yanglu 陳錫輅 und Cha Qichang 查岐昌, Hrsg. 2016 [1754]: *[Qianlong] Guide fu zhi 36 juan [乾隆] 歸德府志 36 卷 ([Qianlong] Chronik der Präfektur Guide, 36 juan)*. Online-Datenbank Diaolong 雕龍 / *Zhongguo Difang zhi* 中國地方誌, via CrossAsia. Nagoya 名古屋 & Taipeh 台北: Kaixi MS 日本凱希多媒體 & tts 大鐸資訊, S. 113–115.

189 Vgl. auch Kapitel 4.7, S. 102.

eine sehr schwache Korrelation, die Steigung kann aber zur textspezifischen Festsetzung eines Schwellenwerts n_t genutzt werden. Anstatt des Achsenabschnitts der Regression (0,25) wird ein Mindestschwellenwert n_θ von 3 festgelegt, um willkürliche Späterdatierungen zu minimieren. Aus den Trainingsdaten ergibt sich damit die Funktion:

$$n_t = 0,0086 \times N + 3$$

Auf dieser Basis kann anstelle des zuvor bestimmten Jahrhunderts c das späteste Jahrhundert als Zeitstempel angenommen werden, für das die Anzahl zugeordneter Namen n_t überschreitet und die Mindestanzahl von 5 Lexem-*types* noch erreicht wird.

Zur Veranschaulichung sei erneut das *Guide fu zhi* herangezogen (Abb. 6.25). Die 2–3-Gramme des Textes weisen 648 Übereinstimmungen mit eindeutigen 3-Zeichen Namen aus der *CBDB* auf. Der Wert von n_t liegt also bei 8,6. 74 Namen sind dem 18. Jh. zugeordnet – dem vorher bereits vergebenen, korrekten Zeitstempel. 6 Namen sind dem 19. Jh. zugeordnet – in diesem Fall kommt es also nicht zur Vergabe eines späteren Zeitstempels. Sowohl dem 19., als auch dem 20. Jh. sind mehr als 5 Lexem-*types* zugeordnet. Wären einem der beiden Jahrhunderte also 9 oder mehr *false positive* Namen zugeordnet, käme es zu einer falschen Späterdatierung.

Mit der oben beschriebenen Vorgehensweise kann anhand der zusätzlichen *NER*-Informationen ein höherer Anteil der 216 *DFZ* (62,5 %) bei einem *MAE* von 72,1 Jahren dem Jahrhundert der Veröffentlichung zugeordnet werden (Abb. 6.27; Tabelle 6.12). Voraussetzung für den Erfolg dieser Herangehensweise bleibt die Erwähnung von Zeitgenoss:innen bzw. ein geringer zeitlicher Abstand zwischen erzählter Zeit und dem Verfassen des Textes. Die Auflistung z. B. von Teilnehmern an lokalen Beamtenprüfungen in einigen *DFZ* schafft dafür eine gute Ausgangssituation, die für andere Textgattungen so nicht erwartet werden kann.

Betrachtung von *temporal expressions*

In Texten erkannte *temporal expressions* können – falls vorhanden – ebenfalls für die zeitliche Einordnung von Texten genutzt werden. Je vollständiger ein solcher Ausdruck ist, desto zuverlässiger verweist er eindeutig auf ein Jahr bzw. sogar ein bestimmtes Datum.[190] In einem Datensatz mit 1–3-Gramm-Häufigkeiten ist die Erkennung solcher Ausdrücke, die typischerweise 4–12 Zeichen lang sind, stark eingeschränkt. Möglich ist aber die Erkennung von Regierungsdevisen (meist zwei Zeichen), gefolgt von einer Ziffer. Bei Eindeutigkeit der Bezeichnung einer so erkannten Regierungsdevise kann ein solcher Ausdruck von drei Zeichen einem Jahr zugeordnet werden.[191] Um Falschzuordnungen durch auftretende *false positives* zu begrenzen, wird ein Schwellenwert von $t_\theta = 4$ unterschiedlichen Vorkommen (*types*) von Zeitausdrücken festgelegt.[192]

In einem Test mit den 216 *DFZ* können 88 % der Texte bei einem *MAE* von 57,5 Jahren korrekt zugeordnet werden, indem sie jeweils dem spätesten Zeitraum mit mindestens vier entsprechenden *temporal expressions* zugeordnet werden.[193] 4,6 % der Texte werden aufgrund von *false*

190 Zur Erkennung von *temporal expressions* in schriftsprachlichen Texten siehe Kapitel 4.8, ab S. 103.
191 Siehe Kapitel 4.8, ab S. 103.
192 t_θ wurde auf Grundlage der Trainingsdaten optimiert. 3,5 % der Texte enthalten 4 oder mehr *temporal expression false positives*, die dem Jahrhundert nach der Veröffentlichung zugeordnet sind. Mit einem niedrigeren Schwellenwert würden mehr Texte fälschlich später datiert, bei einem höheren Wert von t_θ wiederum deutlich weniger Texte noch dem Jahrhundert der Veröffentlichung zugeordnet.
193 Auch hier können *false positives* auftreten, vgl. auch Kapitel 4.8, S. 103.

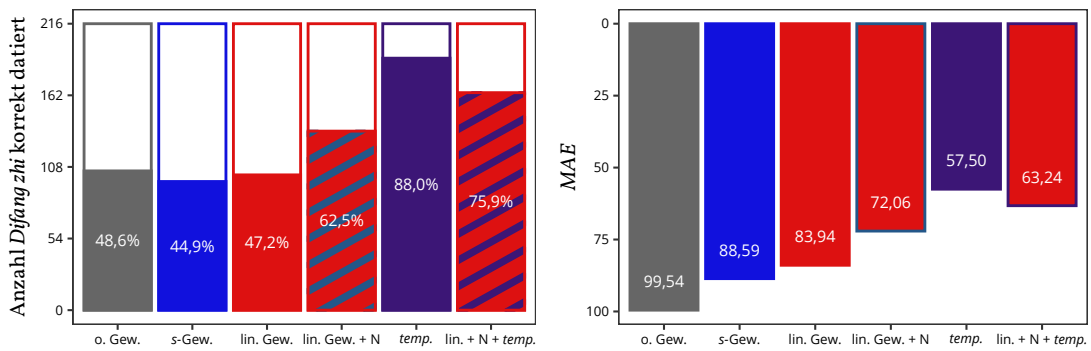
positives zu spät, 6,5 % zu früh eingeordnet. Zwei Texte (0,9 %) enthalten zu wenige *temporal expressions*.

Im Beispiel des *Guide fu zhi* (Abb. 6.25, S. 201) werden 20 unterschiedliche Jahresangaben erkannt, die dem 18. Jh. zugeordnet sind und in die Regierungszeiten der Kaiser Yongzheng 雍正 (reg. 1709–1722) und Qianlong 乾隆 (reg. 1733–1796) fallen. Dem 19. Jh. ist ein *false positive* zugeordnet, dem 20. Jh. keines.

Für den *DFZ*-Datensatz ist diese primitive Herangehensweise den bisher betrachteten überlegen, da sie als historiographische Texte einen hohen Anteil an *temporal expressions* enthalten. Die sehr hohe *Accuracy* hängt auch damit zusammen, dass Texte mit einem Abstand von mehr als 50 Jahren zwischen Veröffentlichung und erzählter Zeit von Trainings- und Testdaten ausgeschlossen sind, um störende Effekte durch spätere Editionen eigentlich älterer Texte zu vermeiden.[194] Da durch die Angabe der Regierungsdevise – anders als bei westlichen Texten – nur ein Zeitpunkt in der Vergangenheit, der Gegenwart oder der nahen Zukunft angegeben werden kann, eignet sich diese Herangehensweise aber grundsätzlich auch zur Datierung von anderen Textsorten. Dabei ist – ähnlich wie bei Personennamen – eine kritische Überprüfung der erkannten *temporal expressions* erforderlich.

Anstelle einer bloßen Betrachtung von *temporal expressions* können diese natürlich auch ergänzend zur Zuordnung auf Grundlage von Lexemen und Namen betrachtet werden. Dabei werden Texte später datiert, wenn einem späteren als dem bisher datierten Jahrhundert 4 oder mehr *temporal expressions* zugeordnet sind. Eine zu späte Datierung auf Basis von Namen oder Lexemen wird dabei nicht korrigiert. Andernfalls würden Texte primär auf die Zeit datiert, über die darin berichtet wird. Für den *DFZ*-Testdatensatz fallen die Ergebnisse mit einer *Accuracy* von 75,9 und einem *MAE* von 63,2 Jahren dann etwas schlechter aus als bei reiner Betrachtung temporaler Ausdrücke.

Ergebnisse



(a) Anteil richtig datierter Texte

(b) Durchschnittliche Abweichung in Jahren *MAE*

Abbildung 6.27 Performance profilbasierter Datierung, *Difangzhi*, 2–3 Zeichen-Lexeme

194 Siehe dazu Abschnitt 6.1.1, S. 158.

In Abb. 6.27 werden die Ergebnisse der beschriebenen Profildatierungen gegenübergestellt. Bei einer reinen Betrachtung von 2–3-Zeichen Lexemen mit und ohne Gewichtungskorrektur, sowie mit der in Abschnitt 6.2.1 eingeführten *s*-Gewichtungskorrektur zeigt sich, dass ein Ausgleich des *HYDCD-Bias* sich grundsätzlich positiv auf den *MAE* auswirkt. Die stark vereinfachende Annahme, dass der Wortschatz in jedem Jahrhundert gleich stark wächst (lineare Gewichtungskorrektur) führt dabei zu den besseren Ergebnissen.[195] Bei ergänzender Verwendung von Personennamen weicht die vorausgesagte Veröffentlichung dabei nur bei 6,4 % der untersuchten *Difangzhi* um mehr als ein Jahrhundert ab, die maximale Abweichung beträgt 277 Jahre. Diese Ergebnisse sind – ohne jede Berücksichtigung von Worthäufigkeiten – durchaus vergleichbar mit denen bei Verwendung von genrespezifischen statistischen Sprachmodellen, obwohl hier ein deutlich längerer Datierungszeitraum von 700 v. u. Z. bis ins 20. Jh. berücksichtigt wird.[196] Bei reiner Betrachtung von *temporal expressions* können die mit Abstand besten Ergebnisse erzielt werden. Sie übertreffen für diese besondere Textgattung sogar eine kombinierte Betrachtung von Lexemen, Namen und Zeitausdrücken.

Im vergleichbaren Ergebnis bei Verwendung eines statistischen Sprachmodells mit *NLLR* werden 59,7 % der Texte korrekt datiert. Der maximale Fehler liegt dann bei 306 Jahren, bei 7,4 % der Texte ist die Abweichung über 100 Jahre. Werden zusätzlich *temporal expressions* mit *NLLR*TE* betrachtet, können 64,4 % der Texte korrekt datiert werden.[197] Da bei der statistischen Datierung der *DFZ* mit einer *chronon*-Dauer von 50 Jahren gearbeitet wird, beträgt der minimale Fehler einer korrekten Datierung E_{min} bei einer korrekten Datierung nur 25 Jahre. Der *MAE* ist daher mit 41,5 bzw. 40,3 Jahren deutlich kleiner.

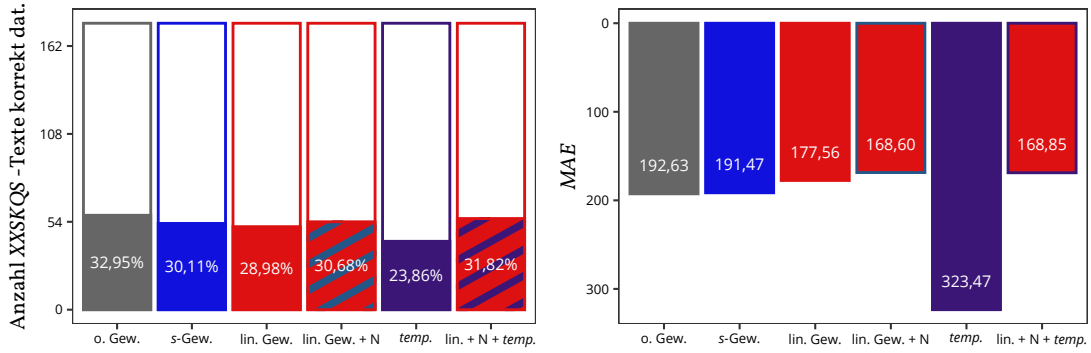
Abgesehen von der Zugänglichkeit für eine philologische Interpretation brauchen temporale Textprofile auch für die automatisierte Datierung den Vergleich mit statistischen Sprachmodellen nicht zu scheuen. Die Regression auf die Anzahl der für das Jahrhundert der Entstehung erwartbaren Lexem-*types* profitiert allerdings ebenfalls von spezifischen Trainingsdaten.

Experimente mit weiteren Korpora

Experimente mit weiteren Testdatensätzen sollen die Eignung der oben beschriebene Methodik für einen erweiterten Testzeitraum und andere Textgenres prüfen. Die Zuordnung von 176 Texten aus den *Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*)[198] zeigt, dass Texte anhand temporaler Textprofile auch dann ungefähr chronologisch eingeordnet werden können, wenn kein passendes Trainingskorpus verwendet wird (Abb. 6.28).[199] Durch die kombinierte Betrachtung von Lexemen, Namen und Zeitausdrücken kann noch eine *Accuracy* von 31,8 erreicht werden – beinah so viel, wie mit einem spezifisch trainierten statistischen Sprachmodell.[200] Der *MAE* liegt dabei zwar mit 168 Jahren deutlich höher, es wird hier aber innerhalb eines deutlich längeren Zeitraums und doppelter *chronon*-Länge datiert.[201] Aufgrund des Stils und vor allem Inhalts einzelner Texte des *XXSKQS*-Korpus kommt es zu Falschdatierungen mit einer Abweichung von bis

195 Erkenntnisse aus der Sprachwandelforschung und die aus dem *HYDCD* extrahierten Daten sprechen allerdings eher für ein *s*-förmiges Wortschatzwachstum, das dem Piotrowski-Gesetz folgt. Siehe dazu Kapitel 2.1 (ab S. 14) und 5.7.2 (ab S. 142).
196 Vgl. Kapitel 6.1, ab S. 156. Die verglichenen *SLM* umfassen den Zeitraum von 1475–1925.
197 Siehe Kapitel 6.1.1, Tabelle 6.1, S. 165, Beobachtungen #4 und #9.
198 Siehe S. 171.
199 Siehe S. 207.
200 Siehe Abschnitt 6.1.2, Tabelle 6.4, S. 171. Mit *SLM*s wurde mit *CS* und *tf-idf* eine *Accuracy* von 33,5 % erreicht.
201 Mit Sprachmodellen konnte mit demselben Datensatz ein *MAE* von 81 Jahren erzielt werden, wobei nur in *chronons* zwischen 1475 und 1925 klassifiziert wurde. Siehe S. 171.

zu 1.773 Jahren. So sind beispielsweise in dem laut Metadaten im Jahr 1823 veröffentlichte *Kaifang shuo* 開方說 des Mathematikers Li Rui 李銳 (1769–1817) keine Lexeme enthalten, die erst nach dem 2. Jh. nachgewiesen sind.[202]

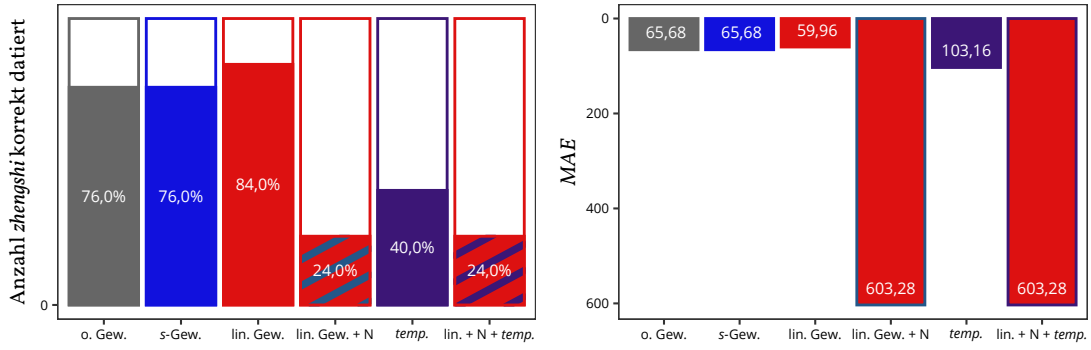


(a) Anteil richtig datierter Texte **(b)** Durchschnittliche Abweichung in Jahren *MAE*

Abbildung 6.28 Performance profilbasierter Datierung, *XXSKQS*, 2–3 Zeichen-Lexeme

Bei alleiniger Betrachtung von *temporal expressions* können nur 23,9 % der Texte dem richtigen Jahrhundert zugeordnet werden, da die untersuchten Texte mehrheitlich keine rezenten oder gar keine Zeitangaben enthalten.

Ein anderes Bild ergibt die Anwendung der Profildatierung auf die Dynastiegeschichten (Abb. 6.29). Dieses Korpus aus nur 25 Texten deckt einen Zeitraum von 91 v. u. Z. bis 1928 ab. Anders als die Datensätze der *DFZ*- und des *XXSKQS* stehen sie als Volltext zur Verfügung, was eine genauere Erkennung von *temporal expressions* ermöglicht.[203]



(a) Anteil richtig datierter Texte **(b)** Durchschnittliche Abweichung in Jahren *MAE*

Abbildung 6.29 Performance Profildatierung, *zhengshi*, 2–3 Zeichen-Lexeme / *temporal expressions*

202 Vgl. Li Rui 李銳 und Li Yingnan 黎应南 2000 [1823]: *Kaifang shuo* 開方說. Online-Datenbank Diaolong 雕龍 / *Xuxiu Siku quan shu* 續修四庫全書, via CrossAsia. Nagoya 名古屋 & Taipeh 台北: Kaixi MS 日本凱希多媒體 & tts 大鐸資訊, Weitere Beispiele für qing-zeitliche Texte, die keine oder kaum zeitgenössische Zeichenkombinationen enthalten, werden weiter unten ab S. 209, sowie in Abschnitt 6.1.3, ab S. 174 diskutiert.

203 Siehe dazu auch Kapitel 2.3, ab S. 20, sowie Kapitel 4.2, S. 65.

Bei der Betrachtung von Lexemen werden mit und ohne Gewichtungskorrektur sehr gute Ergebnisse erzielt, die bei Verwendung der linearen Korrektur am besten sind. Dies ist allerdings nicht repräsentativ, denn durch die intensive Nutzung dieses Korpus bei der Kompilation des *HYDCD* und das zusätzlich damit durchgeführte Training der Lexemdatenbank werden hier Optimalbedingungen ermöglicht,[204] die einer Identität von Trainings- und Testdatensatz ähneln.

Durch die Berücksichtigung von Namen werden die Ergebnisse deutlich verschlechtert. 76 % der Texte werden aufgrund von *false positives*, also Zeichensequenzen, die zufällig identisch mit Namen sind, sowie späteren Personen gleichen Namens, zu spät datiert.[205] Auch die Verwendung von *temporal expressions* ist weniger erfolgreich als bei den *DFZ*. Da ein größerer Zeitraum zwischen erzählter Zeit und Kompilation der Texte liegen kann,[206] werden die Texte häufig zu früh datiert. Die *Liao shi* 遼史 wurde z. B. 1343 fertiggestellt, die späteste darin erkannte *temporal expression* ist aber dem 11. Jh. zugeordnet.[207] Eine wegen *false positives* zu späte Datierung ist unwahrscheinlich, aber ebenfalls möglich. [208]

Zusammenfassung und Einschränkungen

Tabelle 6.12 Ergebnisüberblick der Datierungsexperimente aus 6.2.5

Korpus	Methode	*A* (%)	*MAE* (Jahre)	E_{max} (Jahre)	zu alt dat. (%)	zu neu (%)
DFZ	ohne Gewichtungskorrektur	48,6	99,5	2.815	38	13,4
	s-Gewichtungskorrektur	44,9	88,6	315,5	43,5	11,6
	lineare Gewichtungskorrektur	47,2	83,9	277	33,8	19
	+ Namen	62,5	72,1	277	17,6	19,9
	4+ *temporal expressions*	88	57,5	481	7,4	4,6
	kombiniert (linear + Namen + *temp.*)	75,9	63,2	277	1,9	2,2
XXSKQS	ohne Gewichtungskorrektur	33	192,6	2.890	38,6	28,4
	s-Gewichtungskorrektur	30,1	191,5	1.346	42	27,8
	lineare Gewichtungskorrektur	29	177,6	1.773	32,4	38,6
	+ Namen	30,7	168,6	1.773	29,5	39,8
	4+ *temporal expressions*	23,9	323,5	1.938	72,7	3,4
	kombiniert (linear + Namen + *temp.*)	31,8	168,9	1.773	28,4	39,8
zhengshi	ohne Gewichtungskorrektur	76	65,7	205	16	8
	s-Gewichtungskorrektur	76	65,7	205	16	8
	lineare Gewichtungskorrektur	84	60	211	12	4
	+ Namen	24	603,3	1.540	0	76
	4+ *temporal expressions*	40	103,2	340	56	4
	kombiniert (linear + Namen + *temp.*)	24	603	1.540	0	76

Die Ergebnisse der Experimente mit automatischer Datierung auf der Grundlage von temporalen Profilen werden in Tabelle 6.12 zusammengefasst. Die mit einem Datensatz von 432 *DFZ*

204 Siehe Kapitel 5.7.4, S. 150; siehe auch 5.5.4, S. 134.
205 Siehe dazu auch Kapitel 4.7, ab S. 97, sowie Abschnitt 6.2.2, ab S. 189.
206 Bei Auswahl der Texte aus dem *DFZ*-Korpus wurden Texte mit einem Abstand von über 50 Jahren zwischen Veröffentlichung und erzählter Zeit ausgeschlossen, um die Aufnahme späterer Auflagen zu minimieren.
207 Die Liao 遼 herrschten von 916–1125.
208 Das *Han shu* 漢書 enthält einige Zeitangaben mit den Äranamen *jianshi* 建始 (32–28 v. u. Z.). In der *DDBC* ist *jianshi* nur für die spätere Yan (*Hou Yan* 後燕, 384–409), eines der Sechzehn Reiche, verzeichnet. Das *Han shu* wird daher auf das 5. statt auf das 2. Jh. datiert.

trainierte automatisierte Analyse der temporalen Profile funktioniert grundsätzlich für alle drei Testkorpora zur ungefähren zeitlichen Einordnung der Texte. Erwartungsgemäß können die *zhengshi*, die bereits zur Erweiterung der zugrundliegenden Lexemdatenbank analysiert wurden, am besten zugeordnet werden. Für den *DFZ*-Testdatensatz können ebenfalls gute Ergebnisse erzielt werden. Selbst von den sehr heterogenen *XXSKQS*-Testdaten kann noch knapp ein Drittel korrekt zugeordnet werden, obwohl weder die Lexemdatenbank, noch der Profildatierungsalgorithmus mit diesem Datensatz trainiert wurden.

