7 Ergebnisse und Ausblick

> *„Though the title suggests that they might actually date texts, in fact they argue that linguistic dating […] is a hopeless enterprise."*[1]
>
> Robert D. Holmstedt

In diesem letzten Kapitel werden die wichtigsten Resultate dieser Arbeit zusammengefasst und abschließend diskutiert. Es wird überdies ein Ausblick auf weiterführende Forschungsansätze gegeben, die Ergebnisse der vorliegenden Untersuchung aufgreifen.

Grundlage jeder linguistischen Datierung von Texten ist ein Verständnis diachroner sprachlicher Entwicklungen. In Kapitel 2 wurde daher zunächst der Sprachwandel im Hinblick auf für die Textdatierung relevante Aspekte beleuchtet. Davon ist der Wortschatzwandel für das Chinesische bislang aber am wenigsten systematisch erforscht. Es wurde auf das Piotrowski-Gesetz Bezug genommen und die bisher spärliche Forschung zu seiner Gültigkeit für das Chinesische betrachtet.

In einer Zusammenfassung des Forschungsstands zu Aspekten der historischen Entwicklung der chinesischen Sprache hat sich angedeutet, dass die Rigidität der chinesischen Schriftsprache bzw. der Konservativismus einiger Textgattungen und Sprachstile die Textdatierung erschweren und einige klassische Texte auch bei sorgfältiger Exegese nicht allein mit Methoden der historischen Sprachwissenschaft datiert werden können. Dies gilt umso mehr für Texte mit einer komplexen Überlieferungsgeschichte, wenn sie aus Fragmenten kompiliert wurden und/oder Inhalt und Form des Textes sich im Laufe der Jahrhunderte verändert hat, so dass teilweise eine Unschärfe in der Datierung von mehreren hundert Jahren akzeptiert werden muss.

Phonologische Veränderungen an der Sprache, die sich bei der Verwendung von Alphabetschriften oft in orthographischen Veränderungen widerspiegeln, werden mit chinesischen Zeichen in der Regel nicht verschriftlicht und stilistische, sowie syntaktische Veränderungen geschehen in der Schriftsprache sehr langsam. Dennoch konnten auch in dem stilistisch recht homogenen Korpus der offiziellen Dynastiegeschichten (*zhengshi* 正史) leichte Veränderungen in der Häufigkeit von Funktionswörtern beobachtet werden, die einen syntaktischen und stilistischen Wandel – unter Vorbehalt des engen Betrachtungsrahmens – nahelegen. Offensichtlichere Veränderungen finden aber im Wortschatz statt, was im selben Textkorpus gezeigt wurde. Als Beispiele für solche lexikalischen Innovationen und Veränderungen wurden die Verwendung einiger Amtstitel und mit dem Buddhismus verbundener Begriffe analysiert. Die Entstehung neuer Wörter ist dabei einfacher nachvollziehbar als ihr Aussterben. Neologismen können Indizien für die Textdatierung auch dann liefern, wenn Struktur und Stil eine hohe Kontinuität aufweisen.

1 Robert D. Holmstedt 2009: *Dating the Language of Ruth: A Study in Method*. Paper presented at the annual meeting of the Canadian Society of Biblical Studies (Ottawa, May 23, 2009). URL: http://individual.utoronto.ca/holmstedt/Holmstedt_DatingLangRuth_CSBSrevAug2009.pdf (besucht am 02. 08. 2021), S. 1. Holmstedt geht es in diesem Zitat um die Datierung althebräischer Texte in Young und Rezetko 2014.

In Kapitel 3 wurden sinologische Aspekte der linguistischen Datierung aufgezeigt und ein systematischer Überblick über die verfügbaren computerlinguistischen Methoden gegeben, die für die Datierung von indoeuropäischen Texten eingesetzt werden. Dabei kommen vor allem statistische Sprachmodelle zum Einsatz, die mit diachronen Textkorpora trainiert werden müssen und denen meist eine *Bag of Words* zugrunde liegt. Weitere Datierungsmethoden greifen auch auf Metadaten und explizite Zeitangaben in Texten zurück.

