VLASTIMIL BROM

Mittelhochdeutsche Wörter eingebettet

Überlegungen und ausgewählte Stichproben des Einsatzes von word embeddings zur Auswertung historischer Textquellen

Abstract

Middle High German words embedded. Some remarks and selected samples on the usage of word embeddings for analysing historical texts

The article deals with word embeddings as one of the current approaches for analysing the textual data (mainly with respect to lexical semantics, pragmatics, discourse linguistics etc.). Special attention is being paid to the applicability for historical texts, with the inherently limited amount of available text sources. The Middle High German Conceptual Database (MhdBDB) is used as data source. In addition, the possibilities of making the queries based on word embeddings available within this infrastructure are briefly discussed.

Keywords: historical texts, Middle High German, semantics, word embeddings, Middle High German Conceptual Database
DOI: doi.org/10.15452/Beitrage.2022.03

1. Wort-Einbettungen (word embeddings) – Grundmerkmale des Verfahrens

1.1 Allgemeines zum Konzept der Wort-Einbettungen

Für die Auswertung komplexer sprachlicher Daten werden immer häufiger die Methoden des maschinellen Lernens eingesetzt, die inzwischen in verschiedensten Domänen erprobt sind. Eine wichtige Rolle spielt hier das Konzept der sog. *word embeddings* (Wort-Einbettungen), das insbesondere für Arbeiten mit lexikalisch-semantischen Fragestellungen relevant ist. Im vorliegenden Beitrag soll

auf einige Aspekte dieser Analyseverfahren eingegangen werden, insbesondere im Hinblick auf die Einsatzmöglichkeiten für historische Texte.

Im Kern handelt es sich (in der aktuellen Umsetzung des Konzepts) um eine Verallgemeinerung der Kookkurrenzverhältnisse von definierten sprachlichen Einheiten im gegebenen Textkorpus, wo die gegenseitigen Relationen mit Hilfe von sog. künstlichen neuronalen Netzen erfasst und in mehrdimensionalen Vektorräumen numerisch repräsentiert werden (zum Konzept und Verfahren vgl. Görz/Schmid/Braun 2021:521–524). Die Berechnung der Wortvektoren ist so konzipiert, dass die Worteinträge mit ähnlichen Kookkurrenzprofilen, d. h. im gewissen Sinne ähnlich gebrauchte, ggf. potenziell inhaltlich benachbarte Wörter, durch Vektoren repräsentiert werden sollen, die im Vektorraum wenig voneinander entfernt sind. Mit geeigneten mathematischen Verfahren lassen sich dann die so modellierten Eigenschaften der verarbeiteten Einheiten abfragen und auswerten.

1.2 Ausgewählte Anwendungen im germanistischen Kontext

Die Vorteile dieses weitgehend automatisierbaren Verfahrens konnten z. B. in der aktualisierten 9. Auflage des ‚Dornseiff' genutzt werden, wo auf diese Weise die Zuordnung von neu hinzukommenden Worteinheiten zu entsprechenden Sachgruppen vorab automatisch ermittelt wurde.[1] Eine anschauliche Darlegung der Möglichkeiten von *word embeddings* liefert Bubenhofer in seiner Studie zur semantischen Äquivalenz in Geburtserzählungen (Bubenhofer 2020), wo in der dezidiert datengeleiteten Verarbeitungsstufe auch eine Methode des *clusterings* von *word-embedding*-Profilen entworfen wird. Dabei wird in abschließenden Ausführungen auch das „Potenzial für eine gebrauchssemantisch, diskurslinguistisch orientierte Analyse" aufgezeigt (Bubenhofer 2020:587), wo die so verallgemeinerten Regularitäten des Sprachgebrauchs den Anliegen und Ergebnissen der onomasiologischen Herangehensweise nahekommen können. Ein Bild vom Potenzial der *word embeddings* ergibt sich auch aus dem Vergleich dieser Verfahren mit einer früheren Analyse derselben Korpusgrundlage, wo sich Bubenhofer anderer Ansätze zur korpusgeleiteten Ermittlung der narrativen Muster bzw. Sprachgebrauchsmuster bedient, v. a. der sog. komplexen n-Gramme (Bubenhofer 2018:363 f.). Die untersuchten Geburtserzählungen wurden als

[1] Vgl. Dornseiff (2020:17*).