Werden nur Lexeme analysiert, zeigt sich anhand der *DFZ* und *XXSKQS*, dass Texte tendenziell eher zu alt als zu neu datiert werden. Durch Berücksichtigung von Namen lässt sich diese Tendenz ausgleichen. Eine Späterdatierung auf Basis von erkannten Personennamen kann jedoch nur funktionieren, wenn Namen von Zeitgenoss:innen der Verfasser:innen darin erwähnt werden. Bei den *zhengshi* führt das zu einer Überkompensierung und einer deutlich verschlechterten Performance, da in großer Zahl *false positives* auftreten. Angesichts der für das 1. Jahrtausend in der *CBDB* nur spärlich vorhandenen biographischen Daten und der fehlenden Möglichkeit einer zuverlässigen Tokenisierung bzw. *NER*, bedürfte die Betrachtung von Namen grundsätzlich einer Anpassung auf Spezifika des zu datierenden Textgenres.

Die Verwendung von *temporal expressions* erweist sich diesbezüglich als deutlich robuster – allerdings nur solange entsprechende Ausdrücke überhaupt vorkommen.

Mithilfe der Gewichtungskorrektur können für die *DFZ* extreme Abweichungen der Zeitstempel von der tatsächlichen Datierung vollständig verhindert werden. Bei einzelnen Texten aus dem *XXSKQS*-Datensatz kommt es allerdings auch damit noch zu massiven Fehldatierungen. Die Ursache hierfür sind Texte, die keine oder nur sehr wenige zeitgenössische Lexeme enthalten. Dies ist besonders problematisch, wenn ein Text weder zeitgenössische Personen, noch rezente Ereignisse referenziert. Durch die Beschränkung der Analyse auf Lexem-*types* mit 2–3 Zeichen und ein-eindeutige 3-Zeichen Namen aus der *CBDB* werden die verfügbaren Informationen zusätzlich reduziert.

Wie sehr sich solche schriftsprachlichen Texte allen Bemühungen um eine computerlinguistische Datierung entziehen können, sei am Beispiel des Qing-zeitlichen *Yuan shan* 原善 aus den *XXSKQS*-Testdaten veranschaulicht.

Die in den *XXSKQS* enthaltene Ausgabe dieses philosophischen Texts von Dai Zhen 戴震 (1724–1777) ist in den Metadaten auf das Jahr 1796 datiert.[209] Mit der oben beschriebenen Methodik würde er bestenfalls auf das 6. Jh. datiert, also um etwa 1.200 Jahre zu früh. Das temporale Profil des Textes (Abb. 6.30, ohne Gewichtungskorrektur) zeigt, wie diese Fehleinschätzung zustande kommt: Es sind keine zeitgenössischen Lexeme nachweisbar. Zudem werden im Text weder Namen von in der *CBDB* verzeichneten Personen genannt, noch lassen sich in den 2–3-Grammen des Textes *temporal expressions* feststellen. Auch die „neuesten" Lexem-*types* im Text (*hongju* 閎鉅, *cuanjue* 竄絕, *mingmei* 明昧 usw.) sind bereits in Ming-zeitlichen Texten nachgewiesen – ihre Anzahl ist aber so gering, dass es sich – ohne weitere Prüfungen – eben auch um in der Datenbank zu spät belegte Einträge handeln könnte. Dass das Profil bereits ab der Han-Zeit stark abflacht ist für einen Text aus dem 18. Jh. eher ungewöhnlich. Eine Ursache dafür ist Dai Zhens Argumentationsweise, der sich zur Darlegung seiner konfuzianisch geprägten

209 *XXSKQS*, # 9a0b80531e87b3dbfa56dfa1f5e7c3e4. Eine gedruckte Fassung des Textes existierte aber schon mindestens 1777. Siehe Cheng Chung-ying 成中英 2019 [1971]: *Tai Chen's Inquiry into Goodness: A Translation of the Yuan Shan, With an Introductory Essay*. Honolulu: University of Hawai'i Press, S. 50.

Standpunkte zahlreichen Zitaten aus Texten der klassischen Periode bedient, u. a. *Yijing* 易經, *Mengzi* 孟子 und *Zuo zhuan* 左傳.[210]

Auch die eigenen Textpassagen schreibt Dai allerdings in einer klassischen Sprache, die mit den statistischen Modellen aus Kapitel 6.1 sogar dem 4. Jh. *v. u. Z.* zugerechnet wird.[211] Damit dürfte der Text gegen jeden Versuch einer rein linguistischen Datierung quasi resistent sein.

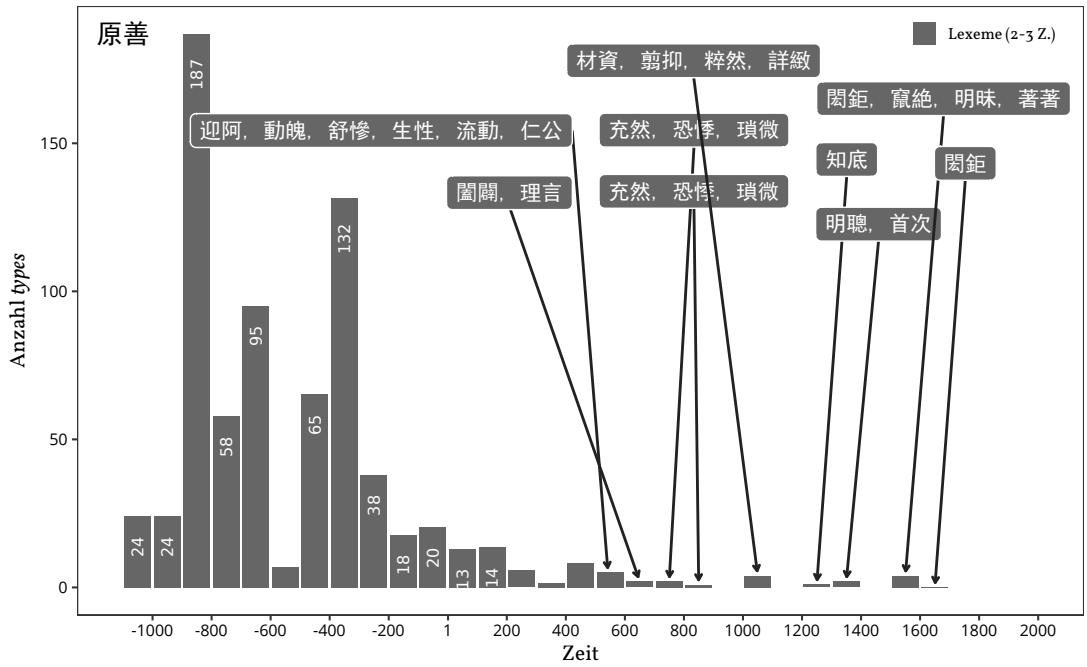


Abbildung 6.30 Temporales Profil des *Yuan shan* 原善

Auch wenn *Yuan shan* durch das Fehlen von Namen und temporalen Ausdrücken sicherlich ein Extrembeispiel darstellt, darf nicht vergessen werden, dass die Machart dieses Textes für die späte Kaiserzeit keineswegs außergewöhnlich ist. Das Beispiel erinnert an eine wichtige Limitation jeder linguistischen Textdatierung: „[T]he absence of any linguistic phenomenon in a book […] proves precious little."[212] Dass Harbsmeier diese Überlegung im Kontext der Datierung des *Lunyu* 論語 formuliert, deutet an, dass diese Problematik sich nicht auf die Datierung einiger spätkaiserzeitlicher Texte beschränkt, sondern für die gesamte Texttradition Relevanz hat.

Ob eine automatisierte Datierung anhand temporaler Profile erfolgreich sein kann, hängt stark von Inhalt und Stil des zu datierenden Textes ab. Das gilt ebenso für die Entscheidung, ob Lexeme, Namen und/oder temporale Ausdrücke betrachtet werden sollten. Unabhängig davon erlaubt die graphische Darstellung – anders als bei rein statistischen Methoden – immer noch eine den individuellen Besonderheiten eines Texts angepasste Interpretation.

210 Für eine ausführliche Darstellung siehe Cheng Chung-ying 成中英 2019 [1971], S. 50–51.
211 Siehe Abb. 6.5, S. 176. Datiert mit *NLLR* u. *NLLR*TE*, Modelle trainiert mit dem Zitatmaterial aus dem *DHYDCD*, vgl. S. 175.
212 Harbsmeier 2019, S. 207.

6.3 Datierung mit dem „durchschnittlichen Lexemalter“ von Texten

„History doesn't proceed in a linear way.“[213]
Francis Fukuyama

Mit den in Kapitel 6.1 und 6.2 vorgestellten Methoden werden Texte aufgrund ihrer Worthäufigkeiten bzw. der nachgewiesenen Lexeme stets vordefinierten Zeiträumen bzw. *chronons* zugeordnet. Intuitiv wird Zeit aber nicht als *kategorisch*, sondern als *kontinuierlich* wahrgenommen.[214]

Mit der auch in Kapitel 6.2 eingesetzten Datenbank lässt sich – wie bereits beschrieben – allen lexikalisierten Zeichenkombinationen eines Textes das (mittlere) Jahr ihrer ältesten Belegstelle zuordnen.[215] Daraus kann eine abstrahierte, absolute chronologische Messgröße für den Text berechnet werden – eine Art „durchschnittliches Wortentstehungsjahr“, im Folgenden durchschnittliches Jahr der Lexikalisierung, bzw. *Average Year of Lexicalization* (*AYL*). Ähnlich wie bei Überlegungen aus der Stylochronometrie, mit der Texte einer Autorin oder eines Autors per Regression auf sprachliche Merkmale datiert werden können,[216] soll untersucht werden, ob und wie mithilfe eines solchen *score* die Entstehungszeit von Texten auch als kontinuierliche Variable berechnet werden kann.

Das *AYL* als durchschnittliche mittlere Datierung *Y* der frühesten Belegstellen von *n* zu bewertenden Lexem-*types w* sei definiert als:

$$AYL = \frac{\sum_{w_1}^{w_n} Y}{n}$$

Zur Veranschaulichung wird die Berechnung des *AYL* für den Satz „昔者莊周夢為蝴蝶。“[217] erläutert. Enthalten sind darin 18 unterschiedliche 2–4-Gramm *types*, von denen drei im *DHYDCD* lexikalisiert und belegt sind:[218]

1. *hudie* 蝴蝶 – Die älteste angegebene Belegstelle stammt überraschenderweise aus einem Tang-zeitlichen Text, *Shi lin ji shi* 士林紀實.[219] Aus der *CBDB* sind die Lebensdaten des Autors Han Wo 韩偓 (842–914)[220] bekannt (siehe dazu auch 5.5.3, 132), so dass das Jahr 878 verwendet wird.[221]

213 Stephen Moss 2011: *Francis Fukuyama: 'Americans are not very good at nation-building'*. URL: https://www.theguardian.com/books/2011/may/23/francis-fukuyama-americans-not-good-nation-building (besucht am 15. 12. 2021).
214 Siehe auch Kapitel 3.3, S. 50.
215 Vgl. auch Kapitel 6.2, ab S. 179.
216 Siehe dazu Kapitel 3.1, S. 41.
217 „Einst träumte Zhuang Zhou, er sei ein Schmetterling.“ Zhuang Zhou 莊周 o. J. [ca. 3. Jh. v. u. Z.] *Zhuangzi* 莊子. *Digitale Ausgabe*. *Guoxue jingdian shuku* 國學經典書庫. Dongyang ligong daxue tushuguan 東洋理工大學圖書館, *juan* 2.14.
218 Die hier implizierten Verarbeitungsschritte sind auf S. 182 beschrieben, insb. Schritte 1–5.
219 *DHYDCD*, 蝴蝶.
220 Vgl. *CBDB*, S. ID 0094717.
221 Selbstverständlich läge mit dem Satz aus *Zhuangzi* 莊子 bereits eine deutlich frühere Belegstelle vor. Sehr wahrscheinlich ist der tatsächliche *Locus classicus* von *hudie* in noch älteren Texten zu finden. Vgl. George A. Kennedy 1964 [1955]: „The Butterfly Case, Part I“. In: *Selected Works of George A. Kennedy*. Hrsg. von Li Tien-yi. New Haven: Far Eastern Publications, S. 274–322, *passim*.

2. *xizhe* 昔者 – das *DHYDCD* gibt das *Yijing* 易經 als *Locus classicus* für *xizhe* an.[222] Der in der Datenbank enthaltene geschätzte ungefähre Entstehungszeitraum ist 850–800 v. .u. Z.,[223] es wird also der Wert -825 verwendet.
3. Zhuang Zhou *meng* 莊周夢 – dieser Ausdruck wird – wenig verwunderlich – mit dem *Zhuangzi* 莊子 belegt,[224] als mittleres Jahr des geschätzten Zeitraums der Entstehung wird hier 300 v. u. Z. angenommen.

Das *AYL* dieses einzelnen Satzes würde also berechnet als:

$$AYL = \frac{878 + (-825) + (-300)}{3} = 115\ v.\,u.\,Z.$$

Berechnet man das *AYL* für alle 2–4-Zeichen-Lexeme des gesamten *Zhuangzi* ergibt sich der Wert -155, für das deutlich ältere *Shangshu* 尚書 -459, für das Song 宋-zeitliche *Meng xi bi tan* 夢溪筆談 231. Das *AYL* kann keineswegs die Entstehungszeit eines Textes angeben, korreliert aber damit: Ältere Texte erhalten tendenziell niedrigere Werte. Mittels Linearregression kann dieser Zusammenhang zwischen *AYL* und Veröffentlichung zu datierender Texte formalisiert werden.

AYL und Textveröffentlichung – experimentelle Formalisierung

Die Eignung und optimale Verwendung des *AYL* zur Schätzung der Entstehungszeit von Texten muss experimentell angenähert werden, da zur Berechnung entweder der gesamte Wortschatz eines Textes, oder unterschiedliche Anteile oder Mengen seiner häufigsten Lexeme betrachtet werden können. Die so berechneten Werte werden mit dem tatsächlichen Jahr der Veröffentlichung von Texten eines bereits datierten Korpus korreliert. Je besser die erzielte Korrelation, desto besser der temporale Informationsgehalt des verwendeten Messwerts.

Zur Untersuchung dieses Zusammenhangs wird eine diachrone Textreihe benötigt, ein homogenes Korpus, das einen möglichst großen Zeitraum abdeckt. Hierfür bieten sich erneut die offiziellen Dynastiegeschichten (*zhengshi* 正史, siehe auch Kapitel 2.3, S. 20) an, deren Fertigstellung sich über den Zeitraum von 91 v. u. Z. bis 1928 erstreckt. Ein *Caveat* ist dabei – wie bereits in Kapitel 6.1 und 6.2 – die Überlagerung zweier temporaler Aspekte: Während die Texte – trotz ihrer strukturellen und stilistischen Anlehnung an das „Vorbild“ *Shiji* 史記 – sprachliche Merkmale aus der jeweiligen Zeit ihres Entstehens aufweisen und stilistische Einflüsse und Vorlieben der Autor:innen bzw. Kompilator:innen vorhanden sind,[225] enthalten sie doch in großem Umfang auch Lexeme, die spezifisch für den vorangegangenen, *erzählten* Zeitraum sind. Eine weitere Ungenauigkeit entsteht durch die Aufnahme von früherem Textmaterial. So wurde z. B. das *Hou Han shu* 後漢書 von Fan Ye 范曄 mehr als zweihundert Jahre *nach* Ende der Han-Dynastie zusammengestellt – allerdings größtenteils aus deutlich früher verfassten, tatsächlich Han-zeitlichen Dokumenten und Texten.[226] Eine Übersichtsdarstellung von inhaltlich abgedeckter Zeitperiode und Veröffentlichung der *zhengshi* findet sich in Kapitel 2.3.[227]

222 *DHYDCD*, 昔者.
223 Diese Angabe basiert auf den Ausführungen von Edward Shaughnessy, siehe Edward Shaughnessy 1993a: „*I ching* 易經 (*Chou I* 周易)“. In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 219.
224 *DHYDCD*, 莊周夢.
225 Siehe auch Kapitel 2.3, ab S. 20.
226 Siehe auch Kapitel 5.7.4. Zur Entstehungsgeschichte des *Hou Han shu* 後漢書 siehe Bielenstein 1954, S. 9–17.
227 Siehe Abb. 2.2, S. 22.

Abb. 6.31 zeigt die Korrelation zwischen *AYL* bei Berücksichtigung *aller* 2–4-Zeichen-Lexem-*types* (kurz $AYL_1^{2--4gram}$) und der Veröffentlichung der 25 *zhengshi* als Linearregression. Obwohl – gegeben durch die sehr unterschiedliche Länge der Korpustexte – die Anzahl der für die Berechnung des *AYL* berücksichtigten Lexeme bzw. Jahresangaben sehr unterschiedlich ist und zwischen 13.778 (*Chen shu* 陳書) und 73.174 Lexemen (*Song shi* 宋史) liegt, kann die Korrelation zwischen *AYL* und Veröffentlichung der Texte mit $R = 0,896$[228] bzw. $R^2 = 0,80$ als vielversprechend angesehen werden. Das *AIC* liegt bei 344,4.[229] Da das Jahr der Textgenese die später im Modell zu errechnende Variable ist, wird es hier auf der *y*-Achse dargestellt.

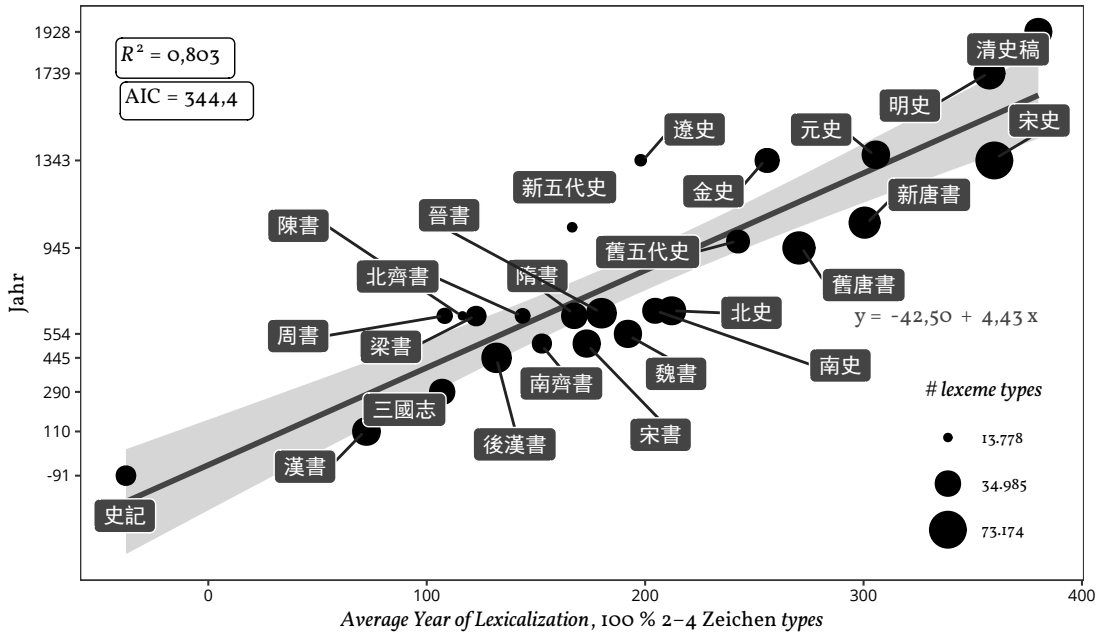


Abbildung 6.31 Korrelation Veröffentlichung *zhengshi*, *AYL* bei 100 % 2–4 Zeichen *types*

Korpus-Beobachtungen mit einer *PCA* legen nahe, dass nicht durch Betrachtung *aller*, sondern nur der *häufigsten types* eine höhere temporale Aussagekraft erzielt werden kann.[230] Zur Op-

228 Der Pearson-Korrelationskoeffizienten *R* misst, wie gut die Messwerte zweier Merkmale in einem linearen Modell miteinander korrelieren. Dabei steht der Wert 1 oder -1 für eine perfekte Abhängigkeit zwischen den Merkmalen, ist der Wert 0, sind sie unkorreliert. Der Korrelationskoeffizient wird aus der Summe der quadrierten Standardabweichungen berechnet. Siehe z. B. Fahrmeir et al. 2013, S. 287.