Für schriftsprachliches Chinesisch ergaben sich daraus zwei wesentliche Herausforderungen, die in Kapitel 4 diskutiert wurden: Diachrone Korpora, wie sie im Optimalfall für die Evaluation von Datierungsmethoden eingesetzt werden, stehen nicht in gewünschter Qualität und Umfang zur Verfügung. Eine diachrone Erprobung von Tokenizern hat zudem gezeigt, dass *state of the art* Segmentierung und *PoS*-Tagging für moderne Texte und neuerdings auch für klassisches Chinesisch in ausreichender Genauigkeit möglich ist. Für den langen, dazwischen liegenden Zeitraum bleiben die Ergebnisse aber unbefriedigend, da Wortgrenzen oft nicht korrekt erkannt werden. Mit einem einfachen *maximum matching*-Ansatz konnten teilweise bessere Erfolge erzielt werden, wenn individuell auf den Text angepasste, diachrone Wörterbücher verwendet wurden. Dies ist jedoch nur möglich, wenn die Entstehungszeit des zu segmentierenden Textes bekannt ist.

Es wurden zudem wichtige Aspekte von Namen und die Verwendung der *China Biographical Database* (*CBDB*) als Quelle für eine datenbankgestützte Erkennung von Namen diskutiert. Aus der Nutzung dieser biographischen Daten ergibt sich ein Potenzial, das für die Textdatierung über das einer herkömmlichen *Named Entity Recognition* hinausgeht, aber auch besondere Herausforderungen mit sich bringt. Es wurde festgestellt, dass dabei Ambiguitäten berücksichtigt werden müssen, die vor allem zweisilbige Namen betreffen. Als problematisch hat sich auch der hohe Anteil homonymer Personen herausgestellt, sowie vor allem *false positives* durch Namens-Zeichenkombinationen, die auch einfache Vorkommen lexikalisierter Bedeutungen dieser Zeichen sein können.

Inspiriert durch die für *MARKUS*[2] genutzte Implementierung der *Dharma Drum Buddhist College Time Authority Database*[3] konnte eine Möglichkeit der Erkennung und zeitlichen Einordnung von Regierungsdevisen erarbeitet werden, die für Zeitangaben in schriftsprachlichen Texten wesentlich sind. Anders als gregorianische Jahreszahlen beziehen sich diese in der Regel auf einen Zeitraum, der aus der Perspektive der Verfasser:in in der Vergangenheit, Gegenwart oder nahen Zukunft liegen muss.

Als Alternative zur Segmentierung wird die Zerlegung der Texte in *n*-Gramme vorgenommen. Hierfür wurde etabliert, dass für das schriftsprachliche Chinesisch die Betrachtung von maximal 1–4 bzw. 2–4 Grammen empfehlenswert ist. Um die Anzahl der *features* einer solchen Textrepräsentation zu begrenzen, wurde die Reduktion dieser *n*-Gramme auf lexikalisierte Wörter, Zeitausdrücke und Namen vorgeschlagen. Zur Erprobung von Datierungsmethoden konnten somit auch *n*-Gramm-Datensätze verwendet werden, wie sie seit 2019 von CrossAsia bereitgestellt werden.[4]

Mit dem Ziel, eine Datengrundlage für Textdatierungsmethoden zu schaffen, bei denen der Aspekt der Wortbildung im Vordergrund steht, wurde in Kapitel 5 aus einer digitalen Ausgabe

2 Ho und DeWeerdt. 2014–.
3 *DDBC*.
4 Siehe CrossAsia, Staatsbibliothek zu Berlin 2019–: *CrossAsia N-Gramm Service*. Website. URL: https://crossasia.org/service/crossasia-lab/crossasia-n-gram-service/ (besucht am 19. 04. 2019).

des *Hanyu da cidian* 漢語大詞典[5] (*DHYDCD*) erstmals eine diachrone Lexemdatenbank für das Chinesische erzeugt. Die Wahl von *Python* für die Implementierung der notwendigen Software hat sich – wie für alle weiteren Programmierungen im Rahmen dieser Arbeit – als adäquat erwiesen. Die Lexikalisierungsdaten aus dem *DHYDCD* konnten mithilfe biblio- und biographischer Daten aus der *CBDB* sowie durch Betrachtung datierter Primärtexte weiter verdichtet werden, so dass eine Datenbank mit 272.892 Lexemen zur Verfügung steht, die mit einer Genauigkeit von ca. 80 Jahren datiert sind. Hinzu kommen über 600.000 zugehörige Textbelegstellen, aus denen nach dem Vorbild von Hoffmann[6] ein diachrones Behelfskorpus generiert wurde. Daraus konnten temporale Sprachmodelle für einen Betrachtungszeitraum von über 2.000 Jahren erzeugt werden.