Vertreter der seriellen Texte ausgewählt, sie sind daher durch ihre thematische, mediale, textsorten- bzw. gattungsspezifische Grundanlage vorgeprägt; aufgrund ihrer Gemeinsamkeiten in vielen Bereichen lassen sich die Gebrauchsmuster bzw. die narrativen Schemata gerade aus der Vielfalt der konkreten Formulierungen herausarbeiten.

Als der wahrscheinlich früheste Einsatz der Wort-Einbettungen im Kontext der Mittelhochdeutschen Begriffsdatenbank (deren Textmaterial hier im Weiteren betrachtet wird) ist der Beitrag von Viehhauser (2020) zu erwähnen, wo sie ergänzend zur Identifizierung bedeutungsnaher Ausdrücke im Rahmen eines komplexeren Verfahrens genutzt werden, nämlich bei der Analyse der Prominenz bestimmter Wortfelder in Minnesang-Texten verschiedener Epochen bzw. einzelner Autoren (Viehhauser 2020:41–42).

1.3 Technische Grundlagen und typische Anwendungen

Die Verfügbarkeit dieser ursprünglich anspruchsvollen Technologien im Hinblick auf die Hardware- sowie Software-Anforderungen ist inzwischen maßgeblich breiter geworden, z. B. ist die Softwarebibliothek Gensim (URL 2; vgl. Řehůřek/Sojka 2010) für die Programmiersprache Python frei nutzbar und kann an üblichen Personalcomputern betrieben werden.

Die so gewonnen Wortvektoren stellen, wie schon erwähnt, eine formalisierte Repräsentation der zu Grunde liegenden Kookkurrenzverhältnisse dar. Da für die so erschlossenen Gebrauchsregularitäten auch die semantischen sowie pragmatischen Aspekte maßgeblich sind, kann man auch auf diese Domänen mittels *word embeddings* zugreifen – insbesondere im Hinblick auf gegenseitige Relationen, die sich in diesem Kontext mathematisch operationalisieren lassen.

Häufige Abfrageformen anhand der *word embeddings* gehen von einem oder mehreren Ausgangseinträgen aus, wobei auf weitere sozusagen analogisch „geschlossen“ wird. Zum gegebenen Worteintrag können so „ähnliche“ bzw. semantisch-pragmatisch „benachbarte“ Einträge gesucht werden, indem die im Vektorraum am wenigsten entfernten Koordinaten anderer Wörter identifiziert werden (i. d. R. mit Kosinus-Ähnlichkeit, anhand des Winkels zwischen den betreffenden Vektoren). Die so erfasste Nähe der Wortvektoren entspricht der Ähnlichkeit von Kookkurrenzprofilen, d. h. dem Auftreten der jeweiligen Wörter in vergleichbaren Kontexten oder Wendungen.

Es sind auch komplexere Verfahren möglich, wo die Relationen zwischen den Wortvektoren aufeinander bezogen werden, so dass auch eine Art „Transposition" möglich ist; übliche Verfahren sind dabei die Addition und Subtraktion der Vektoren. Eine Umsetzung dieses Verfahrens stellt z. B. die Ergänzung von Paaren von Wörtern, die eine gleichartige Relation zueinander aufweisen wie andere, vorgegebene Einträge. Dies ist insbesondere bei ausgeprägt strukturierten Sachbereichen gut nachvollziehbar, wie bei verschiedenen stereotypen, auch realienbezogenen Zuordnungen. Z. B. die Analogie-Ergänzung: *Frankreich : Paris = Italien : (?)*. In der einschlägigen Vektor-Arithmetik handelt es sich, schematisch, informell ausgedrückt, um das folgende Verfahren: $\vec{v}$ *(Paris)* − $\vec{v}$ *(Frankreich)* + $\vec{v}$ *(Italien)*, es wird sozusagen von ‚Paris' der Begriffsgehalt ‚Frankreich' subtrahiert und zum Ergebnis wird ‚Italien' addiert; der resultierende Vektor sollte sich bei einem gut funktionierenden Modell in der Nähe des erwarteten $\vec{v}$ *(Rom)* befinden (die eigentlichen als Ausgangspunkte eingegebenen Vektoren werden ggf. in den Ergebnissen ausgeblendet).