229 Akaike's *Information Criterion* (AIC, Akaikes Informationskriterium) ist ein Kriterium zur Auswahl statistischer Modelle, das 1973 von Hirotsugu Akaike vorgeschlagen wurde und das die Varianz der Residuen mehrerer Modelle vergleicht. Dabei ist das Modell mit dem geringsten *AIC* zu bevorzugen. Es eignet sich daher nur zum direkten, relativen Vergleich ähnlicher Modelle, da es kein absolutes Maß für die Qualität eines statistischen Modells darstellt. Siehe Jan deLeeuw 1992: „Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Hrsg. von Samuel Kotz und Norman L. Johnson. Bd. I: Foundations and Basic Theory. New York: Springer, S. 599–609, S. 607; bzw. die ursprüngliche Veröffentlichung: Akaike Hirotsugu 赤池弘次 1992 [1973]: „Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Hrsg. von Samuel Kotz und Norman L. Johnson. Bd. I: Foundations and Basic Theory. New York: Springer, S. 610–624.

230 Siehe auch Kapitel 3.1, S. 41; siehe auch T. Schalmey 2021, S. 254–255.

timierung der Korrelation kann das *AYL* für die *zhengshi* experimentell mit unterschiedlichen Parametern berechnet werden.[231] Folgende Optionen werden hierfür erörtert:

1. **Gewichtung**. Lassen sich durch die Berücksichtigung von Worthäufigkeiten bessere Ergebnisse erzielen, oder sollte lediglich betrachtet werden, *welche types* in einem Text vorkommen? Das *AYL* kann hierzu mit der Worthäufigkeit gewichtet werden (*Frequency Weighted Average Year of Lexicalization*, *FWAYL*).
2. **Gewichtungskorrektur**. Mit der bereits in Abschnitt 6.2.1 (ab S. 185) verwendeten *Gewichtungskorrektur* kann zudem ein *Standardized Average Year of Lexicalization* (*SAYL*) berechnet werden. Das durch die Kompilation des *DHYDCD* gegebene *Bias*[232] kann so reduziert werden, gleichzeitig findet aber auch eine Verzerrung der Daten statt.
3. Berücksichtigung von **Einzelzeichen**: Die Betrachtung von 2–4 Gramm-Lexemen im Vergleich zur Betrachtung von 1–4 Gramm-Lexemen. Dass für einen Teil der im *DHYDCD* lexikalisierten Einzelzeichen sehr frühe Belegstellen vorhanden sind, während spätestens ab dem Mittelalter nur noch wenig neue Schriftzeichen lexikalisiert werden[233] und die Ergebnisse von Kapitel 6.1 und 6.2 sprechen dafür, Vorkommen von Einzelzeichen aufgrund ihrer geringeren temporalen Entropie zu vernachlässigen.
4. **Anteil oder Anzahl?** Sollen *alle* Lexeme eines Textes, oder nur die häufigsten Lexeme für die Berechnung des *AYL* berücksichtigt werden? Sollte dafür stets dieselbe *Anzahl* an *types* berücksichtigt werden, oder ein festgelegter *Anteil*?
5. In diesem Kontext sollte zudem der Umgang mit der stark unterschiedlichen Länge der Korpustexte evaluiert werden. Während das *Chen shu* als kürzester Text „nur“ knapp 200.000 Schriftzeichen umfasst, hat die Geschichte der Song-Dynastie (*Song shi*) deutlich über 4 Mio. Zeichen.
6. Ist es zweckmäßig, **Interpunktion** in Texten beizubehalten, da sie teilweise Informationen zu Wortgrenzen enthält?

Gewichtung mit Worthäufigkeiten

Die in den Kapiteln 6.1 und 6.2 gemachten Beobachtungen implizieren, dass aufgrund der stilistischen Rigidität einiger schriftsprachlicher Textgattungen eine chronologische Einord-

231 Solche Experimente sind nicht unüblich. Vgl. z. B. Matthew L. Jockers 2013: *Macroanalysis: Digital Methods and Literary History*. Topics in The Digital Humanities. Urbana, Chicago und Springfield: University of Illinois Press, S. 63–104. Jockers zeigt in seinem Buch, wie durch die Betrachtung unterschiedlicher Messungen von (teils einzelnen) Worthäufigkeiten von Texten verschiedene Charakteristika sichtbar gemacht werden können. Er stellt fest, dass unterschiedliche Merkmale und Signale sich jeweils mehr oder weniger eignen, um Genres, Autoren, Geschlecht der Autorin bzw. des Autors, die Entstehungszeit von Texten oder die Texte selbst mit computerlinguistischen Mitteln voneinander zu unterscheiden. Vierthaler experimentiert mit dem *zhengshi*-Korpus, um die beste Metrik zur Genreunterscheidung von spätkaiserzeitlichen chinesischen Texten zu finden. Siehe Vierthaler 2016a, S. 8; Vierthaler bezieht sich dabei auf einen Beitrag von Christof Schöch. Ihn treibt die Frage um, welche Messwerte (z. B. bei unterschiedlicher Länge der Worthäufigkeitslisten) zu verwenden sind, um Autoren oder Genres im Rahmen einer *PCA* jeweils besser unterscheiden zu können. Christof Schöch 2012: „Author or genre? Assessing the quality of cluster analysis graphs in two-dimensional classification problems“. In: *The Dragonfly's Gaze: Computational analysis of literary texts*. URL: https://dragonfly.hypotheses.org/148 (besucht am 30. 12. 2018), Er kommt zu dem Ergebnis, dass sinnvoll ist, 750 oder mehr der häufigsten Wörter zu betrachten, um französische Stücke aus dem 17. Jh. nach Genre zu clustern. Dabei ließen sich Signale für Autorschaft und Signale für das Genre allerdings nicht klar trennen – dieses Beispiel zeigt, wie spezifisch am Ende eines solchen Experiments der Erkenntnisgewinn sein kann und dass am Ende für jeden unterschiedlichen Fall eigene Experimente durchgeführt werden sollten.

232 Siehe dazu Kapitel 5.7.2, ab S. 142.

233 Siehe dazu auch Kapitel 5.7, ab S. 138 bzw. Abb. 5.12, S. 147.

nung besser auf Basis der vorkommenden Lexeme, als über deren Häufigkeit funktioniert. Andererseits lassen sich über einen längeren Zeitraum hinweg durchaus Veränderungen z. B. von Häufigkeiten wichtiger Funktionswörter beobachten.[234] Auch wurde gezeigt, dass diese Veränderungen für die zeitliche Zuordnung von klassischen chinesischen Texten eine Rolle spielen können.[235] Es sollte daher geprüft werden, ob sich durch Einbeziehung ihrer Häufigkeit im jeweiligen Text als Gewicht für die einzelnen *types* eine stärkere Korrelation erzielen lässt. Hierzu wird das *Frequency Weighted Average Year of Lexicalization FWAYL* berechnet als:

$$FWAYL = \frac{\sum_{w_1}^{w_n} Y\times \mid t \mid}{\mid T \mid}$$

D. h. die Summe aller „Lexikalisierungsjahre" *Y* der enthaltenen Wort-*types w*, jeweils multipliziert mit der Häufigkeit des jeweiligen *type*s im untersuchten Text, geteilt durch die Gesamtanzahl *T* der *tokens t*. Die erzielte Korrelation zum Erscheinungsjahr der Korpustexte bei Berücksichtigung aller 2–4 Zeichen-Lexeme ist immer noch sehr gut, erreicht mit $R = 0,864$ bzw. $R^2 = 0,746$ aber nicht die des ungewichteten Modells (Tabelle 6.13, S. 215). Es bleibt für die *AYL*-Datierung eines Textes entscheidender, *welche* Wörter (häufig) darin vorkommen, als *wie häufig* diese Wörter im zu datierenden Text enthalten sind.

Eine mögliche Erklärung für die schwächere Korrelation ist, dass die *zhengshi* genretypische Lexeme aufweisen, die in allen Texten häufig sind und es zum Teil über den gesamten Betrachtungszeitraum (1. Jh. v. u. Z. bis 20. Jh.) auch bleiben.[236] Bei Berücksichtigung der Häufigkeit werden also auch zahlreiche *types* stark gewichtet, deren Häufigkeit relativ konstant ist. Hinzu kommt die Tatsache, dass alte Wörter tendenziell häufiger sind.[237]

Gewichtungskorrektur

Das *Bias*, das durch die Auswahl der Einträge und Belegstellen im *DHYDCD* entsteht, führt zu einer ungleichen Gewichtung der Lexikalisierungszeiträume.[238] Um den Effekt dieser ungewollten Gewichtung auf die Berechnung des *AYL* zu reduzieren, kann ein gewichtungskorrigiertes durchschnittliches Wortentstehungsjahr (*Standardized Average Year of Lexicalization*, *SAYL*) berechnet werden:

$$SAYL = \frac{\sum_{c_1}^{c_n} |V_c| \times g_c \times (c + 50)}{\sum_{c_1}^{c_n} |V_c| \times g_c}$$

234 Siehe dazu Kapitel 2.3, ab S. 20.
235 Siehe Kapitel 3.1, S. 41. Siehe auch Yamada Takahito 山田崇仁 2004.
236 Von 1.000 der häufigsten 2–4-Zeichen-Lexeme des ältesten (*Shiji* 史記), des neuesten (*Qing shi gao* 清史稿), sowie eines mittleren Korpustextes (*Jiu Tang shu* 舊唐書) treten 12 % in allen drei Texten auf, in *Shiji* und *Jiu Tang shu* sogar 30 %, *Jiu Tang shu* und *Qing shi gao* 25 %. Betrachtet man 1–4-Zeichen-Lexeme ergibt sich sogar eine Übereinstimmung von 50,3 % in allen drei Texten. Dass eine hohe Ähnlichkeit von Wortschatz und -häufigkeit innerhalb des *zhengshi* Korpus besteht, hat auch die Studie von Vierthaler bereits eindrücklich gezeigt. Siehe Vierthaler 2016a, z. B. S. 26. In einer *Principal Component Analysis* der 1.000 häufigsten Zeichen bilden die Texte eindeutige *Cluster* abseits von anderen untersuchten Genres.
237 Siehe auch Kapitel 6.2, S. 181.
238 Siehe Kapitel 6.2.1, ab S. 185.

Die Anzahl der jedem Jahrhundert *c* zugeordneten *types* ($|V|$), multipliziert mit dem Gewicht *g* des jeweiligen Jahrhunderts und dem mittleren (50.) Jahr, geteilt durch die genau so gewichtete Gesamtanzahl an *types*. Zur Gewichtung kann sowohl ein linear gleichförmiges, als auch ein *s*-förmiges Wortschatzwachstum angenommen werden.[239] In beiden Fällen lässt sich keine Verbesserung der Korrelation erzielen, bei Verwendung der linearen Gewichtungskorrektur verschlechtert sie sich leicht auf $R = 0,859$ (Tabelle 6.13), mit *s*-Gewichtungskorrektur bleibt sie mit $R = 0,895$ nahezu identisch (Tabelle 6.13). Zur Berechnung des *SAYL* müssen die Daten pro Jahrhundert aggregiert werden, um das ebenfalls nur auf Jahrhunderte genau berechnete *Bias* zur Gewichtung verwenden zu können. Es ist zu vermuten, dass die so entstehende Unschärfe eventuelle positive Effekte der Gewichtungskorrektur wieder ausgleicht.

Eine geringfügig kleinere Maximalabweichung im *SAYL* Modell mit *s*-Gewichtungskorrektur, rechtfertigt kaum die gegenüber dem *AYL* erhöhte Komplexität.[240]

Tabelle 6.13 Vergleich linearer Modelle

Modell	R	R^2	*AIC*	Δ_{max}	Regress.gerade
Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,896	0,803	344,4	510	$y = -42 + 4,4x$
Frequency Weighted Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,864	0,746	350,7	593	$y = 630 + 4,9x$
Standardized Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,859	0,738	351,5	585	$y = 1200 + 4,5x$
s-Standardized Average Year of Lexicalization, 100 % 2–4 Zeichen *types*	0,895	0,8	344,7	489	$y = -210 + 4,5x$

Berücksichtigung von Einzelzeichen und Anteil verwendeter Lexeme

In den bisherigen *AYL*-Berechnungen wurden nur Lexem-Einträge mit 2–4 Zeichen Länge berücksichtigt. Verwendet man zusätzlich einzelne Zeichen, die im *DHYDCD* ebenfalls mit datierbaren Belegen aufgeführt sind (*dan zi tiaomu* 單字条目), stehen mehr verwertbare Daten zur Verfügung. Das *AYL* kann zudem auch dann berechnet werden, wenn ein Text kaum 2–4-Zeichen-Lexeme enthält. Gegen die Verwendung sprechen bisherige Erkenntnisse über die deutlich geringere temporale Entropie von Einzelzeichen. Ein großer Teil des heute verwendeten, standardisierten Zeichenrepertoires ist bereits sehr früh nachweisbar, während nach der Han-Zeit nur noch wenige Zeichen hinzu kommen.[241]

Um die Verwendung von Einzelzeichen für die Berechnung des *AYL* zu evaluieren, lohnt es sich, zeitgleich zwei weitere Aspekte zu betrachten. Wenn bestenfalls nur ein zu bestimmender **Anteil** der *häufigsten* Lexem-*types* verwendet werden sollte, kann dieser – abhängig von der Länge der zu berücksichtigenden Lexem-*types* – unterschiedlich ausfallen. Da die Korpustexte überdies unterschiedlich lang sind, führt aber ein fixer *Anteil* an Lexemen zu stark unterschied-

239 Erläuterungen und Berechnung der Gewichte siehe unter Abschnitt 6.2.1, ab S. 185.

240 Eine Diskussion zum Vorzug einfacherer Modelle im Sinne von Ockhams Rasiermesser findet sich z. B. in Malcolm Forster und Elliot Sober 1994: „How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions“. In: *The British Journal for the Philosophy of Science* 45.1, S. 1–35.

241 Siehe dazu Kapitel 5.7, v. a. S. 147. Eine völlig andere Situation ergibt sich, wenn man zu bestimmten Zeiten oder Orten verwendete Zeichenvarianten (*yiti zi* 異體字 u. ä.) betrachtet, wie sie in großen Zeichenwörterbüchern wie dem *Hanyu da zidian* 漢語大字典 (*Großes Lexikon chinesischer Schriftzeichen*) gesammelt werden. Vgl. dazu auch die etwas irreführende Darstellung in Best und Zhu Jinyang 2006, S. 208–209.

lichen Grundmengen für die Berechnung. Parallel sollte also die Verwendung einer fixen **Anzahl** an *types* geprüft werden.

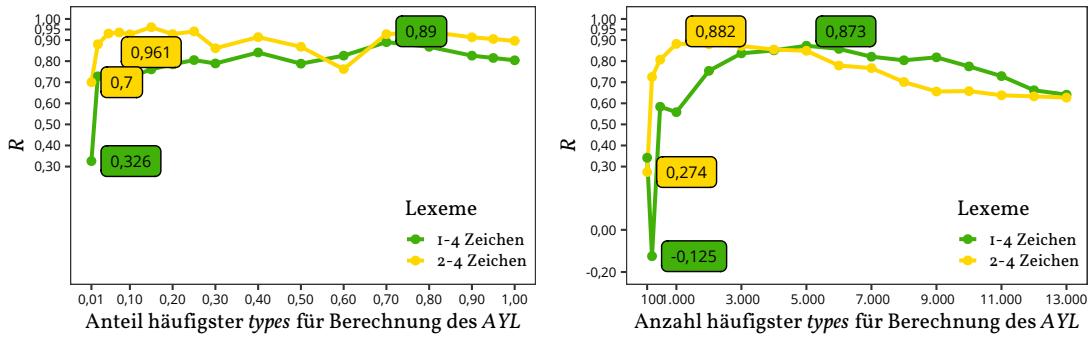


Abbildung 6.32 Vergleich linearer Modelle, *AYL* mit fixem *Anteil* und *Anzahl* von 1–4 und 2–4 Zeichen-Lexemen

Das *AYL* wird für unterschiedliche Anteile (1 %, 2,5 %, 5 %, 10 %, 15 %, 20 %, 30 %, … 90 %, 95 %, 100 %) und Mengen (100, 500, 1.000, 2.000, … 13.000[242]) an jeweils häufigsten Lexemen neu berechnet und *R* für die Korrelation mit dem Jahr der Veröffentlichung bestimmt. Um einen kompakten Überblick zu ermöglichen, zeigt Abb. 6.32 die Veränderung dieser Korrelation für lineare Regressionsmodelle mit einem steigenden Anteil (links) bzw. einer zunehmenden Anzahl berücksichtigter *types*.[243]

Ab etwa 1.000 bzw. 2,5 % der *types* lässt sich für den Zeitraum von 91 v. u. Z. bis zur Veröffentlichung des *Qing shi gao* 清史稿 1928 ein starker linearer Zusammenhang zwischen *AYL* und dem Jahr der Veröffentlichung herstellen. Wie erwartet lassen sich insgesamt mit 2–4-Zeichen-Modellen (gelb) bessere Ergebnisse erzielen als mit 1–4 Zeichen (grün). Die Verwendung eines *relativen Anteils* an häufigsten Lexemen (links) scheint überdies etwas bessere Korrelationen zu ermöglichen, als dies bei einer festgelegten *Anzahl types* (rechts) der Fall ist.

Für beide Fälle ist ein Optimalbereich der Berechnungsgrundlage für die temporale Aussagekraft des *AYL* erkennbar. Die stärkste Korrelation zur Veröffentlichung des Textes ergibt sich hier mit 15 % der häufigsten 2–4 Gramm *types*. Dabei werden zur Berechnung des *AYL* für das *Chen shu* 2.066, im Fall der *Song shi* 10.976 Lexemdatierungen zugrunde gelegt. Trotz dieser Diskrepanz wird mit relativen Anteilen eine bessere Korrelation erzielt. Bei Verwendung einer absoluten Anzahl – z. B. 2.000 – häufigsten *types* werden bei Betrachtung kürzerer bzw. weniger diverser Texte ungleich mehr seltenere Lexeme berücksichtigt, was ursächlich für die Verschlechterung der Korrelation sein kann. Bei Verwendung eines größer werdenden Anteils seltener *types* nehmen die Korrelationen allgemein wieder ab. Bei Modellen mit einer festgelegten Anzahl an 1–4 Grammen ist derselbe Effekt mit einer leichten Verschiebung zu beobachten, da bei der Betrachtung von Einzelzeichen potenziell erst mit mehr Daten sehr seltene *types* berücksichtigt werden.[244]

242 Um eine für alle Texte identische Anzahl zu ermöglichen, stellt die mit 13.778 von allen Korpustexten geringste Anzahl *types* des *Chen shu* die Obergrenze für dieses Experiment dar.

243 Auf die explizite Darstellung der insgesamt 64 Regressionsmodelle wie in Abb. 6.31 (S. 212) wird hier bewusst verzichtet.

244 Da die Texte nicht linear segmentiert sind, werden Einzelzeichen hier in jedem Vorkommen gezählt und nicht nur dann, wenn sie als eigenständiges Wort vorkommen.