Die Datenbank hat zudem die Gewinnung neuer Erkenntnisse über Machart sowie Stärken und Schwächen des *HYDCD* ermöglicht. Eine Analyse der verwendeten Belegstellen konnte lexikographische Präferenzen für bestimmte Texte und Textgattungen aufdecken, in der eine tiefe Verwurzelung des Wörterbuchs in der schriftsprachlichen Texttradition zum Ausdruck kommt.

Mit einer chronologischen Analyse der Lexikalisierungsdaten wurde überdies neues Licht auf die Geschichte des chinesischen Wortschatzes geworfen. Was in der Forschungsliteratur als unscharf getrennte Phasen der Sprachentwicklung dargestellt wird, ließ sich als kontinuierliche, logistische Entwicklung darstellen. Diese Beobachtung legt nahe, dass das Piotrowski-Gesetz auch für die Modellierung des Wortschatzwachstums des Chinesischen geeignet ist. Ob Schwankungen in der Aufnahme neuen Vokabulars bestimmte (Entlehnungs-)Wellen oder Krisen widerspiegeln, ließ sich dabei nicht eindeutig klären. Dass Krisen einen immensen sprachlichen Innovationsschub bewirken können, zeigt sich aktuell in einer Vielzahl von Wortbildungen wie „Flockdown“ oder „Lollitest“.[7] In der retrospektiven Lexikographie werden solche aber nur erfasst, wenn sie sich längerfristig im Sprachgebrauch durchsetzen können.

Es konnte verdeutlicht werden, dass bereits ab dem 4. Jh. v. u. Z. ein hoher Anteil nicht nur zwei, sondern auch drei- und viersilbiger Wortbildungen belegt sind. Während dabei anfangs noch eine – erwartbare – Präferenz für tetra- gegenüber trisyllabischen Lexemen besteht, ist der Anteil an drei- und viersilbigen Wortbildungen ab dem 4./5. Jh. weitestgehend identisch. Mehr als die Hälfte der Wörter, sogar in modernen Texten, bleiben einsilbig,[8] aber über 80 % der Einträge im *DHYDCD* sind disyllabische Wortbildungen. Entgegen dem Eindruck, der durch von Jahrhundert zu Jahrhundert wachsende Zeichenwörterbücher geweckt werden kann, sind fast alle auch heute gebräuchlichen Schriftzeichen – von vereinfachten Formen abgesehen – bereits in der Han-Zeit vorhanden, während sich neue Zeichen – im Gegensatz zu mehrsilbigen Wortbildungen – nur schwer durchsetzen können.

In Kapitel 6.1 wurden für westliche Sprachen erfolgreiche Methoden der Textdatierung auf Basis temporaler Sprachmodelle implementiert und auf ihre Eignung für schriftsprachliche chinesische Texte getestet. Solche Ansätze, bei denen die Datierung als Kategorisierungsproblem aufgefasst wird, haben sich als auf beliebige Entwicklungsstufen des Chinesischen übertragbar erwiesen. Mit dem *N-gram dataset of Chinese local gazetteers* (*Zhongguo Difangzhi* 中國地方誌, *DFZ*) konnten z. B. mehr als 60 % der Testdaten aus demselben Korpus einem korrekten

5 *DHYDCD*.
6 Siehe Hoffmann 2004.
7 Eine umfassende Sammlung von Neologismen, die mit der Coronapandemie in Verbindung stehen, findet sich in: Leibniz-Institut für Deutsche Sprache (IDS) 2020.
8 Siehe auch Breiter 1994, 224ff.

chronon, hier einem Zeitraum von 50 Jahren, bei einem Betrachtungsfenster von 450 Jahren, korrekt zugeordnet werden. Die durchschnittliche Genauigkeit der Datierung lag bei etwa 42 Jahren.[9] Als Ähnlichkeitsmaße haben sich dabei die *Normalized-Log-Likelihood-Ratio* bzw. die Kullback-Leibler-Divergenz am stärksten bewährt. Durch Gewichtung der betrachteten *types* mittels *Temporaler Entropie* konnte nur teilweise eine leichte Verbesserung der *Accuracy* erzielt werden. Bei einer Analyse von Glättungsmethoden hat sich die naive Annahme einer Häufigkeit von *unseen events*, die niedriger als die geringste im Korpus beobachtete Häufigkeit ist, als am effektivsten herausgestellt.