Ein berühmtes, häufig zitiertes Beispiel auf derselben Grundlage bezieht sich auf die englischen Personenbezeichnungen (vgl. Mikolov/Yih/Zweig 2013:746, 749): ‚king' − ‚man' + ‚woman' = ‚queen'. Diese „Gleichung" zeigt bei den morphologisch eigenständigen und größtenteils unverwandten englischen Wörtern auf eine eindrucksvolle Weise die Erfassung der semantischen Relationen, und zwar unabhängig von den ausdrucksseitigen Eigenschaften (die bei diesem konkreten Verfahren außer Acht bleiben). Anzumerken bleibt, dass die Analogie-Beziehungen dieser Art natürlich nur für bestimmte Bereiche der Lexik mit ihrer komplexen Struktur und Dynamik zutreffen. In den anschließenden Forschungsarbeiten zur Sprachverarbeitung mit den Mitteln des maschinellen Lernens werden weitere Ansätze diskutiert und anders akzentuierte Verfahren vorgeschlagen (vgl. z. B. Drozd/Gladkova/Matsuoka 2016); aus der Perspektive der linguistischen Diskursanalyse werden wiederum Vorbehalte geäußert, dass die an strukturalistischen Ansätzen basierenden Lern- und Evaluationsverfahren in Verbindung mit *word embeddings* manchen Spracherscheinungen und ihrem diskursiven Charakter nicht gerecht werden (so Bubenhofer 2020:567 f.; vgl. auch URL 7).

2. Wort-Einbettungen anhand des Textmaterials früherer Sprachstufen

2.1 Illustrative Verarbeitung der Datengrundlage der Mittelhochdeutschen Begriffsdatenbank

Im Folgenden soll versucht werden, die Anwendbarkeit dieser Verfahren für ältere Sprachstufen mit ihren überlieferten Textquellen zu überprüfen; konkret wird die Text- und Datengrundlage der Mittelhochdeutschen Begriffsdatenbank (MhdBDB) (URL 1)[2] genutzt. Mit Rücksicht auf bessere Übersichtlichkeit werden hierzu die Lemma-Grundformen anhand der datenbankinternen Lemmatisierung herangezogen, die anstelle der zugehörigen konkreten Wortformen eingelesen werden. Die Kookkurrenzverhältnisse bleiben dabei für die Auswertung erhalten, in der Darstellung der Lexeme stehen allerdings die weitgehend normalisierten mittelhochdeutschen Grundformen. Für die Behandlung der Lexik und Semantik erscheint diese Vereinheitlichung für praktische Handhabbarkeit als angemessen, da man hier z. B. die Formen- oder Phrasenbildung zunächst nicht verfolgt; insbesondere vermeidet dies aber die Zersplitterung der Worteinheiten infolge der großen Vielfalt an Form- u. v. a. Schreibvarianten, durch die die mittelhochdeutschen Texte geprägt sind.

Ein Umstand, mit dem zu rechnen ist, ist die von vornherein eingeschränkte verfügbare Textgrundlage, die sich nicht einfach erweitern lässt. Selbst bei der möglichen Heranziehung von weiteren Texten aus anderen Quellen bleibt das überlieferte und verfügbare Textmaterial einer historischen Sprachstufe notgedrungen maßgeblich begrenzt im Vergleich zur Gegenwartssprache – im vorliegenden Fall handelt es sich um mehr als 7 Millionen laufende Wortformen, die in der MhdBDB erfasst und zugleich lemmatisiert sind.[3] Zum Vergleich: Die Größe von dem oben erwähnten spezialisierten Korpus von Bubenhofer beträgt

[2] Neben der öffentlichen Web-Schnittstelle der Datenbank wird auch ihre intern verfügbare Abfragefunktionalität genutzt; für ihre Bereitstellung und allseitige fachliche Unterstützung möchte ich mich bei den Mitarbeitern der MhdBDB herzlich bedanken.