Zur Ermittlung eines fixen Anteils der häufigsten Lexeme eines Textes müssen aus seinen *n*-Grammen zuerst *alle* infrage kommenden Lexeme ermittelt werden. Um Rechenleistung einzusparen, kann gleichermaßen auch direkt mit den *n*-Gramm-Frequenzlisten gearbeitet werden.[245] Sehr starke Korrelationen (*R* von 0,955, 0,967, 0,96) lassen sich bei Verwendung eines relativen Anteils im Bereich von 1 %, 3 %, 5 % an häufigsten *2–4*-Gramm-*types* erzielen. Da bei kürzeren Texten die Menge der tatsächlich berücksichtigten Lexeme mit steigender Anzahl berücksichtigter *n*-Gramm *types* immer weniger zunimmt, lässt sich auch mit einer festgelegten Anzahl z. B. der 200.000 häufigsten 2–4-Gramm-*types* eine annähernd gleich starke Korrelation mit *R* = 0,952 beobachten.[246]

Eine inzestuöse Korrelation?

Zwei Faktoren tragen hier sicherlich zur starken Korrelation bei: Als dynastiespezifische Geschichtstexte enthalten die *zhengshi* zwangsläufig eine große Anzahl zeitspezifischer oder -typischer Lexeme. Ein großer Teil der 25 Dynastiegeschichten ist zudem im *DHYDCD* mit zahlreichen Belegstellen präsent.[247] Dadurch wird automatisch ein Teil der in eben diesen Texten gefundenen Lexeme genau dem Zeitraum zugerechnet, in den der Text datiert werden soll. Das impliziert eine problematische Identität von Test- und Trainingsdaten, deren Auswirkungen sich in der Praxis aber als marginal erweisen. Zur Veranschaulichung sei hier das „Paradebeispiel“ des *Hou Han shu* (*HHS*) angeführt: Die verwendete Ausgabe enthält 43.705 unterschiedliche Lexeme von 2–4 Zeichen Länge.

Um die tatsächliche Auswirkung der „selbstevidenten“ Lexeme auf die Berechnung des *AYL* und damit auf die Korrelation zu bewerten, können für die Berechnung die knapp 8.000 *types* (etwa 18 %), für die das *HHS* selbst als erste Belegstelle angegeben wird, weggelassen werden. Bei Berücksichtigung aller *types* würde das *AYL* tatsächlich stark beeinflusst. Da es sich bereits als zielführend erwiesen hat, nur die häufigsten *types* eines Texts zu betrachten, ist die Hebelwirkung der dann verbleibenden „inzestuösen“ *types* auf die Berechnung des *AYL de facto* allerdings überraschend gering.

Verwendet man z. B. 3 % Prozent der häufigsten 2–4-Gramme zur Berechnung des *AYL*, sind lediglich etwa 8 % der verwendeten Wort-*types* des *HHS* „selbstevident“.[248] Die übrigen *zhengshi* wurden von den Kompilator:innen des *DHYDCD* weniger häufig als *Locus classicus* herangezogen, so dass der Effekt noch geringer bzw. nahezu bedeutungslos wird.[249]

245 Da – im Gegensatz zu 2–*n*-Grammen – ein sehr hoher Anteil der vorkommenden Zeichen im *DHYDCD* lexikalisiert und belegt sind, kann mit einer geringeren Anzahl häufigster 1–*n*-Gramme eine gute Korrelation zum Erscheinungsjahr der Texte erreicht werden. Insgesamt können mit 2–4-Grammen aber bessere Korrelationen erzielt werden als mit 1–4-Grammen.

246 Bei einer Betrachtung von 2–4-Grammen haben die untersuchten Texte zwischen 349.232 und 6.614.725 unterschiedliche *n*-Gramm *types*. Werden für alle Korpustexte nur die häufigsten 1 % *n*-Gramm-*types* untersucht, sind etwa 16 % davon tatsächlich im *DHYDCD* lexikalisiert. Durchschnittlich 19.000 *n*-Gramm-*types* enthalten etwa 3.000 Lexem-*types* pro Korpustext. Die *n*-Gramm *type-token*-Relation (*TTR*, Diversifikationsquotient) der Texte liegt zwischen 145.795 und 211.005 *types* pro 100.000 Zeichen. So enthält die verwendete Ausgabe des *Shiji* 史記 569.000 Zeichen, die etwas über eine Million 2–4-Gramm-*types* bilden. Untersucht man die häufigsten 1 % bzw. 10.300 davon, sind davon wiederum 1.595 (15 %) im *DHYDCD* lexikalisiert und belegt.

247 Siehe Kapitel 5.7.4, ab S. 150. Insbesondere bei *Shiji* 史記, *Han shu* 漢書 und *Hou Han shu* 後漢書 ist Vorsicht geboten, da es sich um die am häufigsten im *DHYDCD* als *Locus classicus* angegebenen Texte handelt – auch einige andere Texte des *zhengshi*-Korpus sind aber häufig im *DHYDCD* zitiert.

248 660 von 8.079 Lexem-*types*, die chronologisch zugeordnet werden können.

249 z. B. sind bei Betrachtung der 3 % häufigsten 2–4-Gramme nur 132 von 11.000 erkannten Lexemen des *Jiu Tang shu* 舊唐書 mit eben diesem Text im *DHYDCD* belegt. Der Einfluss auf die Berechnung des *AYL* ist damit marginal. Bei anderen *zhengshi* fällt der Einfluss teilweise noch geringer aus.

Berechnet man die oben gezeigten 2–4-Gramm-Modelle für das *AYL* unter Ausschluss der jeweiligen *Locus classicus*-Lexeme erneut, ist tatsächlich nur eine marginale Verringerung von *R* um etwa 0,01 bis 0,03 zu beobachten. Wie zu erwarten ist die Auswirkung auf Modelle, die auf einem geringen prozentualen Anteil basieren besonders gering, z. B. von 0,967 auf 0,956 bei Verwendung von 3 % der häufigsten 2–4-Gramme. Mit 100.000 der häufigsten 2–4-Gramm-*types* verschlechtert sich der Korrelationskoeffizient *R* von 0,95 auf 0,926. Insofern kann Entwarnung gegeben werden: Der experimentelle Ausschluss von „selbstevidenten" Lexemen zeigt, dass die Verortung der *zhengshi* im *DHYDCD* Korrelationen insgesamt nur marginal begünstigt. Die Korrelationen zwischen *AYL* und der Entstehung von Texten bestehen auch unabhängig von diesem Einfluss. Da in einem „realen" Anwendungsfall der zu datierende Text möglicherweise weder bekannt, noch im *DHYDCD* als *Locus classicus* angegeben wäre, bleibt diese Art des Ausschlusses eine theoretische Überlegung.

Nutzt man hingegen die mit zusätzlichen Belegen aus dem *zhengshi*-Korpus erweiterte diachrone Lexemdatenbank zur Berechnung des *AYL*,[250] kann erwartungsgemäß eine weitere Verbesserung der Korrelation erreicht werden. Mit 1,5 % der häufigsten 2–4-Gramme der *zhengshi* wird ein theoretischer Korrelationskoeffizient von $R = 0{,}968$ erreicht. Für die Datierung der Dynastiegeschichten käme das endgültig einer Identität von Trainings- und Testdaten nahe – zum Zweck der Datierung anderer, nicht im *DHYDCD* verorteter Texte kann diese Erweiterung hingegen sinnvoll sein.

Umgang mit der unterschiedlichen Länge der Texte

Die Korrelation zwischen *AYL* und Veröffentlichung von Texten bei Verwendung eines fixen Anteils an häufigsten Lexemen scheint auch bei stark unterschiedlicher Länge der betrachteten Texte verhältnismäßig stabil zu sein. Zusätzlich kann geprüft werden, wie sich die Zerteilung der Korpustexte in gleich lange Abschnitte auswirkt.[251] Da die verwendete Version des *Chen shu* 176.000 und einige weitere Texte ebenfalls kaum mehr als 200.000 Zeichen aufweisen, wird hier mit Abschnitten zwischen 10.000 und maximal 150.000 Zeichen Länge experimentiert. Die Werte für *R* aus diesen Experimenten werden in Tabelle 6.14 aufgeführt.

Mit Abschnitten von 10.000 Zeichen können die Werte für das *AYL* der einzelnen Abschnitte sehr weit auseinander liegen. Würde die Entstehungszeit des *Qing shi gao* mithilfe einer Linearregression auf das jeweils niedrigste und höchste *AYL* der einzelnen Textpartitionen geschätzt, ergäbe sich ein Spielraum von über 2.000 Jahren, von ca. 360 bis zum Jahr 2450. Bei Verwendung längerer Textabschnitte nimmt diese Spannbreite ab und die Korrelation zu: Mit Abschnitten von 50.000 Zeichen kann ein Wert von $R = 0{,}835$ (100.000: 0,857, 150.000: 0,877) erreicht werden. Die einzelnen Abschnitte des *Qing shi gao* würden auf den Zeitraum zwischen 680 und 2280 datiert.

Wird anstatt der Einzelbeobachtungen für die jeweiligen Textpartitionen der *AYL*-Mittelwert bzw. Median für alle Abschnitte eines Textes berechnet, können stärkere Korrelationen auch schon mit kürzeren Abschnitten ab 10.000 Zeichen erzielt werden.[252] Ein sehr guter Wert von *R* kann bei Verwendung eines optimierten Anteils von ca. 5–10 % der häufigsten 2–4 Gramme und Abschnitten ab 50.000 Zeichen erzielt werden.

250 Siehe dazu Kapitel 5.5.4, ab S. 134.

251 Diese Vorgehensweise ist auch bei anderen quantitativen Untersuchungen üblich. Vgl. z. B. Binongo und Smith 1999, S. 448; zitiert in Vierthaler 2016a, S. 7–8. Vierthaler unterteilt seine Texte in Abschnitte von je 10.000 Zeichen. Der letzte Abschnitt wird dabei in aller Regel verworfen.

252 Mit dem Mittelwert aller Beobachtungen werden minimal bessere Korrelationen erzielt als mit dem Median.

Tabelle 6.14 Korrelationskoeffizient (*R*) für Anteile häufigster 2–4-Gramme aus 10.000–150.000-Zeichen Abschnitten (unter „all“ wird *R* bei Betrachtung aller Einzelbeobachtungen angegeben)

Abschn.	10.000 Zeichen			50.000 Zeichen			100.000 Zeichen			150.000 Zeichen		
% 2–4 Gramme	all	mean	med.	all	mean	med.	all	mean	med.	all	mean	med.
5 %	0,684	0,921	0,919	0,795	0,951	0,937	0,822	**0,962**	0,949	0,852	**0,959**	0,948
10 %	0,722	**0,935**	0,926	0,808	**0,955**	0,942	0,83	**0,957**	0,944	0,862	**0,954**	0,949
20 %	0,739	0,932	0,926	0,816	0,948	0,933	0,838	0,947	0,931	0,868	0,944	0,941
30 %	0,746	0,925	0,926	0,819	0,939	0,928	0,839	0,938	0,924	0,867	0,936	0,934
40 %	0,751	0,918	0,916	0,825	0,935	0,916	0,842	0,931	0,919	0,872	0,933	0,933
50 %	0,761	0,923	0,92	0,826	0,932	0,913	0,846	0,93	0,916	0,874	0,931	0,931
60 %	0,767	0,921	0,918	0,827	0,927	0,908	0,848	0,928	0,915	0,875	0,929	0,927
70 %	0,77	0,921	0,918	0,828	0,924	0,903	0,851	0,928	0,915	0,878	0,924	0,92
80 %	0,772	0,92	0,916	0,833	0,925	0,908	0,852	0,925	0,915	0,877	0,923	0,915
90 %	0,772	0,919	0,918	0,834	0,924	0,91	0,853	0,923	0,915	0,877	0,922	0,918
100 %	0,777	0,921	0,92	0,836	0,924	0,908	0,855	0,921	0,913	0,877	0,92	0,915

Insgesamt verschlechtert das Partitionieren der Texte bei Betrachtung der *AYL*-Mittelwerte für ausreichend lange Textabschnitte die Korrelation zwischen *AYL* und Textgenese nicht. Eine Verbesserung, die diese Maßnahme rechtfertigen würde, stellt sich aber ebenfalls nicht ein. Dass ein *R* von 0,935 bereits mit dem *AYL*-Mittelwert aller Partitionen von 10.000 Zeichen erreicht werden kann, deutet aber an, dass eine temporale Aussagekraft des *AYL* für die chronologische Einordnung von Texten bei dieser Textlänge bereits gegeben ist.

Löschung von Interpunktionszeichen

Die in chinesischen *DH*-Anwendungen verbreitete Entfernung der Interpunktionszeichen[253] wirkt sich kaum auf die Korrelation des *AYL* mit der Textgenese aus. Da Interpunktion zumindest an Satzenden *false positives*, die sonst durch die verwendete *n*-Gramm-Segmentierung entstehen, verhindert, sind Vorteile durch die Löschung der Interpunktion im gegebenen Kontext aber auch nicht erwartbar. Beim Vergleich von *n*-Gramm-Modellen mit und ohne Entfernung der Interpunktion lassen sich minimale Unterschiede bei der Korrelation feststellen. Diese sind vermutlich überwiegend auf eine eine Verschiebung des Optimalbereichs für den den Anteil der zu betrachtenden *types* zurückzuführen, da bei gleicher Länge der betrachteten *n*-Gramm-Häufigkeitslisten nach Entfernen der Interpunktion mehr vermeintliche Lexeme erkannt werden.[254] Die Entfernung der Interpunktion aus allen Texten mag für gemischte Korpora, d. h. solche, die sowohl Texte mit, als auch ohne Interpunktion enthalten, allerdings sinnvoll sein, um Verzerrungen vorzubeugen.

253 Siehe z. B. Vierthaler 2016a, S. 7.

254 Da etliche zuvor gebildete, häufige *n*-Gramme mit ，。 und 、etc. wegfallen, können nach Entfernung der Interpunktion in den (z. B.) 100.000 häufigsten *n*-Grammen mehr tatsächliche Lexeme ausgemacht werden. z. B. sind 8.522 der häufigsten 100.000 2–4-Gramme der interpungierten Version des *Shiji* 史記 im *DHYDCD* lexikalisiert. Nach Entfernung der Interpunktion sind es 9.350, da häufige 2–4-Zeichen Kombinationen mit Interpunktionszeichen wie „曰「“, „。」“, „也。“ oder „之。“ wegfallen und tatsächliche Wörter „nachrücken“.

Zusammenfassung

1. Die stärksten Korrelationen zwischen *AYL* und Entstehung von Texten lassen sich für das betrachtete *zhengshi*-Korpus mit der Berücksichtigung eines optimierten Anteils von etwa 15 % der häufigsten 2–4 Zeichen Lexem-*types* bzw. ca. 1–3 % der häufigsten 2–4-Gramme erzielen.
2. Statt eines festen Anteils häufigster Lexeme kann auch ein fester Anteil oder eine Anzahl an häufigsten *n*-Grammen festgesetzt werden.[255]
3. Die Korrelation verschlechtert sich sowohl bei der Betrachtung von zu vielen seltenen *types*, also auch dann, wenn zu wenige *types* berücksichtigt werden.
4. Vorkommende Einzelzeichen-*types* (Unigramme) eignen sich weniger zur Datierung der Texte als 2–4-Gramm-*types*, da die meisten Schriftzeichen bereits sehr früh lexikalisiert wurden.[256]
5. Die Gewichtung von Lexemen mit der Häufigkeit ihres Auftretens führt ebenfalls nicht zu stärkeren Korrelationen zwischen Textgenese und (*FW*)*AYL*. Dennoch ist die Worthäufigkeit ein sinnvolles Kriterium für die Auswahl der zu betrachtenden *types*.
6. Ein Ausgleich des *DHYDCD-Bias* durch eine Gewichtungskorrektur bietet einen Mehrwert für Visualisierungen,[257] ihr Nutzen für die Berechnung des (*S*)*AYL* wird aber durch die entstehende Unschärfe eliminiert.
7. Als Durchschnittswert ist das *AYL* robust gegenüber unterschiedlichen Textlängen bzw. Mengen an *types*, aus der es berechnet wird. Eine Partitionierung von Texten in Abschnitte gleicher Länge ist daher nicht unbedingt erforderlich. Mit dem arithmetischen Mittel der *AYL*-Berechnungen für gleich lange Textpartitionen ab einer Länge von 10.000 Zeichen kann ebenfalls eine starke Korrelation zur Entstehung der Texte festgestellt werden.
8. Die Entfernung von Interpunktionszeichen ist nur zu empfehlen, wenn Texte mit und ohne Interpunktion miteinander verglichen werden sollen.

6.3.1 Ein optimiertes *AYL*-Regressionsmodell

Nach Analyse der Parameter für die Berechnung des *AYL* kann ein lineares Regressionsmodell mit optimierter Korrelation zwischen *AYL* und Entstehung der Korpustexte bei Berücksichtigung von 15 % der häufigsten 2–4-Zeichen-Lexeme detaillierter betrachtet werden. Abb. 6.33 zeigt dieses Modell mit $R = 0{,}961$ und einem *AIC* von 320,7.

Steigung (4,22) und Achsenabschnitt (*Intercept* 639,8) sind beide hoch signifikant.[258] Mit der Regressionsgeraden des Modells (Abb. 6.33) kann das *AYL* der häufigsten Lexem-*types* auf die ungefähre Zeit der Entstehung $Y_{proj.}$ eines geeigneten Eingabetextes projiziert werden. Mit 15 % der häufigsten 2–4-Zeichen-Lexemen also: $Y_{proj.} = 4,22 \times AYL + 639,8$.

255 Da bei einer *n*-Gramm-Segmentierung die Häufigkeit der vorhandenen Lexem-*types* erst nach Ermitteln aller *n*-Gramm-*types* berechnet werden kann, ist die Betrachtung von Lexem-*types* mit Performanceeinbußen verbunden.

256 Siehe v. a. auch Abb. 5.11, S. 147.

257 Siehe dazu Abschnitt 6.2.1, ab S. 185.

258 Das Modell wird in *R* mit `lm` (*linear model*) berechnet. *t*: Die Nullhypothese besagt: *AYL* hat keinen Zusammenhang mit dem tatsächlichen Entstehungsjahr des Textes. Je niedriger *P* ist, desto geringer ist die Wahrscheinlichkeit, dass die Nullhypothese stimmt. Bei Werten kleiner als 0,05 kann man die Nullhypothese „kein Zusammenhang" in der Regel verwerfen. Die Signifikanzwerte *p* belaufen sich für das oben gezeigten Modell auf $2e^{-16}$ für den Achsenabschnitt und $2,47e^{-14}$ für die Steigung.

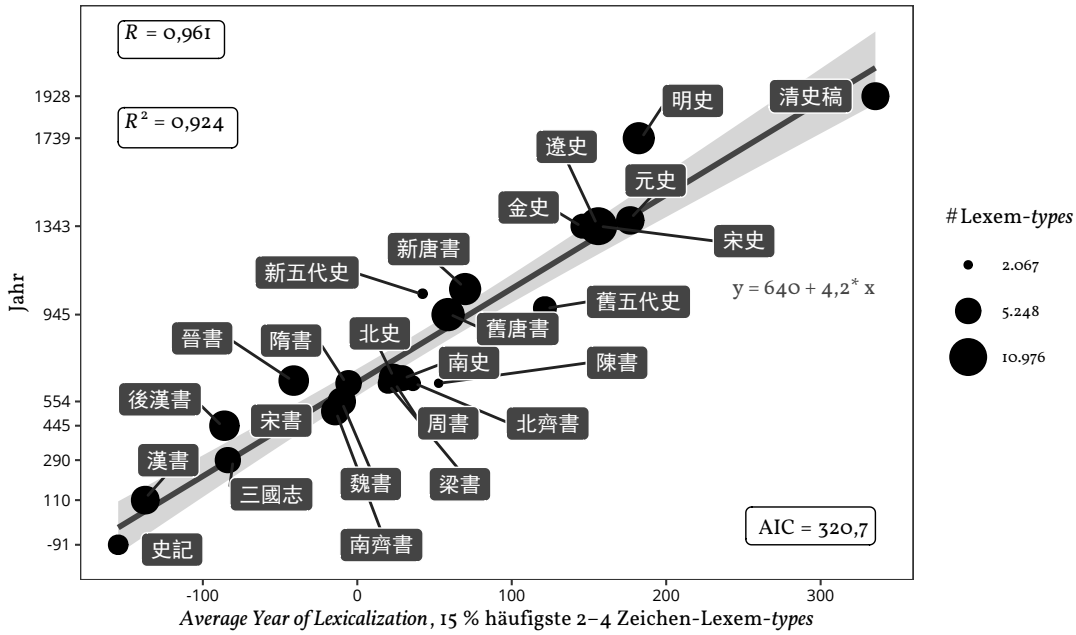


Abbildung 6.33 Korrelation Veröffentlichung *zhengshi*, *AYL* mit 15 % 2–4 Zeichen Lexem-*types*.