Um die zeitliche Einordnung von Texten auch über Genregrenzen hinweg und für einen größeren Zeitraum zu ermöglichen, wurden zudem temporale Sprachmodelle aus den Belegstellen im *DHYDCD* erzeugt. Die Klassifizierung von Texten in *chronons* von 100 Jahren bei einem Zeithorizont von über 2.000 Jahren kann damit grundsätzlich als vielversprechend angesehen werden. Wie Tests mit dem *N-gram dataset of Xu xiu si ku quan shu* 續修四庫全書 (*XXSKQS*) gezeigt haben, ist die Datierung schriftsprachlicher Texte mit solchen Sprachmodellen gleichzeitig aber zutiefst problematisch. Dies liegt primär am Modellcharakter einiger antiker Texte, deren Stil und damit Grammatik über Jahrtausende als Vorbild gedient haben. Dass für westliche Sprachen erfolgreiche Datierungsmethoden an solchen Texten scheitern, verdeutlicht die beeindruckende Kontinuität der Tradition bestimmter schriftsprachlicher Textgattungen. Das rigide Schriftsystem, das von kurzlebigen orthographischen Anpassungen aufgrund von phonologischem Wandel weitestgehend unberührt bleibt, verstärkt diesen Effekt.

In Kapitel 6.2 wurde mit temporalen Textprofilen eine neue lexikographische Methode eingeführt, die eine temporale Einordnung von Texten auf Basis der frühesten Belege der enthaltenen Lexeme, sowie den vorkommenden Namen und Zeitangaben ermöglicht. Die gewählte Darstellung, die die Anzahl der pro Jahrhundert zugeordneten *types* als Balkendiagramm visualisiert, kann als Hilfestellung für qualitative Analysen genutzt werden. Anhand der Profile lässt sich ablesen, wieviele *types* durch Textbelege bzw. biographische Daten und Zeitangaben mit einem Jahrhundert verbunden sind, was Rückschlüsse sowohl über Inhalt als auch Genese des Textes zulässt – unabhängig von Genre oder Alter des Eingabetextes.

Mithilfe von Trainingsdaten können diese temporalen Profile auch für die automatische Datierung von Texten eingesetzt werden. Bei Experimenten mit dem *DFZ*-Datensatz wurden dabei Ergebnisse erzielt, die mit denen bei Verwendung temporaler Sprachmodelle vergleichbar sind. Etwa 50 % der untersuchten Texte konnten auf etwa 100 Jahre genau eingeordnet werden, wobei der Betrachtungszeitraum über 2.000 Jahre betrug. Bei zusätzlicher Betrachtung von Personennamen konnte der Anteil der korrekt datierten Texte auf über 60 % erhöht werden. Bei einer einfachen Betrachtung von *temporal expressions* konnten in diesem Korpus sogar 88 % der Texte korrekt zugeordnet werden. Die Besonderheit, dass Zeitangaben in Form von Regierungsdevisen erfolgen, hat sich dabei als nützlich herausgestellt, da – im Gegensatz zu westlichsprachigen Texten – damit keine Referenz auf die ferne Zukunft erfolgen kann.

Experimente mit dem *XXSKQS*-Datensatz konnten zudem zeigen, dass eine automatisierte lexikographische Datierung auch unabhängig von Trainingsdaten erfolgreich sein kann.

Neben der ungefähren zeitlichen Einordnung von Dokumenten kann das Konzept der temporalen Textprofile auch zur Unterstützung bei der Analyse von vermuteten Fälschungen, bzw. Texten von fragwürdiger Provenienz oder ungeklärter Autorschaft herangezogen werden. Da die zur Verfügung stehenden Lexikalisierungsdaten große Ungenauigkeiten aufweisen können

9 Siehe Kapitel 6.1.1, ab S. 158.

und die Verwendung von *n*-Grammen zur Segmentierung der Texte einen hohen Anteil an *false positives* mit sich bringt, erfordert dieser Anwendungsbereich ein hohes Maß zusätzlicher, sorgfältiger Recherche.[10]

In Kapitel 6.3 wurde zuletzt ein experimenteller Ansatz betrachtet, der die bereits in 6.2 verwendeten Lexikalisierungsdaten aus dem *DHYDCD* auf einen einzigen Messwert abstrahiert: das Jahr der durchschnittlichen Lexikalisierung der enthaltenen Lexeme. Mit den *DHYDCD*-nahen *zhengshi* 正史 konnte dabei ein verblüffender Zusammenhang zwischen diesem *Average Year of Lexicalization* (*AYL*) und der ursprünglichen Entstehungszeit beobachtet werden, der im betrachteten Zeitraum linear zu sein scheint. Es wurde unter anderem festgestellt, dass der statistische Zusammenhang zwischen Textentstehung und *AYL* besser ist, wenn Unigramme unberücksichtigt bleiben und stattdessen nur ein geringer Anteil der häufigsten 2–4-Gramme betrachtet wird. Weitere Experimente legen aber nahe, dass diese Art der Modellierung nur innerhalb der engen Grenzen eines homogenen Korpus gut funktioniert.