[3] Vgl. *MhdBDB – Statistik* (URL 8) (generiert am 12.11.2021). Die angegebene Zahl beläuft sich zu diesem Zeitpunkt auf 7 373 332 laufende Wörter, die lemmatisiert sind (die Gesamtsumme der Spalte „Wörter“ in der Tabelle „Status der Lemmatisierung / Disambiguierung“); nicht einbezogen werden dabei wie erwähnt nicht lemmatisierte Wortformen sowie Interpunktionszeichen.

über 12 Millionen laufende Wörter (Bubenhofer 2020:568 f.). In beiden Fällen handelt es sich um verhältnismäßig niedrige Zahlen, generell arbeitet man z. B. bei der Berechnung der *word embeddings* der Gegenwartssprache mit viel umfangreicheren Textgrundlagen.[4]

Die Lemma-Einträge werden aus den einzelnen Texten in der Reihenfolge der entsprechenden laufenden Wörter eingelesen, die Satzeinheiten werden dabei anhand der Interpunktion gegliedert (namentlich . ; ! ? : ohne Berücksichtigung von Komma, wo mehrdeutige Verwendungen vorliegen – auch in Auflistungen u. a.). Die Berechnung der Wortvektoren folgt sonst (d. h. neben der Anpassung der Darstellung an Lemma-Grundformen) dem vorgesehenen Ansatz bei Wortvektoren (word2vec), es wird die Sequenz der Satzabschnitte mit den zugehörigen Worteinheiten verarbeitet, wie es einem rohen Text weitgehend entsprechen würde. Keine zusätzlichen MhdBDB-Informationen (wie z. B. die Zuordnung von Wortklassen, Unterscheidung der Sememe, zugeordnete Begriffskategorien) werden auf dieser Ebene berücksichtigt, so dass die eigenständige Leistung des Verfahrens sichtbar ist. Aus demselben Grund wurde auf weitere mögliche „bereinigende" Modifizierung der Datengrundlage verzichtet, z. B. die Ausklammerung von „Stoppwörtern", Ausschluss von Eigennamen u. a.

Die Parameter der Berechnung wurden vorerst weitgehend bei den vorgegebenen Ausgangswerten der Softwarebibliothek Gensim, im word2vec-Modell, belassen (vgl. URL 4) (unter anderem die Anzahl der Dimensionen der Vektoren – vector_size=100; die Wortumgebung/Textfenster zur Berücksichtigung der Kookkurrenzen – window=5); gesenkt wurde jedoch die Mindestfrequenz der einbezogenen Einträge – anstelle des Ausgangswertes min_count=5 wurde die Mindestanzahl 2 angesetzt.

[4] Vgl. z. B. die Angaben zu bereits früher publizierten Modellen zum Englischen, die auf Textdaten im Umfang von Milliarden laufenden Wörtern basieren, vgl. die Daten für zwei Projekte bei Fares/Kutuzov/Oepen/Velldal (2017:273): „… English Wikipedia dump from September 2016 (about 2 billion word tokens) and Gigaword Fifth Edition (about 4.8 billion word tokens)". – Bereits 2013, bei der Publizierung des Entwurfs von Mikolov u. a. wird der Umfang von einer Billion laufenden Wörtern in Aussicht gestellt als eine zu verarbeitende Datenmenge; vgl. Mikolov/Chen/Corrado/Dean (2013:10): „...it should be possible to train the CBOW and Skip-gram models even on corpora with one trillion words ...".