Der Standardfehler dieser Regression beträgt 136,6 [Jahre].[259]

Die Diagnoseplots in Abb. 6.34 ermöglichen ein genaueres Verständnis des Regressionsmodells.[260] Die Darstellung der Residuen (oben links) macht deutlich, welche Texte in dem verwendeten Modell am stärksten von der gewünschten Datierung abweichen. Die stärkste Abweichung vom Idealbild des Modells stellt dabei die *Ming shi* 明史 dar, die auf das Jahr 1409 datiert würde – 330 Jahre zu früh. Diese Datierung fällt aber in den im Text beschriebenen Zeitraum (1368–1644). Der erst 1739 veröffentlichte Text wurde zudem bereits 1645 in Auftrag gegeben und trotz der großen, teils herrschaftspolitisch bedingten Verzögerung naturgemäß aus Ming-zeitlichen Materialen kompiliert.[261] Im Hinblick darauf, dass die Geschichte der Ming also im Verhältnis zur beschriebenen Zeitperiode spät fertiggestellt wurde,[262] passt die Abweichung im Modell zu Inhalt und Textgeschichte. Dieser Erklärungsansatz passt auch zu den übrigen, deutlich zu alt datierten *zhengshi*: *Xin Wudai shi* 新五代史, *Jin shu* 晉書 und *Hou Han shu* 後漢書. Die in der Datierung um 221 Jahre abweichende *Xin Wudai shi* wurde von

259 Der *Residual standard error* gibt die Wurzel des durchschnittlichen Quadrats der Residuen an, d. h. hier die ungefähre durchschnittliche Abweichung des Alters eines zu datierenden Textes vom Idealbild der Regressionsgeraden.

260 *R* erzeugt mit der Funktion `plot` vier Diagnoseplots: *Residuals vs. Fitted* (Residuen gg. angepasste Werte), *Normal* Q-Q (Verteilung der Residuen), *Scale-Location* (ähnlich wie bei *Residuals vs Fitted* werden die Residuen gg. die angepassten Werte dargestellt, erstere werden allerdings normalisiert), sowie *Residuals vs Leverage* (Residuen und ihre Hebelwirkung auf das Modell anhand des Cook'schen Abstands). Die hier wiedergegebenen Plots wurden aus ästhetischen Gründen mit *ggplot2* erzeugt. Zur Erzeugung von Standard-Diagnoseplots mit `ggplot` siehe Raju Rimal 2014: *Diagnostic Plots using ggplot2*. Website. URL: `https://rpubs.com/therimalaya/43190` (besucht am 07. 11. 2018).

261 Siehe Thomas Wilson 1994: „Confucian Sectarianism and the Compilation of the Ming History“. In: *Late Imperial China* 15, S. 53–84. DOI: `10.1353/late.1994.0002`, S. 62; siehe auch Edward L. Farmer et al. 1994: *Ming History: An Introductory Guide to Research*. Ming Studies Research Series 3. Minneapolis: Center for Early Modern History, University of Minnesota, S. 72.

262 Siehe dazu auch Abb. 2.2, S. 22.

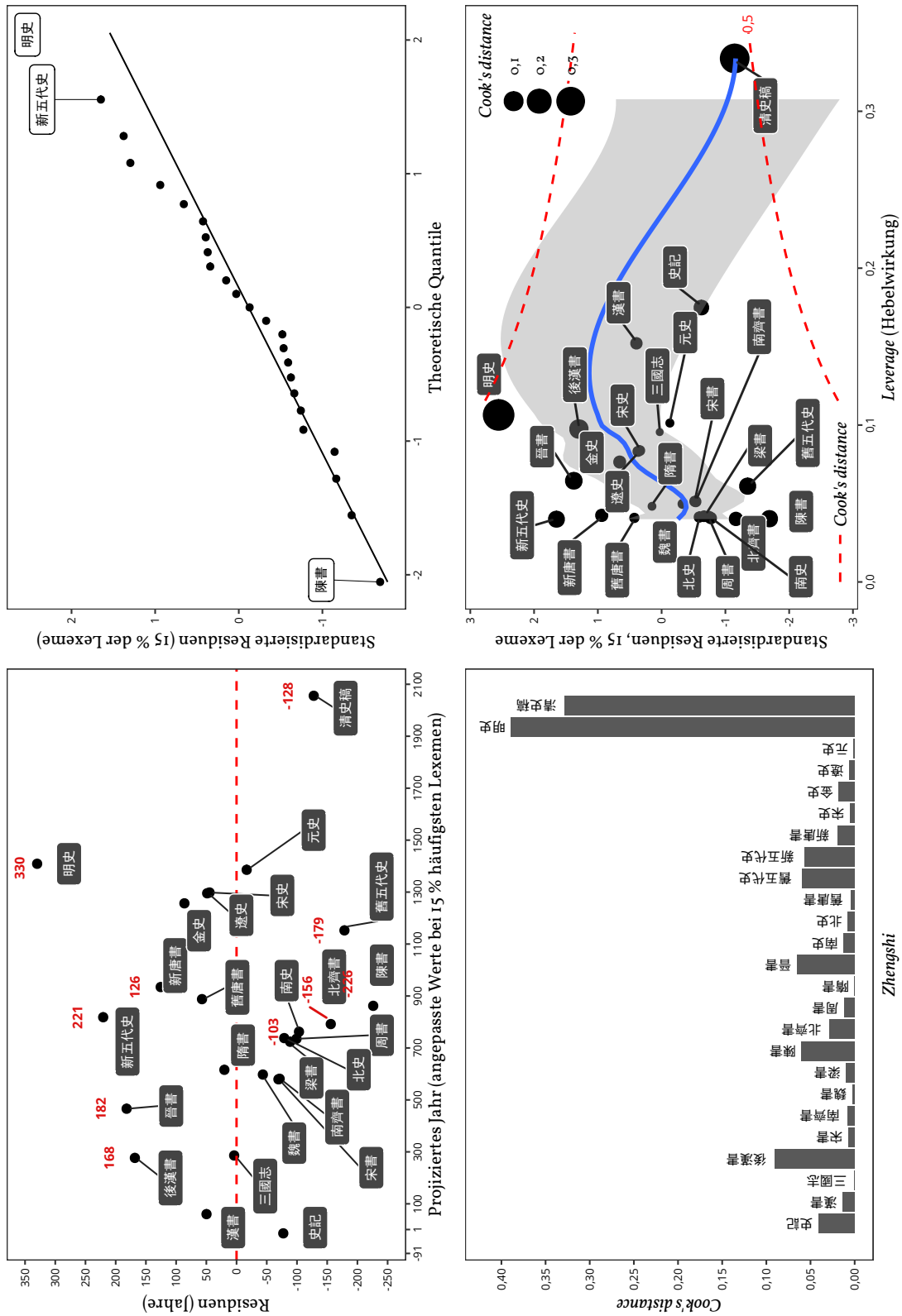


Abbildung 6.34 Diagnoseplots (Residuen, *Normal Q-Q*, *Cook´s distance*, *Leverage*)

Tabelle 6.15 Datierung unterschiedlicher Texte mit *AYL*, 15 % der häufigsten 2–4 Zeichen-Lexeme

Text	*AYL*-Datierung	tatsächliche Datierung	Δ_{min}
Dao de jing 道德經	-763	ca. –500—-400	-263
Zhuangzi 莊子	-116	ca. -400—200	+84
Zhongjing 忠經	81	ca. 320–960	-239
Wen xuan 文選	14	520–530 [ca. -300–500]	-
Zi zhi tong jian 資治通鑑	627	1084	-460
Meng xi bi tan 夢溪筆談	1282	1088	+194
Sanguo zhi yanyi 三國志演義	1333	1300–1400	-
Shui hu zhuan 水滸傳	2626	ca. 1320–1372	+1.254
Jin ping mei 金瓶梅	2785	ca. 1596–1610	+1.175
Hong lou meng 紅樓夢	2854	ca. 1750	ca. +1.104
Ru lin wai shi 儒林外史	2609	1750	+940
Xi you ji 西遊記	2266	1592	+674

Ouyang Xiu 歐陽修 (1007–1072) als konzisere Neufassung der *Jiu Wudai shi* 舊五代史 (damals 五代史) verfasst.[263] Ouyang Xiu, der auch einer der Verfasser des im Modell ebenfalls als „zu alt“ geschätzten *Xin Tang shu* 新唐書 ist,[264] orientierte sich dabei stilistisch stark an antiken Vorbildern.[265]

Auf der anderen Seite lassen sich im Modell stark *zu neu* datierte Texte nicht mit derselben Kohärenz erklären. *Qing shi gao* und *Jiu Wudai shi* sind sehr zeitnah nach Ende des beschriebenen Zeitraums entstanden, bei dem 226 Jahre zu neu datierten *Chen shu* liegen aber fast 50 Jahre zwischen dem Ende der Chen-Dynastie (557–589) und der Fertigstellung im Jahr 636, sehr ähnlich verhält es sich bei dem 156 Jahre zu neu datierten *Bei Qi shu* 北齊書. Beide Texte sind überdies deutlich kürzer und enthalten dadurch deutlich weniger *types* als der Durchschnitt des Korpus. Grundsätzlich sind die meisten Abweichungen normal verteilt (Abb. 6.34, oben rechts). *Shiji* und *Qing shi gao* haben als jeweils ältester und neuester Text den größten Einfluss auf Steigung und Achsenabschnitt der Regressionsgeraden, wohingegen die größten Residuen durch ihre Lage im mittleren Beobachtungsbereich verhältnismäßig wenig Einfluss auf das Modell nehmen können (Abb. 6.34, unten). Der Cook'sche Abstand aller Beobachtungen liegt unter 0,5, so dass keine zu starke Hebelwirkung auf das Modell besteht.

Auch wenn sich für die Abweichungen im gegebenen Regressionsmodell stimmige Erklärungen finden lassen, muss vermutet werden, dass die starke Korrelation zwischen *AYL* und Entstehung der zu datierenden Texte nur durch den ebenfalls stark temporal konnotierten Inhalt der *zhengshi* möglich wird. Zudem muss eine Überanpassung der berechneten Modelle auf das gegebene Textkorpus geprüft und infrage gestellt werden, ob das *AYL* überhaupt für die Datierung anderer Text(gattungen) infrage kommt. Aufgrund der geringen Größe des Korpus kann auch innerhalb der *zhengshi* keine sinnvolle Aufteilung in Test- und Trainingsdaten gemacht werden, die Grundvoraussetzung für eine solide Evaluierung der *AYL*-Projektion als Datierungsmethode wäre.

263 Siehe Davis 2004, S. xlvii.

264 Im Ggstz. zur *XWDS* war die Kompilation des *Xin Tang shu* allerdings keine private Unternehmung, sondern wurde gemeinsam mit Song Qi 宋祁 (998–1061) und weiteren Gelehrten auf Geheiß des Hofes kompiliert. Siehe auch Wilkinson 2000, S. 820.

265 Ouyang Xiu wird dabei eine „literary revolution to replace the awkward ‚contemporary‘ prose current in his day with the ‚classical‘ style of the Spring and Autumn period“ nachgesagt. Siehe Davis 2004, S. xliv.

Versucht man mit der ermittelten Funktion einige eklektisch ausgewählte Texte unterschiedlicher Genres zu datieren (Tabelle 6.15), ergibt sich ein sehr durchwachsenes Bild. Für klassische bzw. schriftsprachliche Texte lassen sich halbwegs sinnvolle Ergebnisse erzielen, die zutreffend sind oder im erwartbaren Rahmen abweichen. So deutet sich ein klarer, richtiger Unterschied zwischen den beiden daoistischen Klassikern *Zhuangzi* 莊子 und dem deutlich älteren *Dao de jing* 道德經 an. Das *Zhongjing* stammt zwar nicht – wie mittels *AYL*-Datierung eingeordnet – aus der Han-Zeit, doch entspricht dies der traditionellen Einordnung.[266] Das Anfang des 6. Jh. zusammengestellte *Wenxuan* 文選 wird zwar auf das Jahr 14 datiert, enthält jedoch Texte aus der Zeit von ca. 300 v. u. Z. bis zum Ende des 5. Jh. Das *Zi zhi tong jian* 資治通鑑 sollte als historiographischer Text besonders gut zu den verwendeten Trainingsdaten des *zhengshi*-Korpus passen, jedoch wird eine um mehr als 400 Jahre zu frühe Datierung projiziert. Eine Ursache dafür ist sicherlich, dass zwei zeitliche Aspekte hier untrennbar ineinander greifen. Zwar wurde der Text im 11. Jh. verfasst, das Vokabular ist aber geprägt durch den langen Zeitraum, über den geschrieben wird: vom 4. Jh. v. u. Z. bis zur zweiten Hälfte des 10. Jh. Wie die Beispiele der *xiaoshuo* 小說 aus dem 14.–18. Jh. zeigen (*Hong lou meng* 紅樓夢, *Shui hu zhuan* 水滸傳, *Jin ping mei* 金瓶梅, *Ru lin wai shi* 儒林外史 und *Xi you ji* 西遊記), werden umgangssprachlichere Texte um viele Jahrhunderte zu neu datiert. Lediglich die Datierung des Romans *Sanguo zhi yanyi* 三國志演義 wirkt zutreffend.[267] Auch hier wirkt sicherlich – diesmal positiv – ein Effekt der Zeit, über die geschrieben wird – das Ende der Han-Zeit und die Zeit der Drei Reiche (ca. 208–280) – auf das Datierungsergebnis ein.

Die Erweiterung des für die *AYL*-Funktion zugrunde gelegten Textkorpus um weite Teile des Loewe-Korpus[268] kann genutzt werden, um zu zeigen, dass für klassische und schriftsprachliche Texte auch dann ein guter Zusammenhang zwischen *AYL* und Textentstehung hergestellt werden kann, wenn keine Beschränkung auf historiographische Texte besteht. Abb. 6.35 zeigt die zugehörige Regressionsgerade.

Da das Loewe-Korpus zahlreiche deutlich kürzere Texte enthält, konnte hier die beste Korrelation mit $R = 0,93$ bei Betrachtung von 90 % der häufigsten 2–4-Zeichen Lexeme hergestellt werden.[269] Trotzdessen fällt der Standardfehler mit 221 [Jahren] deutlich höher aus als bei einem auf *zhengshi* spezialisierten Modell.

6.3.2 *AYL* mit unterschiedlichen Korpora

Die Erkenntnisse aus 6.3.1 suggerieren, dass die besten Ergebnisse erzielt werden können, wenn für unterschiedliche Textgattungen jeweils eigene Trainingsdaten verwendet werden. Um dies fundiert zu evaluieren, sind verschiedene größere, diachrone Textkorpora erforderlich, z. B. die Abteilungen des *Xu xiu si ku quan shu* 續修四庫全書.[270] Diese eignen sich – wie festgestellt werden musste – als Korpus für die Evaluation von Datierungsmethoden allerdings nur bedingt, da zahlreiche spätere Ausgaben antiker Texte enthalten sind.[271]

266 Die Datierung des *Zhongjing* 忠經 wird in Abschnitt 6.2.4, ab S. 6.2.4, diskutiert. Der Text wird traditionell einem Ma Rong 馬融 (79–166) zugeschrieben, dies gilt aber als widerlegt.

267 Hier wurden die sechs „klassische chinesische Romane" nach der Definition in Hsia Chih-tsing 夏志清 1968, ausgewählt. Zur Analyse wurden hier die *Project Gutenberg*-Versionen verwendet.

268 Das Loewe-Korpus enthält digitale Fassungen der Texte, die in der von Michael Loewe herausgegebenen Bibliographie *Early Chinese Texts* vorgestellt werden. Siehe Kapitel 4.2, siehe auch Loewe 1993; siehe auch T. Schalmey 2009, S. 104–106.

269 Die Gleichung der Regressionsgeraden lautet: $Y_{proj} = 2,57 \times AYL + 78,6$.

270 *XXSKQS*.

271 Siehe u. a. Kapitel 6.1.3, S. 174.

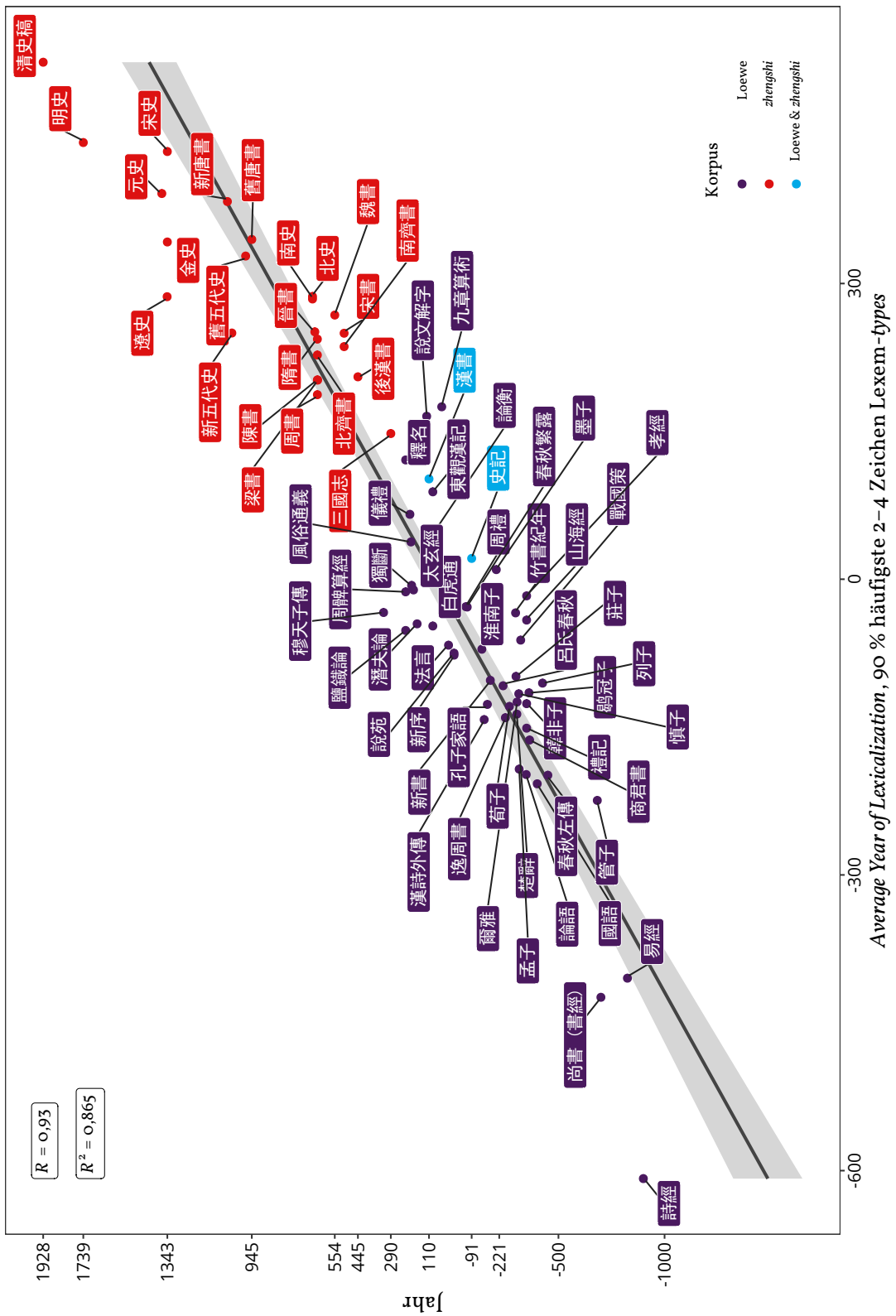


Abbildung 6.35 Korrelation Veröffentlichung Loewe / *zhengshi*, AYL mit 90 % 2–4 Zeichen Lexem-*types*

Obwohl – wie auch schon bei den *zhengshi* – die Problematik der Vermischung der Aspekte Entstehungszeit und Textinhalt besteht, lohnt sich aber – in Ermangelung geeigneterer Korpora – ein Blick auf die Verwendung des *AYL* im Kontext des *Difangzhi* 地方誌-Korpus[272] (*DFZ*).

Abb. 6.36b stellt ein wie in 6.3.1 optimiertes Regressionsmodell für 432 aus diesem Korpus zufällig ausgewählte Texte dar.[273] Da nur max. 3-Gramm-Daten zur Verfügung stehen, können nur Lexeme mit 2–3 Zeichen Länge betrachtet werden. Die stärkste Korrelation ergibt sich dabei mit ca. 70–90 % der häufigsten Lexem-*types* der Texte.[274] Im Gegensatz zu den *zhengshi* sind auch deutlich kürzere Texte mit teils nur wenigen hundert *types* enthalten. Bei Gegenüberstellung mit einem *zhengshi*-Modell (Abb. 6.36a) mit identischen Parametern, ergibt sich für die *DFZ* eine deutlich flachere Regressionsgerade mit höherem Achsenabschnitt. Als Referenz ist der relevante Abschnitt der Regressionsgeraden aus Abb. 6.36a ohne Datenpunkte in Abb. 6.36b erneut eingezeichnet.
Die gestrichelten grünen Linien geben in beiden Graphiken einen Toleranzbereich von 100 Jahren an. Datenpunkte innerhalb dieses Bereichs würden „richtig" datiert, wenn eine Genauigkeit von ±100 Jahren angestrebt wird.