Zuletzt wurden in Kapitel 6.4 die drei für das schriftsprachliche Chinesische untersuchten Datierungsmethoden gegenübergestellt und ein *user interface* vorgestellt, das für einen hochgeladenen *Plain Text* einen Vergleich der Ergebnisse im Browser ermöglicht.

Unabhängig von der gewählten Methodik wurden die Limitationen, schriftsprachliche chinesische Texte (computer)linguistisch zu datieren, immer wieder sichtbar. Bei der automatisierten Textdatierung kann der korrekte Zeitstempel für zu datierende Texte um viele hundert Jahre, in Einzelfällen sogar mehr als ein Jahrtausend, verfehlt werden. Davon betroffene Texte weisen tendenziell ein hohes Maß an Intertextualität auf, bilden ein „Mosaik von Zitaten“[11] älterer Schriften oder kommentieren diese.

Dass solche Texte fälschlich etwa der Zeit zugeordnet werden, deren Syntax und Stil darin imitiert oder aufgegriffen wird, kann in gewisser Hinsicht auch als Bestätigung dafür gewertet werden, dass die chronologische Einordnung grundsätzlich funktioniert. Manche Autor:innen kommen gänzlich ohne zeitgenössisches Vokabular aus, was entsprechende Texte – auch bei Recherche der einzelnen *types* aus temporalen Textprofilen – gegen eine rein linguistische Analyse vollkommen resistent macht. Wegen des inhaltlichen und sprachlichen Vergangenheitsbezugs könnten solche Texte in gewisser Hinsicht auch als „transtemporal“ bezeichnet werden.[12] Für die Klärung der Authentizität oder Entstehungszeit solcher Dokumente verbleibt nur der Verweis auf die qualitative, inhaltliche und kontextuelle Analyse, sowie die zusätzlichen Möglichkeiten, die gegebenenfalls für physisch vorliegende Exemplare zur Verfügung stehen: Bestimmung des Alters des verwendeten Papiers bzw. Trägers, der verwendeten Farbe, Analyse von Kalligraphie-, Zeichen- und Druckstilen,[13] sowie Tabuzeichen.[14] Derartige Untersuchungen erlauben Rückschlüsse aber stets nur auf die vorliegende Kopie, auf eine Ausgabe, die Manifestation eines Textes, nicht über seine ursprüngliche Genese.

10 Siehe zur Genauigkeit der Lexemdatierungsdaten und zur Polysemie der Lexeme auch Kapitel 5.7.1, ab S. 139. Viele der erkannten Zeichenkombinationen sind zudem im geg. Kontext keine „Wörter“, vgl. auch Kapitel 2.3 und 5.5.4.

11 Julia Kristeva 1972 [1965]: „Bachtin, das Wort, der Dialog und der Roman“. In: *Zur linguistischen Basis der Literaturwissenschaft II*. Hrsg. von Jens Ihwe. Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven Bd. 3. Frankfurt am Main: Athenäum, S. 345–375, S. 348. Kristevas Formulierung scheint hier sehr zutreffend, obwohl sie ihrem literaturwissenschaftlichen Kontext entrissen ist.

12 Der Begriff der Transtemporalität im Kontext von Anspielungen auf die Vergangenheit ist entlehnt aus Christian Soffel 2020: „Transcultural Aspects in Chang Yi-Jen's 張以仁 Poetry“. In: *Interface – Journal of European Languages and Literatures* 12.2, S. 113–137. doi: 10.6667/interface.12.2020.111, S. 133.