2.2 Einfache Abfragen zur Wortähnlichkeit anhand der Wort-Einbettungen

In den folgenden illustrativen Proben werden die Ergebnisse einiger Abfragen dargelegt; in Anlehnung an die Terminologie der benutzten Software-Bibliothek Gensim werden unter „+“ die Wortvektoren aufgelistet, die positiv zum Ergebnisvektor beitragen, bzw. die zueinander addiert werden sollen, unter „−“ die zu subtrahierenden Wortvektoren. Die Ergebnisse werden absteigend nach Kosinus-Ähnlichkeit sortiert, d. h. zuerst werden die nächsten „Treffer“ angeführt, mit der kleinsten Winkel-Entfernung von der vorgegebenen bzw. errechneten Position im Vektorraum[5] (im gegebenen Kontext liegen die Werte zwischen 1, bei idealer Übereinstimmung, und −1, bei entgegengesetzten Vektoren; bei 0 handelt es sich um orthogonale Vektoren; diese repräsentieren völlig unähnliche, einander inhaltlich „meidende“ Wörter).

Ziemlich überzeugend erscheinen in manchen Sachgebieten einfache Auflistungen von bedeutungsnahen Ausdrücken zu einem Ausgangswort, z. B.:[6]

> +[ros] −[]: [*phert: 0.8302, swert: 0.6519, satel: 0.6481, kastelân: 0.6375, glavîn: 0.6233, harnas: 0.6100, stegereif: 0.6059, schilt: 0.6009, sper: 0.5914, hëlm: 0.5891, gewæfen: 0.5727, wâpen: 0.5691, poinder: 0.5661, halsberc: 0.5652, zoum: 0.5617, stange: 0.5594, tjoste: 0.5502, ...*]

Pferd wird als der am engsten benachbarte Ausdruck zum weitgehend synonymen *Ross* identifiziert, außerdem hebt sich dieser Eintrag auch mit dem Ähnlichkeitsscore deutlich von den weiteren aufgelisteten Wörtern ab. Bei diesen ist die weitere Abfolge nach der Ähnlichkeit durch kleinere gegenseitige Abstände geprägt. Die Beziehung zum Ausgangswort ist da generell nachvollziehbar, wobei unterschiedliche Teilbereiche zu erkennen sind – Pferdegeschirr, (Aus)Rüstung, Turnier, Waffen, Kampf u. a.

---

5 Die Vektorlänge wird in diesem Verfahren nicht berücksichtigt; der Vektorraum ist ferner nicht auf eine vorgegebene, definierte Weise segmentiert (z. B. nach sachlich-thematischen Gesichtspunkten); eine Art Orientierung im Modell ist erst durch Verhältnisse zu jeweils gewählten Bezugswörtern der Abfrage möglich.

6 Hier und im Weiteren werden die Abfrage-Ergebnisse anhand der MhdBDB Textdaten und der berechneten Wort-Einbettungen als objektsprachliche Zitate angeführt; der Umfang der Proben wird individuell je nach Darstellungszusammenhang angesetzt; die eckigen Klammern markieren hier die geordneten Listen der angeführten Einheiten. Die Ausgangseinträge der Abfragen werden jeweils am Anfang angegeben, mit Inversion des Schrift-Formats, d. h. nicht kursiviert.

Bei einem Wochentagsnamen als Ausgangswort ist die ermittelte Nachbarschaft ebenfalls sehr anschaulich und auch individuell nachvollziehbar:

> +[sunnetac] −[]: [*mittewoche: 0.8769, mântac: 0.8734, vrîtac: 0.8714, sameztac: 0.8555, phingeste: 0.8549, nône: 0.8241, ôstern: 0.7710, wîhnaht: 0.7596, phingestac: 0.7496, Iunius: 0.7375, ertac: 0.7370, Idus: 0.7364, vigilia: 0.7339, donerstac: 0.7208, calendae: 0.7201, zweinzegest: 0.7177, dienstac: 0.7082, gesungener: 0.7045, Bartholomeus: 0.6989, perchttag: 0.6942, ...*]

Neben Wochentagsnamen mit generell sehr hohen Ähnlichkeitsscores ist auch bei vielen weiteren der Bezug zu Zeit bzw. die potenzielle Verwendung bei einer Datierung naheliegend (Feiertage, liturgische Stunden, Zahlwörter, Heiligennamen u. a.).