Während für die *zhengshi* das lineare Modell mit sehr wenigen Datenpunkten und einem sehr langen Betrachtungszeitraum von 2.019 Jahren weiter sehr gut zu funktionieren scheint, ist in der rechten Graphik eine lineare Tendenz bei deutlich breiterer Streuung erkennbar. Bei einer Gruppe von *zhengshi*-Texten, die alle im Zeitraum zwischen 635 u. 659 herausgegeben wurden, ist allerdings eine ähnliche Streuung zu beobachten: *Liang shu* 梁書, *Chen shu* 陳書, *Bei Qi shu* 北齊書, *Zhou shu* 周書, *Sui shu* 隋書, *Jin shu* 晉書, *Nan shi* 南史 und *Bei shi* 北史 wurden in einem relativ kurzen Zeitraum von 25 Jahren während der Tang 唐 veröffentlicht. Die dem Modell angepassten Werte für die *AYL*-Datenpunkte in Abb. 6.36a für diese Texte ergeben einen Bereich von über 380 Jahren (501–883). Der Standardfehler der Regression des *zhengshi*-Modells beträgt dabei 171 [Jahre], für die *DFZ* ist er mit 95 trotz der breiten Streuung insgesamt deutlich niedriger. Die größten Abweichungen sind bei beiden Korpora sehr ähnlich: Die Geschichte der Liao (*Liao shi* 遼史) aus dem *zhengshi*-Korpus würde 364 Jahre zu früh auf das Jahr 980 datiert, die laut *DFZ*-Metadaten 1341 veröffentlichte Chronik der Präfektur Kunshan (*Kunshan qun zhi* 崑山郡志) 377 Jahre zu spät auf das Jahr 1718.

Bei mit dem *DFZ*-Korpus berechneten Modellen ist die Korrelation zwischen Veröffentlichung der Texte und *AYL* deutlich geringer, als dies bei den *zhengshi* der Fall war. Hierin spiegelt sich nicht nur die angesprochene Streuung mit deutlich mehr Texten aus einem deutlich kürzeren Betrachtungszeitraum wider, sondern auch die quasi nicht vorhandene Verortung der Texte im *DHYDCD* und eine hohe stilistische Rigidität der Textgattung.

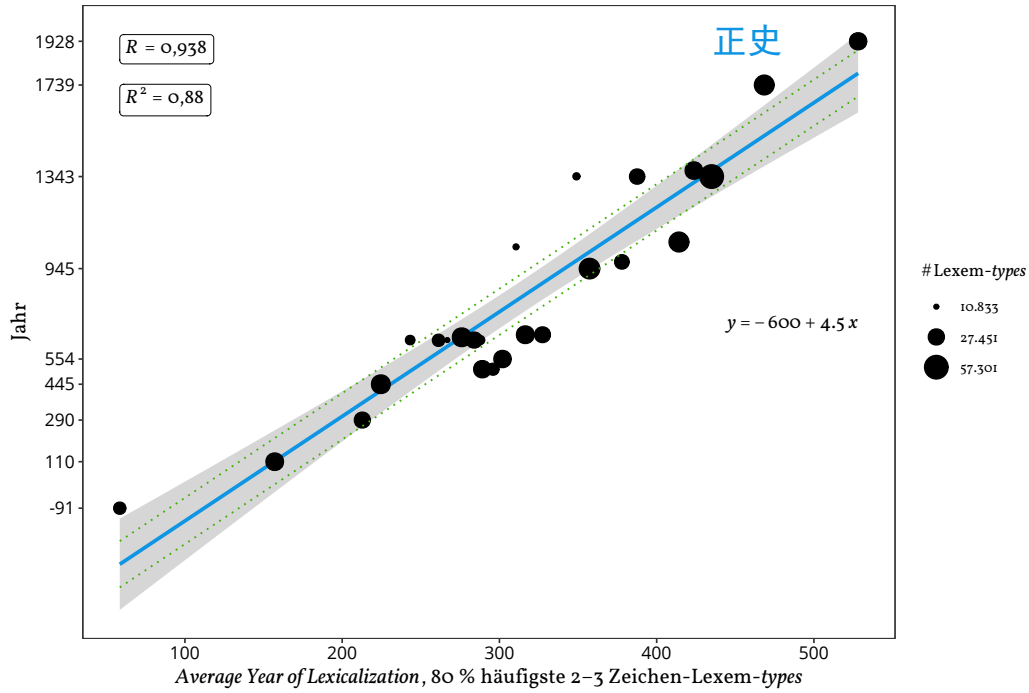
Die Größe des Korpus erlaubt es, das *AYL* als Datierungsmethode mit separaten Trainings- und Testdaten einem Praxistest zu unterziehen, der zumindest ansatzweise einen Vergleich mit den in den Kapiteln 6.1 und 6.2 vorgestellten Methoden ermöglicht.

Datiert man anhand der mit 432 *DFZ*-Texten trainierten Funktion aus Abb. 6.36b ($y = 1062{,}9 + 1{,}6 \times AYL$) die 216 Testdatensätze aus Kapitel 6.1.1, werden mit einem *MAE* von 94 Jahren bei einer

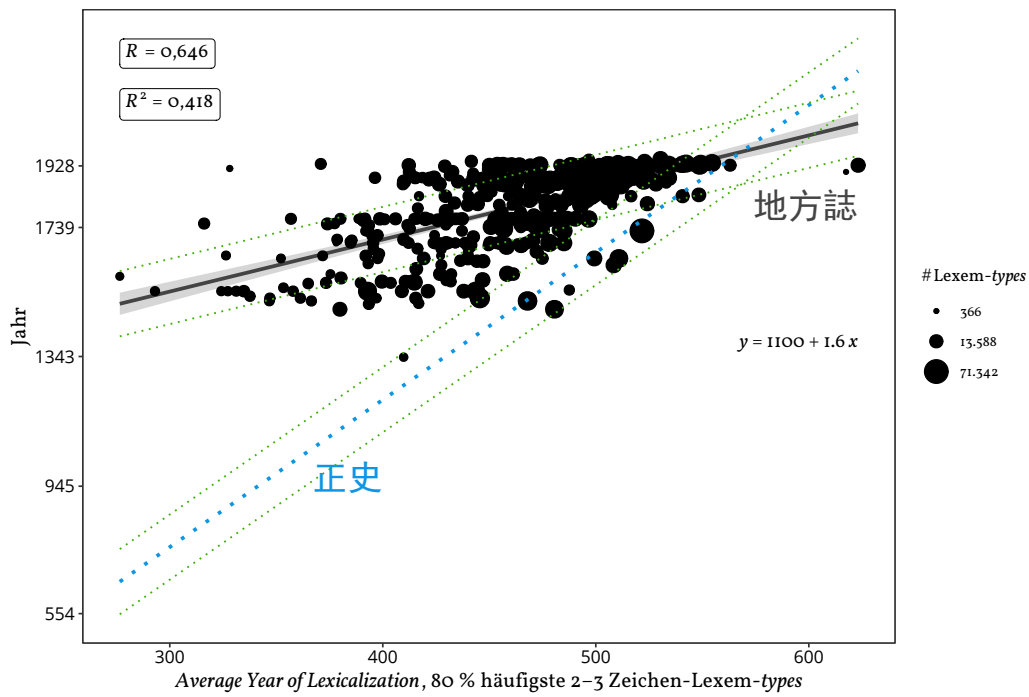
272 *DFZ*.

273 Um die Vergleichbarkeit der Ergebnisdaten zu erhöhen, wird der Testdatensatz aus Kapitel 6.1.1 mit 216 Texten ausgeschlossen.

274 Dass bei Verwendung von 15 % der häufigsten *types* ein schwaches *R* von nur 0,35 erreicht wird, ist sicherlich auf die deutlich geringere Länge einiger Texte zurückzuführen. Werden nur die Beobachtungen zu Texten mit min. 20.000 *types* berücksichtigt, erhöht sich der Korrelationskoeffizient *R* auf den – immer noch schwachen – Wert von 0,583.



(a) *zhengshi* 正史



(b) 432 *DFZ* 地方誌

Abbildung 6.36 Vergleich linearer Modelle, 80 % häufigste 2–3-Zeichen-Lexeme

Toleranz von ±100 Jahren 60,9 % der Texte korrekt datiert (Abb. 6.37), bei einer Toleranz von genau einem Jahrhundert (±50 Jahren) sind es immerhin noch 31 %.

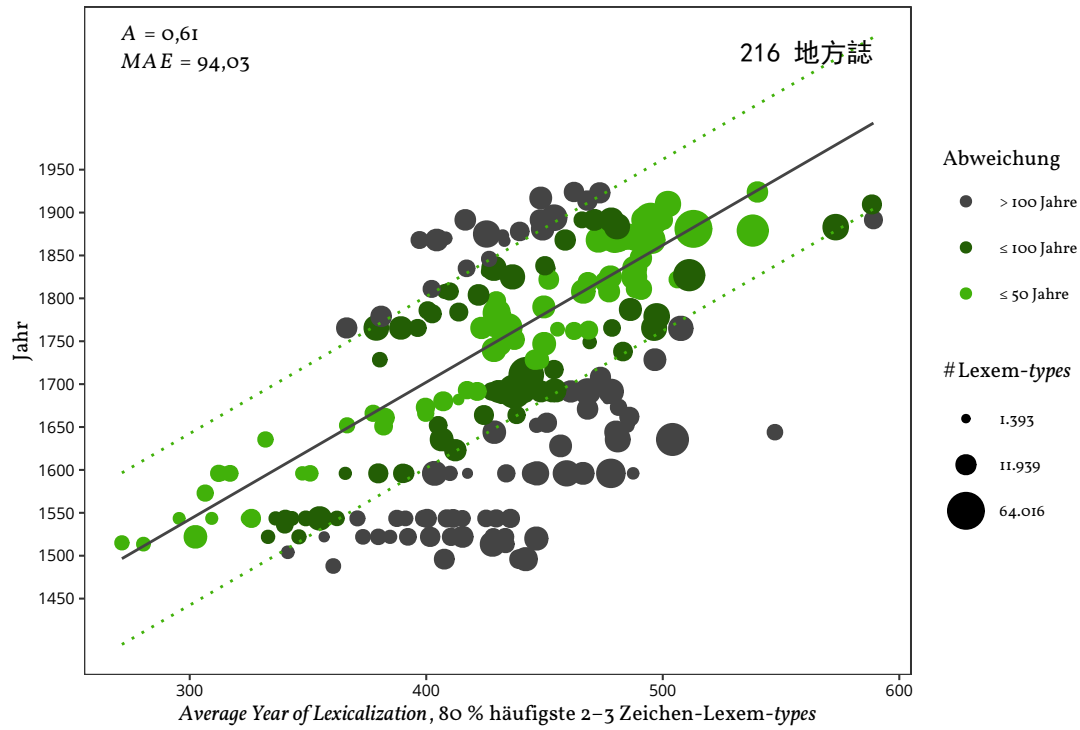


Abbildung 6.37 Datierungsergebnis *Difangzhi* mit 80 % häufigsten 2–3-Zeichen-Lexemen

Ordnet man dieselben *DFZ*-Texte unter Verwendung der Funktion aus dem *zhengshi*-Vergleichsmodell ein, würden bei einer durchschnittlichen Abweichung (*MAE*) von 360 Jahren nur 7,9 % der Texte bei einer Toleranz von ±100 Jahren korrekt zugeordnet.[275]

Für die Datierung von Texten anderer Genres bzw. außerhalb des Trainingskorpus ist ein lineares *AYL*-Modell also ungeeignet. Das gilt umso mehr, wenn mit schriftsprachlichen Trainingsdaten literarische bzw. v. a. umgangssprachliche Texte datiert werden. Das *AYL* kann ohne passende Trainingsdaten für das Chinesische also eher als eine Art relativer Zeit-Stil-Indikator eingesetzt werden. Die Verwendung mit dem Zweck der absoluten Textdatierung scheint nur sinnvoll, wenn Texte desselben Genres untereinander verglichen werden bzw. diachrone Trainingsdaten für genau diesen Texttypus vorliegen. Die Datierung mithilfe statistischer Sprachmodelle ist dann aber als etwas aussichtsreicher anzusehen.

275 Eine identische *Accuracy* bei einem *MAE* von 319 Jahren ergibt sich bei Anwendung der *zhengshi*-Funktion für die 432 Texte in Abb. 6.36b. Die korrekt zugeordneten Texte sind dabei zumeist verhältnismäßig lang, mit einer großen Anzahl an *types*. Da der Ausschluss kurzer Texte aber keinen wesentlichen Einfluss auf der Berechnung des *DFZ*-Modells hat, lassen sich daraus keine relevanten Schlüsse ziehen.

6.4 Untersuchte Datierungsmethoden im Überblick

In diesem Abschnitt werden die wesentlichen Unterschiede der in den Kapiteln 6.1, 6.2 und 6.3 untersuchten Methoden zur Textdatierung im Hinblick auf Potenzial und Limitationen bei der Betrachtung schriftsprachlicher chinesischer Texte zusammenfassend verglichen. Einige Ergebnisse werden – soweit möglich – anhand von *Accuracy* und *MAE* gegenübergestellt.

Bei der Datierung mithilfe statistischer Sprachmodelle (*Statistical Language Models*, *SLM*s) können Texte anhand unterschiedlicher Ähnlichkeitsmaße mit anderen Texten, oder mit aggregierten *chronons* verglichen werden. Beides funktioniert für schriftsprachliches Chinesisch grundsätzlich gut, wie in unterschiedlichen Experimenten gezeigt werden konnte. Von mehreren untersuchten Ähnlichkeitsmaßen hat sich – neben der *KLD* – die *NLLR* als am besten geeignet erwiesen. Zur Behandlung von *unseen events* hat sich ein einfaches *Smoothing* als am effizientesten herausgestellt. Dabei wird angenommen, dass ein *unseen event* weniger häufig ist als das seltenste *type* desjenigen *chronons* mit den meisten *types*.[276]

Zur Verwendung von *SLM*s sind für den gesamten Zeitraum, aus dem Texte datiert werden sollen, umfangreiche Trainingsdaten erforderlich. Die besten Ergebnisse werden dann erzielt, wenn die Trainings- und Testdaten aus demselben, stilistisch homogenen Korpus stammen, wie das bei den *difangzhi* 地方誌 (*DFZ*) der Fall ist. Da bis dato kaum entsprechende diachrone, schriftsprachliche Korpora digital vorliegen, ist die Nutzung von *SLM*s in der Praxis der linguistischen Datierung für schriftsprachliches Chinesisch zunächst nur eingeschränkt möglich. Eine – wenngleich grobe – Datierung mit genreübergreifenden Trainingsdaten ist jedoch möglich. Hierfür konnten die Belegstellen aus dem *DHYDCD* als Trainingsdaten verwendet werden.

Die in dieser Arbeit vorgestellte Datierung anhand von temporalen Textprofilen stützt sich im Wesentlichen auf diachrone Lexikalisierungsdaten, die in Kapitel 5.5[277] ebenfalls aus dem *DHYDCD* extrahiert wurden. Damit werden – grob vereinfacht – Lexeme, d. h. lexikalisierte 2–4-Zeichen-Kombinationen mit dem Jahr in Verbindung gebracht, in welchem ihr *Locus classicus* veröffentlicht wurde. Ein temporales Textprofil zeigt die Zuordnung aller in einem Text enthaltenen Lexeme zum Jahrhundert ihrer frühesten Verwendung als Balkendiagramm. Zusätzlich können Personen- und Ortsnamen, sowie temporale Ausdrücke in die Darstellung miteinbezogen werden. Um die zugrunde liegenden Daten für eine automatisierte zeitliche Einordnung des Textes zu nutzen, sind ebenfalls Trainingsdaten erforderlich. Die Abhängigkeit von einzelnen Genres oder Zeiträumen ist dabei aber deutlich geringer. Im Gegensatz zu den übrigen untersuchten Methoden erfordert die Interpretation der Textprofile nur geringe mathematische bzw. statistische Vorkenntnisse. Durch die transparente Darstellung ist die Nutzung der Profile für die Suche nach Hinweisen für die Textdatierung auch dann noch möglich, wenn ein statistischer Ansatz wenig aussichtsreich ist. Dies kann vor allem dann hilfreich sein, wenn der zu datierende Text sehr kurz ist, z. B. für Gedichte, oder bewusst in einem klassischen Stil verfasst wurde. Dieser lexikographische Ansatz spiegelt dabei eine traditionelle Methodik der historischen Linguistik wider – das Aufspüren von Anachronismen.

In Kapitel 6.3 wurde zuletzt damit experimentiert, die Lexikalisierungsdaten auf eine einzelne Messgröße, das durchschnittliche Jahr der Lexikalisierung der in einem Text enthaltenen Lexeme (*Average Year of Lexicalization*, *AYL*), zu reduzieren. Als Datengrundlage dienen auch hier die diachronen Lexikalisierungsdaten aus dem *DHYDCD*. Es konnte gezeigt werden, dass bei

276 Je nach Art und Umfang der Trainingsdaten sollte der *Smoothing*-Parameter λ zur Bestimmung der optimierten Häufigkeit des *unseen event* angepasst werden. Siehe dazu die Abschnitte 6.1.1, ab S. 164, sowie 6.1.3, ab S. 171.

277 Siehe ab S. 120.

Betrachtung einer homogenen, diachronen Textreihe ein linearer Zusammenhang zwischen der Entstehung der Texte und dem *AYL* besteht. Für seine Berechnung muss ein zu optimierender Anteil der häufigsten, im Text enthaltenen 2–4-Zeichen-Lexeme betrachtet werden. Eine starke Abhängigkeit besteht dabei erneut zu den stilistischen Eigenschaften der zu datierenden Texte. Für eine praktische Nutzung ist das Vorhandensein passender Trainingsdaten daher zwingend erforderlich. Die Anzahl der dafür erforderlichen Texte sollte – anders als bei der Verwendung von *SLM*s – allerdings überschaubar sein. Mit einer Linearregression kann anhand dieser Trainingsdaten eine Funktion berechnet werden, die das *AYL* auf die geschätzte Entstehung des Textes projiziert. Eine stabile Korrelation zwischen Textentstehung und *AYL* kann aber nur für längere Texte ab etwa 10.000 Zeichen (ca. 10 Seiten)[278] erreicht werden. Sind Genre und grobe Entstehungszeit eines zu datierenden Textes unbekannt oder stehen keine ausreichenden Trainingsdaten zur Verfügung, kann das *AYL* lediglich als grober Zeit-Stil-Indikator im direkten Vergleich zwischen ähnlichen Texten betrachtet werden.

Bei allen oben erläuterten Methoden wird von der Datierung von *Plain Text* bzw. *n*-Gramm Häufigkeitslisten ausgegangen. In allen Fällen sollte geprüft werden, ob die verwendeten Ausgaben Kommentare enthalten – auch wenn diese stilistisch nicht unbedingt moderner sind als der zu datierende Haupttext, können sie einerseits einen bedeutend größeren Anteil einnehmen als der eigentliche Text, andererseits suchen die Kommentator:innen oft den Text mit zeitgenössische(re)n Ausdrücken zu erklären, was grundsätzlich äußerst problematisch für den Versuch einer rein linguistischen bzw. lexikographischen Datierung ist.[279] Für eine philologische Untersuchung hingegen kann die Kommentartradition durchaus zur Beleuchtung datierungsrelevanter Aspekte beitragen, da sie Aufschluss über die Überlieferungsgeschichte geben kann. Auch Kommentare können dabei natürlich – wie ein Text selbst – eine spätere Fälschung sein.

Dass alle beschriebenen Datierungsmethoden anhand des *DFZ* Korpus getestet wurden, erlaubt einen groben Vergleich der Ergebnisse. In Tabelle 6.16 sind einige der in den Kapiteln 6.1, 6.2 und 6.3 durchgeführten Experimente gegenübergestellt. Dafür werden jeweils Ergebnisse mit der besten erzielten *Accuracy* (A_{max}) ausgewählt und der zugehörige *mean average error* (*MAE*) angegeben.
Auf den ersten Blick sind die Ergebnisse, die sich mit *SLM*- und lexikographischer Datierung erzielen lassen, sehr ähnlich. In beiden Fällen kann eine *Accuracy* von ca. 50 % erreicht und die Texte mit einem *MAE* von ca. 50–80 Jahren datiert werden. Während *SLM*s dafür auch Einzelzeichen und Worthäufigkeiten betrachten, werden bei der Erstellung von temporalen Profilen nur Lexeme mit einer Mindestlänge von 2 Zeichen berücksichtigt, unabhängig von ihrer Häufigkeit im untersuchten Text.