13 Vgl. aber Galambos 2006.

14 Siehe dazu Kapitel 4.3, sowie v. a. Adamek 2012.

Wichtige Texte der schriftsprachlichen bzw. klassischen Texttradition, wie das *Shijing* 詩經 (*Buch der Lieder*) oder das *Shangshu* 尚書 (*Buch der Urkunden*), deren Datierung besonders intensiv diskutiert wird, wurden vollkommen ungeachtet unserer heutigen, eurozentrischen Auffassung von Autorschaft über einen Zeitraum von mehreren hundert Jahren bis zum Erreichen ihrer heutigen Form immer wieder verändert oder ergänzt. Das Streben nach einer präzisen Datierung solcher Textgewebe wirkt in diesem Kontext ebenso unpassend, wie es die Zuschreibung zu einer einzigen Autor:in wäre.[15] Zumindest müsste die ohnehin komplexe Frage nach der Datierung durch zusätzliche Aspekte verkompliziert werden.[16] Der Versuch einer genauen, linguistischen Datierung solcher Texte gerät zu einer Art Henne-Ei-Problem, wenn Trainingsdaten bzw. Primärquellen selbst nur sehr vage datiert sind, denn „eine robuste lexikalische Chronologie setzt zunächst die Existenz einer großen Reihe von sicher datierten Texten voraus."[17] Gesicherte Belege lassen sich teilweise aus archäologischen Funden gewinnen, deren Alter naturwissenschaftlich bestimmt, oder aus dem Kontext des Fundorts ermittelt werden kann.[18] Tatsächlich kann durch solche Funde aber nur – wenn überhaupt – das *Mindest*alter eines Texts korrigiert werden, da sie keinen Aufschluss über die ursprüngliche Textgenese erlauben.

7.1 Ausblick

Naheliegend für eine Weiterentwicklung der in Kapitel 6.2 vorgestellten Methodik zur zeitlichen Einordnung von Texten ist eine kontinuierliche Verbesserung der Datengrundlage. Dies kann durch eine „mitlernende" Datenbank geschehen, in die laufend zusätzliche Lexeme und Namen sowie Belegstellen – auch zu bereits vorhandenen Einträgen – aus sicher datierten Texten aufgenommen werden. Die Verwendung der Profile für die Analyse moderner chinesischer Texte könnte dann ebenfalls erwogen werden. Auch ein *crowd sourcing* mit Ergänzung von Worteinträgen und Belegen durch Anwender:innen ist denkbar. Die Zuverlässigkeit der Textprofile könnte so immer weiter verbessert und letzlich auch eine feinere Granularität erreicht werden als der momentan noch unscharfe Zeitraum von 100 Jahren. Dazu müsste allerdings die Toleranzschwelle, wie viele „zu neue" Lexeme ein Text enthalten darf, ebenfalls kontinuierlich trainiert werden. Unlösbar bleibt die Tatsache, dass viele schriftsprachliche Texte bewusst in einem antiken Stil verfasst sind und wenig bis gar keine zeitgenössischen Lexeme enthalten.

Neben der möglichst genauen Sammlung von *Loci classici* sollte auch die systematische Erfassung des Lebenszyklus von Wörtern angestrebt werden. Die im Rahmen von Kapitel 5.5 gesammelten Belege können dafür einen Anfang bilden und perspektivisch auch helfen, Fragen über das Verschwinden von Wörtern zu beantworten. Überdies ist auch das Potenzial des *DHYDCD* als Datenquelle noch nicht ausgeschöpft, da neben den hier verwendeten Informationen zum lexikalischen Sprachwandel auch Belege zu unterschiedlichen Bedeutungen gegeben

15 Vgl. z. B. Edward Shaughnessy 1993b: „*Shang shu* 尚書 (*Shu ching* 書經)". In: *Early Chinese Texts: A Bibliographical Guide*. Hrsg. von Michael Loewe. Berkeley: SSEC und IEAS, S. 376–389, v. a. S. 376–380; vgl. z. B. Boltz 2007, S. 70–71; für eine Diskussion des Autoschaftsbegriffs für Asien siehe auch Christian Schwermann und Raji C. Steineck 2014: „Introduction". In: *That Wonderful Composite Called Author*. Hrsg. von Christian Schwermann und Raji C. Steineck. East Asian Comparative Literature and Culture 4. Leiden: Brill, S. 1–29, v. a. S. 20–26.