2.3 Komplexere Abfragen

Das oben zitierte eindrucksvolle Beispiel der „Arithmetik" bei der Erfassung der semantischen Relationen mittels *word embeddings* (‚king' − ‚man' + ‚woman' = ‚queen') ließe sich auch am mittelhochdeutschen Sprachmaterial veranschaulichen:

> +[künic, vrouwe] −[herre]: [*küniginne: 0.5561, maget: 0.5152, herzoginne: 0.4687, Ginover: 0.4657, massenîe: 0.4657, vürste: 0.4637, tavelrundære: 0.4626, amîe: 0.4622, keiserinne: 0.4515, wîp: 0.4363, Kriemhilt: 0.4358, Arabel: 0.4232, ingesinde: 0.4222, vürstinne: 0.4205, Secundille: 0.4182, juncvrouwe: 0.4167, Anjou: 0.4157, minneclîch: 0.4145, tavelrunde: 0.4143, Prünhilt: 0.4141, Herzeloide: 0.4106, gotinne: 0.4095, admirât: 0.4074, knabe: 0.4022, Kudrun: 0.3995, Britanienlant: 0.3988, Helena: 0.3978, Sigelint: 0.3962, Waleis: 0.3921, wirtinne: 0.3883, marcgrævinne: 0.3857, ...*]

‚Königin' steht hier als das nächstliegende Ergebnis der „Gleichung", plausibel erscheinen da auch weitere feminine Ableitungen aus diesem Rahmen und auch die Eigennamen der hoch rangierenden Frauenfiguren der (nicht nur) mittelhochdeutschen Literatur. Bei weiteren Wörtern dieser Aufzählung könnte man mit gewisser Vereinfachung annehmen, dass für sie jeweils ein Aspekt im semantischen Gehalt dominant ist im Hinblick auf die vorliegende Abfrage (z. B. *vürste*, *admirât; wîp* für ‚herrscherlich' bzw. ‚weiblich'). In einzelnen Fällen wie *tavelrundaere*, *ingesinde*, *knabe* wäre wohl an eine Art „Begleitpersonal" im königlichen Milieu (in den zu Grunde liegenden Darstellungen) zu denken.

Anzumerken ist, dass die Abfrage der Wortähnlichkeit anhand der als ständisch unmarkiert angesehenen geschlechtsspezifischen Personenbezeichnungen *wîp* und *man* deutlich unterschiedliche, eher unerwartete Ergebnisse liefert:

> +[künic, wîp] −[man]: [*swester: 0.4710, vater: 0.4534, tohter: 0.4310, sun: 0.4300, vrouwe: 0.4186, keiser: 0.4063, Pyrrus: 0.3982, Hesiona: 0.3955, vürste: 0.3938, Risine: 0.3874, amîe: 0.3871, niftel: 0.3803, küniginne: 0.3793, oeheim: 0.3762, keiserinne: 0.3753, ...*]

Die Analogie-Arithmetik anhand der Wortvektoren scheint in diesem Fall zu einer Art Verdrängung oder Abschwächung des inhaltlichen Aspekts ‚herrscherlich' zu führen; auch die (vorgesehene) Geschlechtsdifferenzierung bzw. -restriktion bleibt weitgehend aus.[7] Was hier hingegen (jedenfalls in den hohen Scorenbereichen) zuverlässig funktioniert ist die allgemeine sachliche Bestimmung – bei allen angegebenen Wörtern handelt es sich um Personenbezeichnungen (samt Eigennamen).