Eine hohe *Accuracy* konnte mit *SLM*s nur dann erreicht werden, wenn spezifische Trainingsdaten zur Erzeugung der Modelle verwendet wurden. Auch der Algorithmus für die Datierung mittels temporalen Textprofilen basiert zwar auf Beobachtungen an Trainingsdaten aus demselben Korpus, jedoch ist die Datierung hier offen über den Zeitraum von 700 v. u. Z. bis ins 20. Jh. angelegt. Bei der Datierung mit *SLM*s ist der Ergebnisraum auf den Zeitraum der Trainingsdaten begrenzt, von 1475–1925. Bei 17 überlappenden 50-Jahre-*chronons* ergibt sich eine

278 Zum Vergleich werden Seiten im Format A4 mit einer Schriftgröße von 12 pt herangezogen.
279 Während in gedruckten bzw. formatierten Ausgaben oder strukturierten Dateiformaten Kommentare in der Regel z. B. durch eine kleinere Schriftgröße bzw. entsprechende Markierungen klar erkennbar sind, ist die Trennung von Kommentaren und Haupttext in *Plain Text*-Formaten erschwert.

Tabelle 6.16 Vergleich der in Kapitel 6 vorgestellten Methoden anhand des *DFZ*-Korpus

Methode	A_{max}	*MAE*	Details	*T*	Training
Statistical Language Models					
50 Jahre *Chronon-SLM*	60,6	40,3	◆ *NLLR*, 1–2 Zeichen-Lexeme & *temporal expressions*	1475–1925	spezifisch, *DFZ*
50 Jahre *Chronon-SLM*	64,4	42,5	◆ *NLLR* * *TE*, 1–2 Zeichen-Lexeme & *temporal expressions*	1475–1925	spezifisch, *DFZ*
50 Jahre *Dokumenten-SLM*	46,3	59,3	◆ *NLLR*, 1–2 Zeichen-Lexeme & *temporal expressions*	1475–1925	spezifisch, *DFZ*
100 Jahre *chronon-SLM*	19,9	136	◆ *CS* * *tf-idf*, 1–2 Gramme	ca. 700 v. u. Z.–2000	unspezifisch, *DHYDCD*
100 Jahre *chronon-SLM*	16,2	140,6	◆ *NLLR* * *TE*, 1–2 Zeichen-Lexeme & *temporal expressions*	ca. 700 v. u. Z.–2000	unspezifisch, *DHYDCD*
Lexikographisch / datenbankgestützt					
100 Jahre Neologismusprofil, korrigierte Gewichtung	47,2	83,9	■ 2–3 Zeichen-Lexeme	ca. 700 v. u. Z.–2000	spezifisch
100 Jahre Temporales Profil, korrigierte Gewichtung	62,5	72,1	■■ 2–3 Zeichen-Lexeme, 3 Z. Namen,	ca. 700 v. u. Z.–2000	spezifisch
100 Jahre Temporales Profil, korrigierte Gewichtung	75,9	63,2	■■■ 2–3 Zeichen-Lexeme, 3 Z. Namen, *temporal expr.*	ca. 700 v. u. Z.–2000	spezifisch
Newest dates in text	88	57,5	■ 4+ *temporal expressions*	ca. 220 v. u. Z.–1912	nein
***Average Year of Lexicalization* / datenbankgestützt**					
$AYL_{0,8}^{2-3}$, Toleranz ±50 Jahre	31	94	▲ 80 % 2–3 Zeichen Lexeme; kontinuierlich	∞/ ca. 100 v. u. Z.–4000	spezifisch, 432 *DFZ*

Baseline-Wahrscheinlichkeit von etwa 11 %, mit der ein Zufallsgenerator die Texte korrekt zuordnen würde, bei temporalen Textprofilen wären es – ohne Überlappungen – lediglich 3,7 %. Andererseits kann dabei – auch aufgrund der Ungenauigkeit der auf den *attestations* im *DHYDCD* aufgebauten diachronen Lexemdatenbank – bislang nur eine Genauigkeit von 100 Jahren angestrebt werden.

Wird der Ergebnisraum für die *SLM*-Datierung ebenfalls auf 700 v. u. Z. bis ins 20. Jh. erweitert, indem Sprachmodelle mit 100-Jahre-*chronons* aus den *DHYDCD attestations* erzeugt werden, kann bei einer *Baseline* von etwa 4 % noch eine *Accuracy* von knapp 20 % erreicht werden, mit einem immer noch beachtlichen *MAE* von 136 Jahren. Ein überwiegender Teil der *DFZ* kann also auch ohne spezifische Trainingsdaten grob korrekt eingeordnet werden. Insgesamt erweist sich der Einsatz temporaler Profile bei der Datierung von Texten mit unspezifischen Trainingsdaten aber als robuster.[280]

Einen Sonderfall stellt die Datierung der *DFZ* auf Basis der spätesten enthaltenen *temporal expressions* dar. Eine Datierung, die sich allein auf Zeitangaben im Text stützt, kann nur dann erfolgreich sein, wenn der zu datierende Text überhaupt konkrete Zeitangaben enthält, vor allem aber dürfen erzählte Zeit und Veröffentlichung nur unweit auseinanderliegen. Diese Bedingungen sind für die Lokalchroniken auch deshalb optimal erfüllt, da eine entsprechende Einschränkung bei der Auswahl von Test- und Trainingsdaten vorgenommen wurde.[281] Es muss davon ausgegangen werden, dass für kaum eine andere Textgattung auf diesem Weg vergleichbare Ergebnisse erzielt werden können. In Verbindung mit temporalen Textprofilen bietet die Analyse vorhandener Zeitausdrücke aber immer einen Mehrwert.

Die Ergebnisse der zeitlichen Einordnung mithilfe des *AYL* sind – trotz des spezifischen Trainings – eindeutig als schwächer zu bewerten. Das Potenzial dieser Herangehensweise liegt also eher in der relativen Einordnung von Texten einer diachronen Reihe, bzw. kann eine Datierungsfunktion auch dann trainiert bzw. optimiert werden, wenn Trainingsdaten nur in geringem Umfang zur Verfügung stehen.

Während die chronologische Einordnung schriftsprachlicher chinesischer Texte mit allen drei untersuchten Methoden grundsätzlich funktioniert, zeigen sich außerhalb der mit dem *DFZ*-Korpus durchgeführten Experimente immer wieder Limitationen einer linguistischen Datierung. Gerade statistische Methoden sind besonders anfällig für eine zu frühe Datierung von Texten, die in einem streng altertümlichen Stil (*guwen* 古文) verfasst sind, dessen Grammatik und Wortwahl für viele Textgattungen bis zum Ende der Kaiserzeit Vorbilcharakter genoss bzw. in Ausnahmefällen immer noch genießt.[282] Auch eine lexikographische Methodik kann nur dann Abhilfe schaffen, wenn der zu datierende Text überhaupt zeitgenössische Lexeme enthält – und die zugrundeliegende Datenbasis ausreichend umfangreich und genau ist.

6.4.1 *VisualTime — user interface* für Datierungsmethoden

Um die wesentlichen hier vorgestellten Methoden zur Datierung (schriftsprachlicher) chinesischer Texte für Sinolog:innen mit geringen computerlinguistischen Vorkenntnissen nutzbar zu machen, wird mit *VisualTime* ein *user interface* im Browser bereitgestellt.[283]

280 Siehe Abschnitt 6.2.5, ab S. 197, v. a. Tabelle 6.12, S. 207.

281 Siehe dazu Abschnitt 6.1.1, S. 158. Ziel dieser Einschränkung ist der Ausschluss von späteren Editionen bzw. die Vermeidung potenzieller Verzerrungen durch diese.

282 Vgl. z. B. Tai und M. K. M. Chan 1999, S. 229.

283 Eine aktuelle Version kann unter `https://visualtime.schalmey.de` getestet werden.

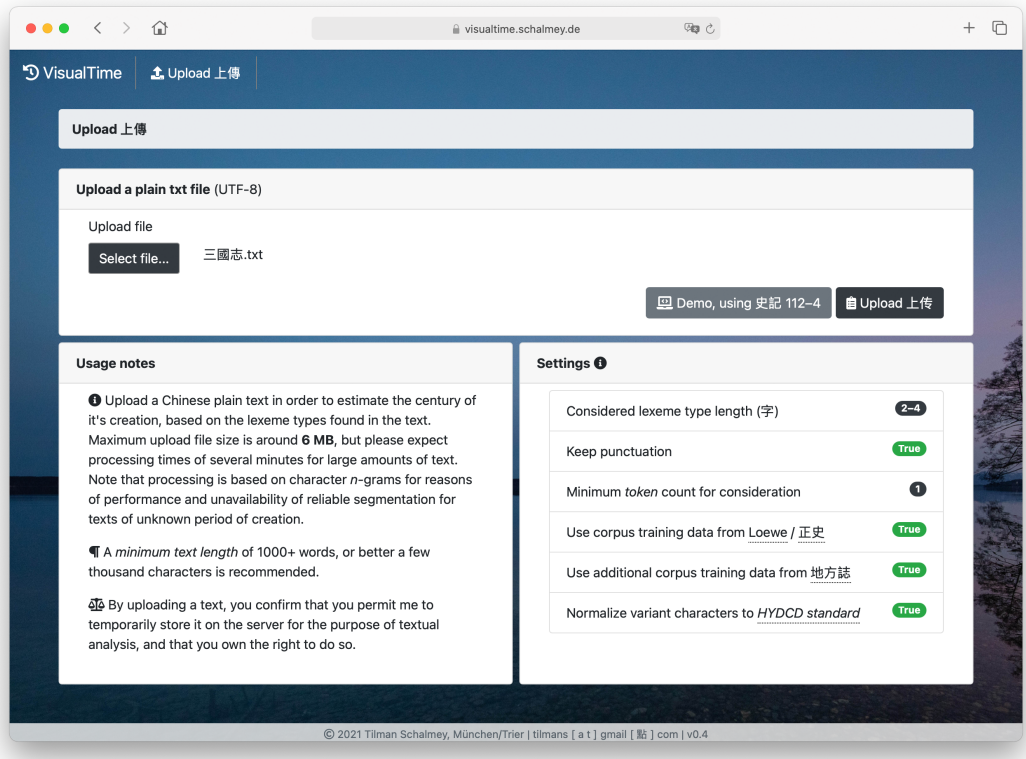


Abbildung 6.38 *VisualTime* Startseite – Datei auswählen und hochladen

Die für Kapitel 6.1, 6.2 und 6.3 entwickelten Komponenten und die dabei gewonnenen Erkenntnisse fließen dabei in eine Anwendung ein, der das *Python Framework Django*[284] zugrunde liegt. Die so erzeugte Benutzer:innenoberfläche ermöglicht den Upload einer *Plain Text* Datei, für deren Inhalt ein temporales Textprofil erzeugt wird. Zusätzlich erfolgt eine *SLM*-Klassifizierung sowie Berechnung des *AYL*. Das resultierende Neologismusprofil und die Ergebnisse der Datierung per Sprachmodell können im Detail erkundet und direkt miteinander verglichen werden.

VisualTime wird zunächst anhand eines Optimalbeispiels erläutert. Das *Sanguo zhi* 三國志 (*Chroniken der Drei Reiche*) ist in den *zhengshi* 正史 Trainingsdaten für die diachrone Lexemdatenbank und dem zur *AYL*-Projektion verwendeten Regressionsmodell enthalten. Zudem wird in 6.799 *DHYDCD*-Einträgen aus dem *Sanguo zhi* zitiert, so dass auch die *chronons* 200–300 und 250–350 des verwendeten temporalen Sprachmodells direkt von diesem Text geprägt sind. Durch diese „Schummelei“, die teilweise Identität von Trainings- und Testdaten, lässt sich die Ausgabe unter Optimalbedingungen veranschaulichen.

284 Django Software Foundation 2005–: *django – The web framework for perfectionists with deadlines*. URL: https://www.djangoproject.com/ (besucht am 12. 10. 2021), *Django* wird hier primär als *templating engine* zur Erzeugung der Seiten mit *HTML* & *CSS* verwendet, die Kernfunktionalität der Verwaltung von Datenmodellen wird hier (noch) nicht genutzt, stattdessen wird die in Kapitel 5.5 (ab S. 120) erzeugte Datenbank angebunden.

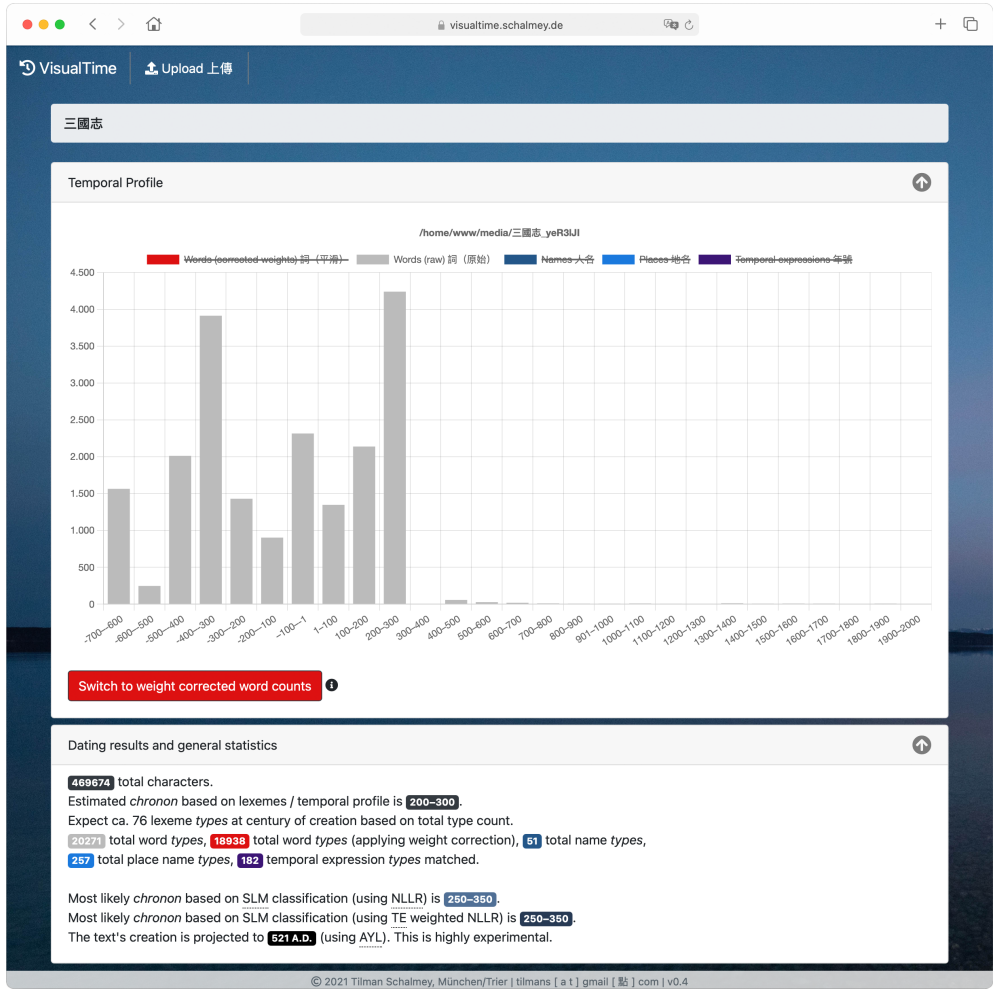


Abbildung 6.39 *VisualTime* Ergebnisseite

Allgemeine Statistiken und Temporales Textprofil

Auf der Ergebnisseite (Abb. 6.39) werden allgemeine Statistiken zum hochgeladenen Text und die Ergebnisse der unterschiedlichen Datierungsmethoden angezeigt. Durch die Verwendung von *Chart.js*,[285] einem *JavaScript Framework* zur Visualisierung von Daten im Browser bzw. innerhalb von Webseiten, kann ein interaktives temporales Profil des Textes mit an- und abklickbaren Lexemen, Namen, Ortsnamen und temporalen Ausdrücken dargestellt werden. Initial werden die im Text ausgemachten Lexeme jahrhundertweise in einem Balkendiagramm angezeigt, wahlweise kann zu einer Darstellung mit Gewichtungskorrektur gewechselt werden.[286] Da der betrachtete Text in den Trainingsdaten für die Lexemdatenbank enthalten ist, kann in beiden Darstellungen die Entstehung im 3. Jh. klar abgelesen werden – *types* aus späteren Jahrhunder-

285 Evert Timberg et al. 2013–: *Chart.js*. GitHub Repository. URL: `https://github.com/chartjs/Chart.js` (besucht am 12. 10. 2021).

286 In Abschnitt 6.2.1, ab S. 182, wird die Auswertung der dieser Darstellung zugrunde liegenden Daten ausführlich erläutert. Ab S. 185 wird auf das hier verwendete *Smoothing* eingegangen.

ten sind kaum enthalten. Entsprechend wählt auch der Datierungsalgorithmus mit 200–300 das korrekte *chronon*.[287]

Abbildung 6.40 *VisualTime* – Temporales Profil des *Sanguo zhi* mit flexibel aktivierbaren *types*

Durch Deaktivieren der Lexemanzeige (*word types*) bzw. Aktivierung der übrigen *types* können Namen oder temporale Ausdrücke separat betrachtet werden (Abb. 6.40). Die im *Sanguo zhi* erkannten *temporal expressions* (unten rechts) zeigen deutlich, dass die im Text beschriebene Zeit primär ins 2.–3. Jh. fällt, was ebenfalls den Tatsachen entspricht. Deutlich problematischer ist die zeitliche Einordnung der im Text erkannten Personen- und Ortsnamen. Es kommt hier zu unterschiedlichen Arten von *false positives*, da Zeichenfolgen fälschlich als Namen erkannt werden können, Personen gleichen Namens abweichende biographische Daten haben, bzw. die als Datenquelle verwendete *CBDB* nicht ausreicht, um eine klare Interpretation zu ermöglichen.[288]

Um die *NER*-Ergebnisse dennoch sinnvoll nutzbar zu machen, kann durch Klick auf entsprechende *Tabs* auf der Ergebnisseite die Liste der erkannten Personennamen ausgeklappt und chronologisch, oder alternativ nach Häufigkeit der Nennung sortiert werden (Abb. 6.41).

Abbildung 6.41 *VisualTime* – Erkannte Namen im *Sanguo zhi* (Ausschnitt)

287 Siehe dazu Abschnitt 6.2.5, ab S. 197.
288 Siehe dazu auch Kapitel 4.7, ab S. 97, sowie Abschnitt 6.2.2, ab S. 189.

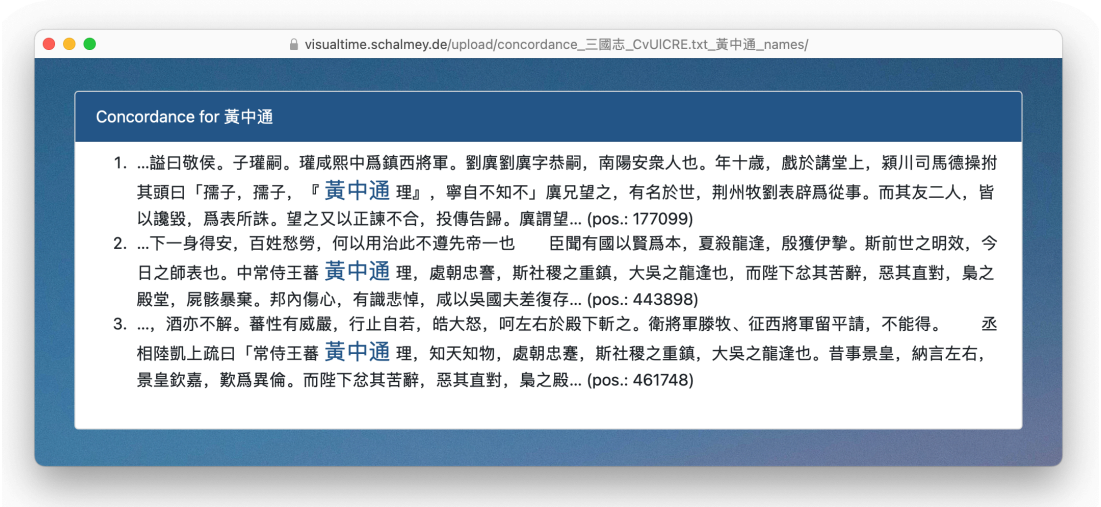


Abbildung 6.42 *VisualTime* – Anzeige aller Textstellen mit Huang Zhongtong 黃中通 im *Sanguo zhi*

Bei im Text häufigen Zeichenfolgen wie Zhuge Liang 諸葛亮 (181–234, 109 Nennungen) handelt sich mit höherer Wahrscheinlichkeit tatsächlich um die jeweilige Person. Zum Abgleich der Daten bzw. für die weitergehende Recherche ist zudem rechts die *CBDB* ID der angezeigten Einträge aufgeführt. Personen mit in der *CBDB* vollständigen biographischen Daten werden blau hervorgehoben.[289] Um zu verifizieren, ob im Text tatsächlich eine Person genannt wird und es sich nicht um Vorkommen einer zufällig mit dem Namen übereinstimmenden Zeichenkombination handelt, kann durch Klick auf den Namen eine Konkordanz der betroffenen Textstellen in einem neuen Fenster angezeigt werden (Abb 6.42).