16 Siehe dazu auch Kapitel 3, S. 36, sowie Harbsmeier 2019, S. 189.

17 Toner und Han Xiwu 2019, S. 36, übersetzt durch den Verfasser.

18 Ein Beispiel sind umgangssprachliche Funde aus Dunhuang 敦煌, siehe z. B. Jing-Schmidt und Hsieh 2019, S. 516; Als weiteres Beispiel seien die 1993 in Guodian 郭店 gefundenen Bambusstreifen genannt, die auf die Mitte des 4.–3. Jh. v. u. Z. datiert werden können. Siehe z. B. Shirley Chan 2019: „Introduction: The Excavated Guodian 郭店 Bamboo Manuscripts". In: *Dao Companion to the Excavated Guodian Bamboo Manuscripts*. Hrsg. von Shirley Chan. Dao Companions to Chinese Philosophy 10. Cham: Springer Nature, S. 1–20. DOI: 10.1007/978-3-030-04633-0_1, S. 4–7.

werden. Aus diesen ließen sich auch umfangreiche Daten zum semantischen Wandel gewinnen. Um sie für die Datierung nutzbar zu machen, wäre selbstverständlich auch eine semantische Analyse der zu datierenden Texte erforderlich.[19]

Dass „es für fast alle europäischen Sprachen historische Wörterbücher gibt, die – oft genaue – Angaben über die erste schriftliche Erwähnung eines jeden Wortes liefern",[20] macht es interessant, Neologismusprofile auch für europäische Sprachen einzusetzen. Mit der digitalen Fassung des *Oxford English Dictionary*[21] oder der Neubearbeitung des *Deutschen Wörterbuchs* von Jacob und Wilhelm Grimm[22] stehen Quellen zur Verfügung, in denen – im Gegensatz zu den oft vagen Angaben im *HYDCD* – die Belege mit Jahreszahlen angegeben werden. Entsprechende Experimente ließen also sogar eine genauere Darstellung zu. Dass die chinesische Sprache stark isolierend und ihre Schrift eine fast zeitlose, von orthographischen Veränderungen verschonte Schreibung ermöglicht, war hier von Vorteil. Ein Abgleich historischer, westlicher Texte mit diachronen Wörterbüchern setzt hingegen eine intensivere Beschäftigung mit der Normalisierung von Schreibweisen und mit Lemmatisierung voraus.

Entwicklungen sowohl in Bereichen des *machine learning*, sowie wachsende digitale Sammlungen schriftsprachlicher, darunter auch mittelchinesischer und spätkaiserzeitlicher Texte lassen in naher Zukunft auch eine befriedigende Tokenisierung und vor allem ein *PoS-Tagging* für solche Texte möglich erscheinen. Die Erfassung von Wortarten hätte dabei ein großes Potenzial in der Erforschung von Sprachwandel, da anstatt bloßer Vorkommen von Zeichen und Zeichenkombinationen Wörter und – zumindest in Teilen – auch ihre unterschiedlichen Bedeutungen betrachtet werden könnten. Eine Analyse der Häufigkeit von *zhi* 之 in Kapitel 2.3 konnte dies nur erahnen lassen.[23] Dass durch *PoS-Tagging*, sowie auch durch die Unterscheidung von Bedeutungen und Berücksichtigung gängiger Wortverbindungen (*collocations*) auch die Aussagekraft temporaler Sprachmodelle verbessert werden kann, wurde von Kanhabua und Nørvåg bereits gezeigt.[24] Auf Datensätze mit *n*-Gramm Häufigkeiten können diese Techniken allerdings nicht angewandt werden.

Ähnlich den in Kapitel 6.1 verwendeten temporalen Sprachmodellen könnten Methoden aus dem Bereich des *machine learning* auch für die Einordnung (schriftsprachlicher) chinesischer Texte getestet werden. Als Einstiegsmöglichkeit bieten sich *support vector machines* an, die bereits im Bereich der Textdatierung eingesetzt wurden.[25]

Spannend wäre dabei auch zu sehen, ob Textdatierungen für modernes (*xiandai* 現代) Chinesisch innerhalb des 20. Jahrhunderts spürbar genauer möglich sind, als dies bei den in relativ große *chronons* eingeordneten schriftsprachlichen Texten der Fall war. Der stark angestiegene sprachliche Wandel, wie ihn die Beobachtungen aus Kapitel 6.1.4 für diesen Zeitraum andeu-

19 Bei Verfügbarkeit passender Trainingsdaten, stehen *machine learning* Technologien zur Verfügung, mit denen sich entsprechende Modelle trainieren ließen. Dazu zählen neben sogenannten *word embeddings* auch die bereits in Kapitel 4.5 (ab S. 77) erwähnten *Transformers*.

20 Alinei 2004, S. 213, übersetzt durch den Verfasser.

21 John A. Simpson, Hrsg. 2014 [2002]: *Oxford English Dictionary, 2nd Edition [Second edition on CD-ROM Version 3.0]*. Oxford University Press. URL: http://njw.me.uk/oed (besucht am 10. 04. 2014).

