3 Linguistische Datierung

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8 Siehe ebd., S. 121.

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

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

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

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

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

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

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

12 Ebd., S. 207.

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

14 Harbsmeier 2019, S. 207.

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

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

17 Siehe ab S. 20.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

26 Valla 2007 [1440], S. 63.

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

28 Ebd., S. 189.

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

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

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

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

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

29 Henningsen 2010, S. 35–36.

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

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

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

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

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

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

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

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

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

3.1 Computerlinguistische Datierung von Texten

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

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

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

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

40 Kanhabua und Nørvåg 2008.

41 Garcia-Fernandez et al. 2011.

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

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

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

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

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

47 Bamman et al. 2017.

48 Graliński et al. 2017.

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

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

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

50 Yamada Takahito 山田崇仁 2004.

51 Yu Xuejin und Wei Huangfu 2019.

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

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

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

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

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

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

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

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

60 Stamou 2007, S. 181.

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

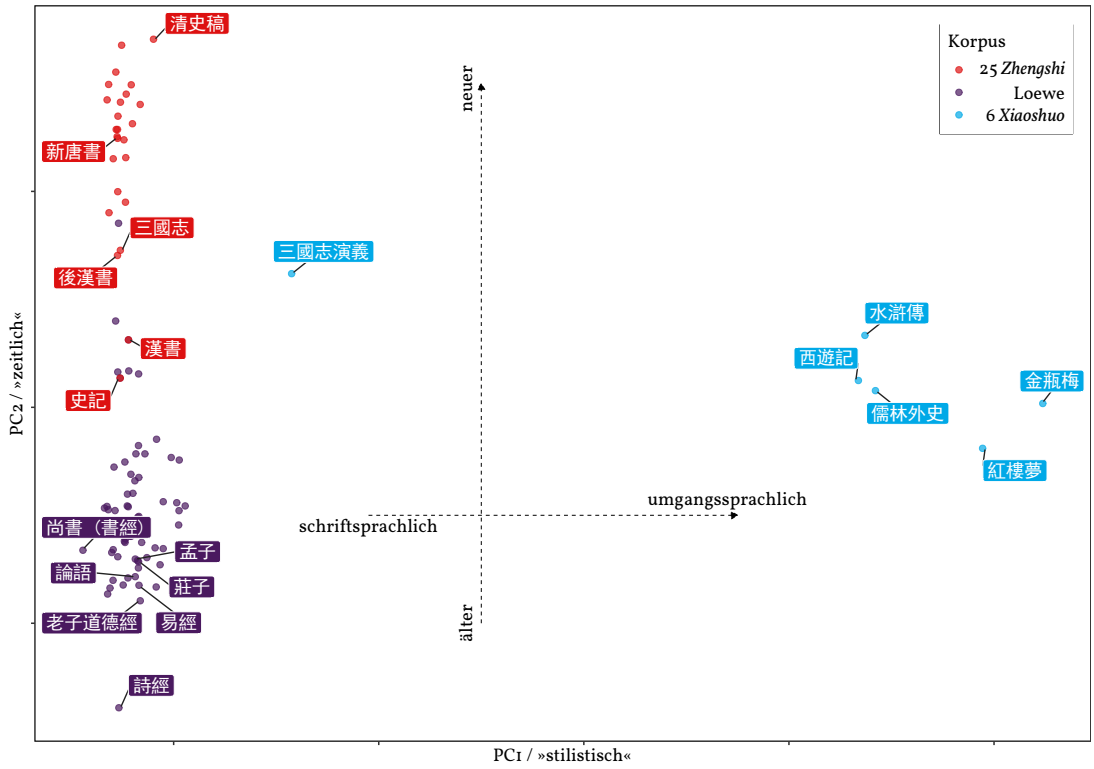

Abbildung 3.1 *PCA*, 1.000 häufigste 1–4 Zeichen Lexeme — *zhengshi* 正史, Loewe und *Xiaoshuo* 小說[65]

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

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

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

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

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

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

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

3.2 Datierung als Kategorisierungsproblem: de Jong, Rode und Hiemstra

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

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

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

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

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

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

68 de Jong, Rode und Hiemstra 2005.

69 Ebd., S. 3.

70 Siehe auch S. 50.

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

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

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

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

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

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

3.3 Systematisierung der bestehenden Ansätze

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

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

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

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

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

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

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

83 Siehe ebd.

84 Siehe dazu Kapitel 4.2, ab S. 62.

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

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

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

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

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

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

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

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

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

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

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

91 Siehe dazu Kapitel 6.3, S. 211.

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

93 Siehe Kanhabua und Nørvåg 2008, S. 359.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

104 Siehe dazu Kapitel 4.8, ab S. 103.

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

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

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

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

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

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

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

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

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

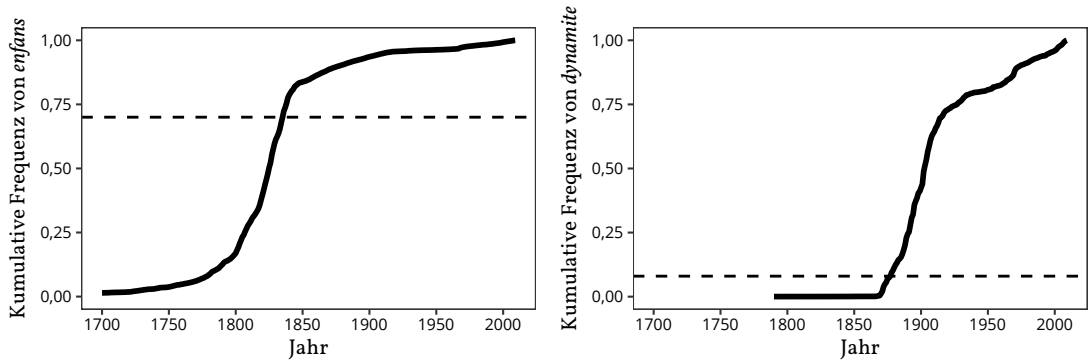

Abbildung 3.2 Kumulative Häufigkeit von „Archaismen“ und „Neologismen“[119]

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

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

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

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

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

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

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

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

121 Siehe Kapitel 2.1, ab S. 14.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Die *NLLR* ist definiert als:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

154 Siehe ebd., S. 368.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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