Für Huang Zhongtong 黃中通 ist in der *CBDB* das sogenannte Indexjahr mit 1679 angegeben,[290] es muss sich hier also – trotz der drei Vorkommen – um *false positives* handeln. Alle drei Textstellen enthalten die Zeichenfolge *huang zhong tongli* 黃中通理, ein feststehender Ausdruck, der auf das *Yijing* 易經 (*Buch der Wandlungen*) zurückgeführt werden kann.[291]

Eine strukturierte Detailanzeige der im Text gefundenen *types* sowie von Vorkommen dieser *types* ist gleichermaßen auch für Lexeme und temporale Ausdrücke möglich.

289 Zu biographischen Angaben in der *CBDB* siehe auch Kapitel 4.7, S. 99.
290 Siehe *CBDB*, Zum Indexjahr in der *CBDB* siehe auch Kapitel 4.7, ab S. 99.
291 Die Farbe gelb in der Mitte steht für einen *junzi* 君子, der schnelle Auffassungsgabe und eine gute Urteilsfähigkeit aufweist. James Legge übersetzt „ 君子黃中通理“ aus dem Text zum Hexagramm *kun* 坤 mit „The superior man (emblemed here) by the yellow and correct (colour), is possessed of comprehension and discrimination.“ James Legge 1882: *The Yi King*. Übers. von James Legge. Sacred Books of the East XVI. Oxford: Clarendon Press, S. 421.

Datierung mit *SLM*

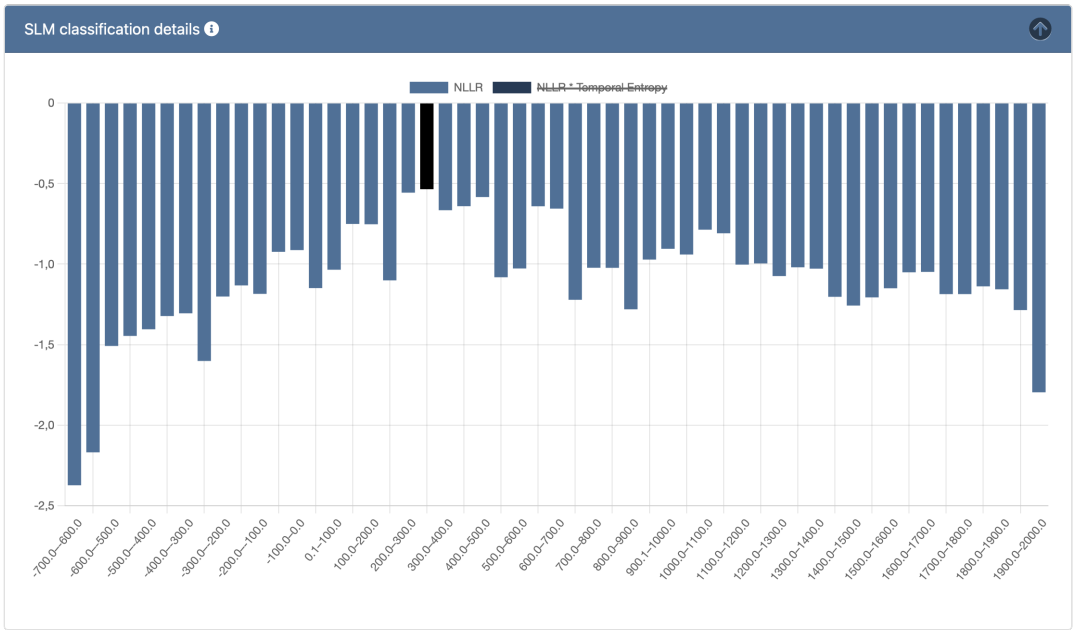


Abbildung 6.43 *VisualTime* – Anzeige *NLLR* des *Sanguo zhi* zu einzelnen *DHYDCD-chronons*

Wie in Abb. 6.39 zu sehen ist, wurde das *Sanguo zhi* mit einem temporalen *SLM* dem wahrscheinlichsten *chronon* 250–350 zugeordnet. Wie bereits angedeutet wird die Richtigkeit dieser Zuordnung stark dadurch begünstigt, dass Sätze aus dem *Sanguo zhi* 30 % der Trainingsdaten dieses *chronon*s ausmachen.[292] Um die Qualität bzw. Verlässlichkeit einer solchen Zuordnung zu analysieren, können die *NLLR*-Ähnlichkeiten des Texts zu allen *chronons* des Modells eingeblendet werden (Abb. 6.43).

Die Detailansicht zeigt eine tendenziell steigende *NLLR* von den frühesten *chronons* hin zur tatsächlichen Fertigstellung des Texts im Jahr 297. Die höchsten Werte erhalten dabei die beiden überlappenden *chronons* 200–300 und 250–350, letzteres in der Graphik farblich hervorgehoben. Für die späteren *chronons* ist wieder eine abnehmende Tendenz zu beobachten. Diese klar ablesbaren Tendenzen und die Nachbarschaft der ähnlichsten *chronons* sprechen für die Verlässlichkeit der Zuordnung.[293] In der Darstellung in Abb. 6.43 kann zu einem mit Temporaler Entropie gewichteten Modell umgeschaltet werden (ohne Abb.).[294]

Zur Veranschaulichung der Funktionsweise und Darstellungen sei zum Vergleich die Ausgabe in *VisualTime* für einen weiteren bekannten Text erläutert. Das *Sanguo zhi yanyi* 三國志演義 (*Die ausführliche und erläuterte Geschichte der Drei Reiche*) ist ein historischer Roman, der Luo Guanzhong

292 Die Trainingsdaten des *chronons* 250–350 setzen sich aus 22.452 Belegstellen aus dem *DHYDCD* zusammen, 6.799 davon sind dem *Sanguo zhi* entnommen. Siehe dazu auch Kapitel 5.6 (ab S. 137), sowie Abschnitt 6.1.3 (ab S. 171) zur Erzeugung und Verwendung des auf dem *DHYDCD* basierenden temporalen Sprachmodells.

293 Vgl. auch de Jong, Rode und Hiemstra 2005, S. 6: „A simple confidence measure for dating could be the relative distance between the score of the top-ranked time partition to the scores of the following ones. A more sophisticated measure could also take into account the level of timely scattering in the top-ranked partitions."

294 Siehe dazu Kapitel 3.3, S. 53.

羅貫中 (ca. 14. Jh.) zugeschrieben wird. Die älteste überlieferte Ausgabe stammt aus dem Jahr 1522.[295] Wie auch beim *Sanguo zhi* dienen die Geschehnisse zur Zeit der Drei Reiche als Grundlage, hier allerdings literarisch und mehr als 1.000 Jahre nach den historischen Ereignissen erzählt.

Anders als das *Sanguo zhi* ist das *Sanguo zhi yanyi* nicht Teil der Trainingsdaten. Zwar ist es im *DHYDCD* 1.644 mal als Beleg präsent, aber in der Datenbank fehlt eine Datierung des Texts, so dass weder das Sprachmodell noch die Lexemdatenbank davon beeinflusst sind.

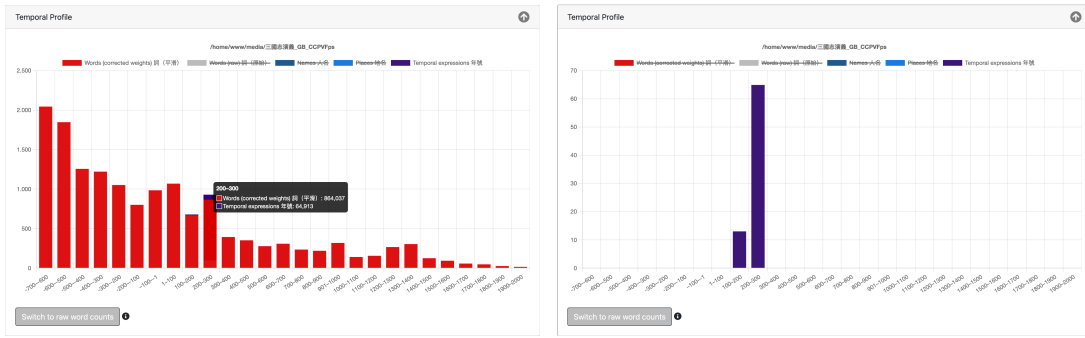


Abbildung 6.44 *VisualTime* – Profil des *Sanguo zhi yanyi*

Abb. 6.44 (links) zeigt das gewichtungskorrigierte Neologismusprofil des *Sanguo zhi yanyi*[296] mit eingeblendeten temporalen Ausdrücken. Eindeutig lässt sich nach Ausblenden der Lexem-*types* (Abb. 6.44 rechts) erkennen, welche Zeit in dem Text behandelt wird: das 2. und v. a. 3. Jh. Ganz anders als in der Darstellung des *Sanguo zhi* (Abb. 6.40) sind aber zahlreiche Lexeme enthalten, die erst nach dem 3. Jh. nachgewiesen sind. Ein deutliche Abnahme ist vom 14. hin zum 15. Jh. erkennbar, was für eine Entstehung des Texts im 14. oder Anfang des 15. Jhs. sprechen würde – zu Lebzeiten von Luo Guanzhong. Der Profil-Datierungsalgorithmus verortet den Roman in das *chronon* 1600–1700.[297]

Das *chronon* mit dem höchsten Wert für die *NLLR* ist 1550–1650, wobei hier ebenfalls die benachbarten *chronons*, v. a. 1600–1700, nahezu identische Werte erreichen (Abb. 6.45). Wie schon bei der Darstellung der *NLLR* für das deutlich ältere *Sanguo zhi* (Abb. 6.43) lässt sich ein zum Jahrhundert der Textentstehung steigender Wert erkennen, der anschließend (hier ab dem 18. Jh.) wieder abnimmt. Allerdings verläuft diese Zunahme ab dem 7. Jh. bereits sehr flach, was auf eine geringe Klarheit bzw. Verlässlichkeit der *SLM*-Datierung hindeutet.

295 Siehe Clemens Treter 2004: „Die Literatur der Ming- und Qing-Zeit". In: *Chinesische Literaturgeschichte*. Hrsg. von Reinhard Emmerich. Stuttgart: Metzler, S. 225–287, S. 239.

296 Verwendete Version: Luo Guanzhong 羅貫中 2007 [?1522]: *Sanguo zhi yanyi* 三國志演義 *(Romance of the Three Kingdoms)*. Project Gutenberg eBook. URL: https://www.gutenberg.org/ebooks/23950 (besucht am 28. 05. 2021).

297 Ohne Abb. Die Datierung erfolgt hier aufgrund der Linearregression auf die für das Jahrhundert der Textentstehung erwartete Anzahl *types* (60) auf Basis der Gesamtzahl erkannter Lexem-*types* (16.021, mit korrigierter Gewichtung 15.127). Siehe dazu Abschnitt 6.2.5, ab S. 197.

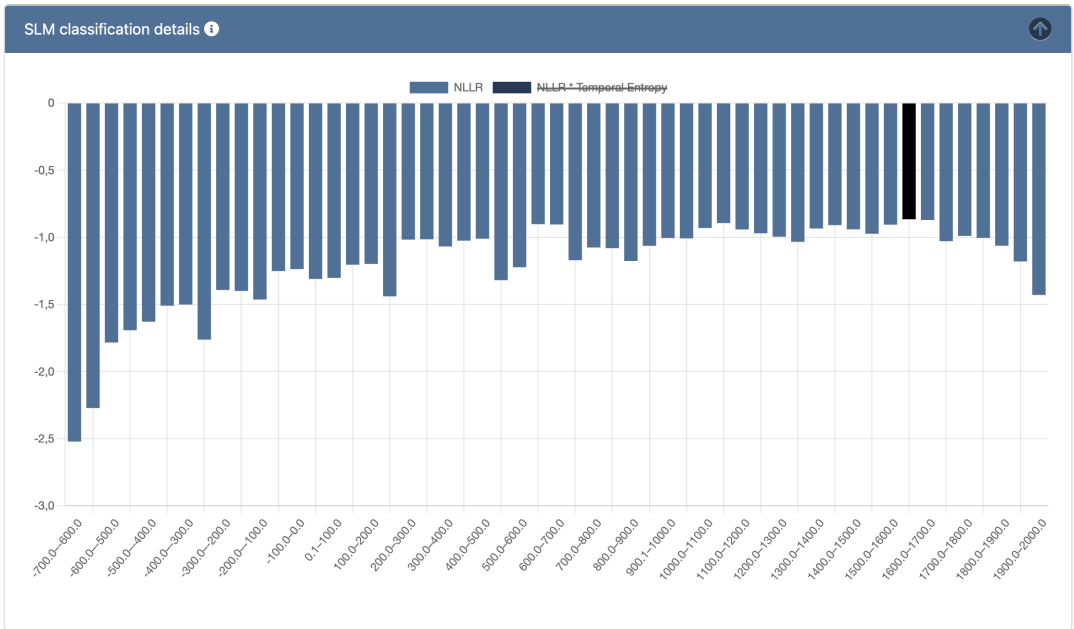


Abbildung 6.45 *VisualTime* – *NLLR*-Werte des *Sanguo zhi yanyi* für die einzelnen *chronons*

In diesem Fall lässt sich aus den genannten Interpretationen kein klares Bild ableiten. Alle drei genannten Datierungen können richtig sein, da weder klar ist, ob Luo der Autor war, wie stark die überlieferte Version von seiner Fassung abweicht und welche Version oder Ausgabe des *Sanguo zhi yanyi* hier genau vorliegt. Das Beispiel zeigt aber, dass sich sowohl mit dem *DHYDCD-SLM*, als auch mit der lexikographischen Datierungsmethode ähnliche Ergebnisse erzielen lassen, die in der Tendenz richtig sind.

Die Detaildarstellung der im Text erkannten *types* erlaubt es, weitere Analysen vorzunehmen. Die Betrachtung der temporalen Ausdrücke (Abb. 6.46) gibt Aufschluss über die Zeitangaben im Text – die späteste Angabe bezieht sich dabei auf das siebte Jahr der Ära Taikang (*Taikang qi nian* 太康七年, 286), während der Regierung von Kaiser Wu 武 der westlichen Jin 晉 (Sima Yan 司馬炎, reg. 266–290). Auffällig und erwartbar ist auch, dass der literarische Text trotz seinem etwas größereren Umfang (581.810 Zeichen) deutlich weniger unterschiedliche Zeitangaben – 78 – enthält als die historiographische Fassung mit 182.

Bei so umfangreichen Texten wenig aussichtsreich ist die Betrachtung der einzelnen Lexem-*types* (Abb. 6.47). Um z. B. Belege für eine Entstehung des Texts im 17. Jh. (oder später) zu suchen, müssen zumindest diejenigen *types* betrachtet werden, die dem 17. (52), 18. (35), 19. (6) und 20. Jh. (2) zugeordnet sind.

Abbildung 6.46 *VisualTime* – Temporale Ausdrücke im *Sanguo zhi yanyi* (Ausschnitt)

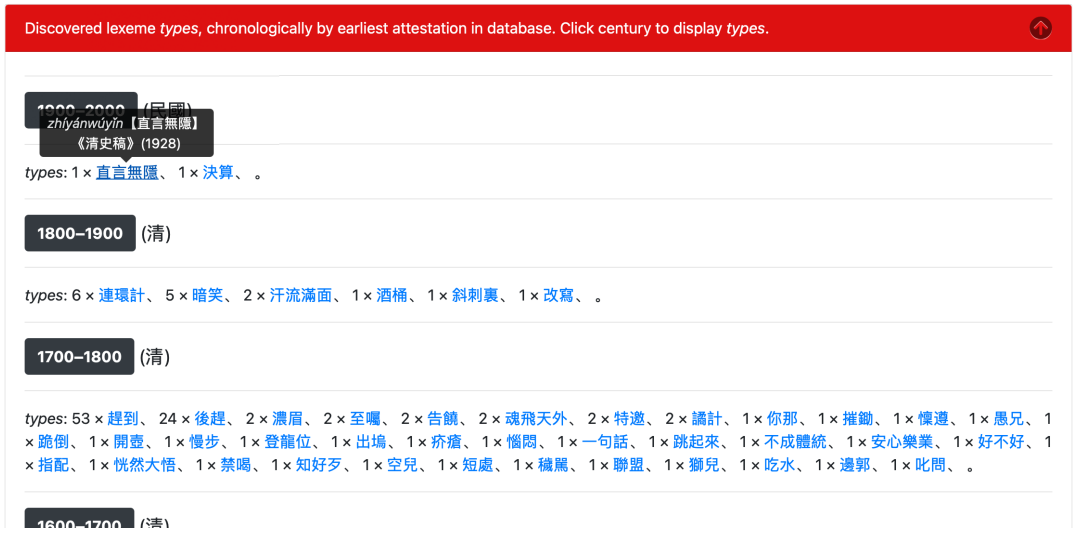


Abbildung 6.47 *VisualTime* – Lexeme im *Sanguo zhi yanyi* (Ausschnitt)

Der „neueste“ lexikalisierte Ausdruck mit 2–4 Zeichen im *Sanguo zhi yanyi* ist *zhi yan wu yin* 直言無隱 („offen seine Meinung sagen“). Wie im Screenshot zu sehen gibt das *HYDCD* als *attestation* das 1928 veröffentlichte *Qing shi gao* an.[298] Es kann davon ausgegangen werden, dass das *Sanguo zhi yanyi* ein deutliches *ante-dating* für diesen Ausdruck ermöglicht.

Das zweite ins 20. Jh. datierte Lexem, *juesuan* 決算 („Abschlussrechnung“), ist ein anschauliches Beispiel für *false positives*, da die beiden Zeichen im *Sanguo zhi yanyi* eher in ihren Einzelbedeutungen (etwa: „Pläne machen“ oder „Pläne ausführen“) zu verstehen sind.[299]

298 Siehe auch *HYDCD*, Bd 1., S. 858.

299 „後人有詩讚玄德曰運籌決算有神功，[…]“: „Die nachfolgenden Generationen machten ein Lied, in welchem sie Xuande 玄德 (i. e. Liu Bei 劉備) lobten: ‚Beim Entwerfen von Strategien und Machen von Plänen brachte er wunder-

Die Unvollständigkeit der diachronen Lexemdatenbank und die Problematik der fehlenden Tokenisierung oder sogar semantischen Analyse des zu datierenden Textes werden an den beiden Beispielen deutlich. Eine eingehende Untersuchung, wie sie z. B. in Abschnitt 6.2.4 anhand des *Zhongjing* 忠經 demonstriert wurde,[300] ist unter diesen Umständen mit zahlreichen *false positives* aufwändig – für kürzere Texte aber ein gangbarer Weg.

Datierung per *AYL*

Auf der Ergebnisseite wird zusätzlich die Datierung mittels *AYL* angezeigt.[301] Um ein breiteres Spektrum an schriftsprachlichen Texten einordnen zu können, wird zur Projektion ein Regressionsmodell mit den Texten der Loewe und *zhengshi*-Korpora trainiert.[302] Wie bereits in Abschnitt 6.3.1 wird zur Berechnung des *AYL* die Lexikalisierung der 90 % häufigsten 2–4-Zeichen Lexem-*types* verwendet. Als Grundlage dient die Datenbankabfrage für die Erzeugung des Neologismusprofils. Da hierbei die zusätzlichen Korpus-Belegstellen verwendet werden,[303] muss auch das Training für die Linearregression entsprechend berechnet werden. Zur Projektion dient auf dieser Basis die Funktion:

$$Y_{proj.} = 2,41 \times AYL_{0,9}^{2--4} + 904,03$$

Für das in den Trainingsdaten des Modells enthaltene *Sanguo zhi* ergibt sich daraus eine geschätzte Entstehung im Jahr 521 (+ 224 Jahre), für das umgangssprachlichere *Sanguo zhi yanyi* das Jahr 962 (- ?560 Jahre). Vor allem letztere Zuordnung ist unbrauchbar und deutet an, dass eine genre- oder stilübergreifende Datierung mit dieser Methodik nicht möglich ist.

Dass Ergebnisse aller untersuchten Methoden verglichen werden können, erleichtert die Einschätzung ihrer Verlässlichkeit. Übereinstimmende oder ähnliche Ergebnisse sprechen für eine korrekte Zuordnung, zwischen lexikographischer und statistischer Datierung stark abweichende Ergebnisse sollten Anlass zur Skepsis sein. Gerade in solchen Fällen ist eine qualitative Untersuchung erforderlich. Die diachrone Detailanzeige der *types* in *VisualTime* stellt dafür eine Hilfestellung dar, auf Basis derer zusätzliche Quellen konsultiert werden können.

bare Meisterleistungen hervor […]'". Luo Guanzhong 羅貫中 2007 [?1522].
300 Siehe S. 192.
301 Siehe dazu auch die Darstellung in Abb. 6.39 auf S. 234.
302 Siehe dazu Abschnitt 6.3.1, v. a. Abb. 6.35, S. 225.
303 Siehe dazu Kapitel 5.5.4, S. 134.