22 *Deutsches Wörterbuch von Jacob* Grimm *und Wilhelm* Grimm*, Neubearbeitung (A–F), Version 01/21* 2021. Digitale Fassung im Wörterbuchnetz des Trier Center for Digital Humanities. URL: https://www.woerterbuchnetz.de/DWB2 (besucht am 30. 05. 2021).

23 Siehe Kapitel 2.3, ab S. 24.

24 Siehe Kanhabua und Nørvåg 2008, S. 361, S. 367–369, siehe auch Kapitel 3.3, ab S. 45.

25 Siehe Garcia-Fernandez et al. 2011, S. 8–10, siehe auch Kapitel 3.3, ab S. 45. Die Verfügbarkeit von *Python*-Bibliotheken wie `scikit-learn` ermöglicht überdies Experimente mit zahlreichen weiteren Methoden, die dem *machine learning* zugerechnet werden können. Vgl. auch Fabian Pedregosa et al. 2011: „Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, S. 2825–2830; Vanderplas 2018.

ten,[26] sollte dies begünstigen. Für solche Analysen bieten sich aufgrund der genauen Datierung diachrone Sammlungen von Zeitungsartikeln, wie etwa aus der *Renmin ribao* 人民日報 (*RMRB*), an. Eine temporale Veränderung der darin behandelten *Themen* dürfte dabei allerdings einfacher zu beobachten sein als rein sprachliche Veränderungen.[27]

Unter dem Eindruck von Diskussionen um den Missbrauch von *Open Source* Software wie dem Apache Webserver im Einsatz bei Ölfirmen oder *NationBuilder* als Plattform für die Brexit-Kampagne,[28] scheint es zuletzt auch angezeigt, mögliche *misuse cases*[29] zu betrachten.[30] Besteht ein solches Dilemma auch für Software zur Datierung von Texten?

Sie ließe sich sicherlich einsetzen, um die Produktion glaubhafterer Fälschungen historischer Texte zu erleichtern, indem entlarvende Anachronismen vorab erkannt werden. Abgesehen vom zweifelhaften Nutzen eines solchen Unterfangens bleibt es der kritischen Leser:in unbenommen, eine Fälschung an anderen, inhaltlichen Kriterien zu erkennen, selbst wenn die Fälscher:in z. B. einen perfekten Song-zeitlichen Wortschatz verwendet. Die Datierungssoftware selbst wäre durch diesen Schwindel allerdings leicht hinters Licht zu führen. Der hier beschriebene *misuse case* kann zugleich ein valider Anwendungsfall sein, etwa wenn authentische Dialoge für fiktionale historische Werke wie Fernsehserien verfasst werden sollen. Näherliegend, gerade für das schriftsprachliche Chinesische, ist aber die Verwendung für die Vergabe von Zeitstempeln innerhalb von *NLP*-Workflows. Die zeitliche Einordnung des Textes kann so zur Verbesserung von *NER* oder der Tokenisierung von Texten beitragen und somit für eine breite Vielfalt computerlinguistischer Anwendungen genauso nützlich sein, wie für eine erleichterte Textlektüre.

26 Siehe ab S. 177.

27 Siehe aber Tang Xuri, Qu Weiguang und Chen Xiaohe 2015, Die Autoren untersuchen anhand von 59 Jahrgängen der *RMRB* den semantischen Wandel einzelner Lexeme wie *touming* 透明. vgl. auch de Jong, Rode und Hiemstra 2005, S. 4.

28 Siehe dazu z. B. Chris Jensen 2018: *Why we need an Open Source License that considers the misuse of our code*. URL: https://hackernoon.com/why-we-need-an-open-source-licence-that-considers-the-misuse-of-our-code-8d19b65d425 (besucht am 27. 09. 2018).

29 „A misuse case is a special kind of use case, describing behavior that the system/entity owner does not want to occur." Guttorm Sindre und Andreas L. Opdahl 2000: „Eliciting Security Requirements by Misuse Cases". In: *Proceedings of TOOLS Pacific*, S. 120–131, S. 122; Zu industrierelevanten *misuse cases* siehe z. B. auch Ian Alexander 2003: „Misuse Cases: Use Cases with Hostile Intent". In: *IEEE Software*, S. 58–66. DOI: 10.1109/MS.2003.1159030.

30 Eine ausführtliche Diskussion zu *Dual Use* und weiteren ethischen Fragen in den *DH* findet sich in Malte Rehbein und Christian Thies 2017: „Ethik". In: *Digital Humanities – Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, S. 353–357, *passim*.