Zum Vergleich kann auch die direkte Abfrage der Ähnlichkeit zum Ausgangswort herangezogen werden, die manche Überschneidungen mit den obigen Einträgen zeigt, jedoch nicht eine durchgehende Äquivalenz:

> +[küniginne] −[]: [*keiserinne: 0.8366, herzoginne: 0.7931, juncvrouwe: 0.7638, maget: 0.7021, wirtinne: 0.6995, vrouwe: 0.6584, vürstinne: 0.6117, marcgrævinne: 0.6016, Kriemhilt: 0.5962, tohter: 0.5808, Arabel: 0.5638, amîe: 0.5637, minneclîch: 0.5624, Kudrun: 0.5624, Giburc: 0.5618, burcgrævinne: 0.5508, grævinne: 0.5338, swester: 0.5318, Isolde: 0.5293, gotinne: 0.5285, ingesinde: 0.5275, Prünhilt: 0.5179, sældebernde: 0.5114, Helena: 0.5106, Uote: 0.5105, ...*]

In lexikalisch-begrifflicher Perspektive erscheinen einige dieser Fälle wohl als dunkle Zuordnungen; erst mit Berücksichtigung des kookkurrenz-basierten Verfahrens sind sie eher nachvollziehbar; die so modellierten Gebrauchsregularitäten sind da gattungsspezifisch geprägt (*tavelrundaere* ist – nicht überraschend –

[7] Eine mögliche Erklärung für diese Erscheinung könnte in dem Umstand gesehen werden, dass die ständisch markierten Ausdrücke den Herrscherbezeichnungen auch im (textuell erfassten) Sprachgebrauch generell näher stehen; dem entsprechen auch die Kosinus-Ähnlichkeitsscores der betreffenden Wörter im benutzten word2vec-Modell: (*künic*, *herre*): 0.1471; (*künic*, *man*): 0.1107; (*küniginne*, *vrouwe*): 0.6584; (*küniginne*, *wîp*): 0.2841 (besonders deutlich ist der Unterschied bei den weiblichen Bezeichnungen); zu beachten ist ferner die stärker ausgeprägte Mehrdeutigkeit von *man*).

vorwiegend in der Artusepik belegt, auch der viel häufigere Ausdruck *knabe* erscheint bevorzugt in den Ritterepen).[8]

Zu den Lieblingsthemen der (nicht nur) historischen Semantik und Lexikologie, denen im Rahmen unterschiedlicher Forschungsansätze Aufmerksamkeit gewidmet wird, zählen wohl aus mehreren Gründen die Verwandtschaftsbezeichnungen.[9] Sie bilden ein reiches, differenziertes Gefüge, in dem vielfältige und (zumindest potenziell) eindeutige Abgrenzungen möglich sind; zugleich erfuhren einige Teilbereiche durchgreifende Wandlungen, bei denen sich unterschiedliche Faktoren auswirkten, sodass genug Spielraum bzw. Untersuchungsmaterial für mehrere Deutungsansätze zur Verfügung steht. Hinzu kommt, dass zumindest einige der zu Grunde liegenden v. a. sozialen Beziehungen eine weitere Geltung haben und oft übereinzelsprachlich vergleichbar sind.

Vor diesem Hintergrund bieten sich die Verwandtschaftsbezeichnungen auch für eine illustrative Darlegung der hier diskutierten Funktionalität der Wortvektoren an. Die bereits vorgestellte Arithmetik lässt sich z. B. an einigen geschlechtsspezifischen Verwandtschaftsbezeichnungen veranschaulichen:

> +[bruoder, wîp, vrouwe] −[man]: [*swester: 0.7243, tohter: 0.6753, niftel: 0.6098, sun: 0.5956, maget: 0.5825, amîe: 0.5818, muoter: 0.5674, juncvrouwe: 0.5562, amîs: 0.5555, gespil: 0.5491, hûsvrouwe: 0.5432, vriundinne: 0.5311, neve: 0.5260, vater: 0.5251, gemahel: 0.5227, geselle: 0.5210, muome: 0.5186, ...*]

Der nächstliegende erste Eintrag ‚Schwester' entspricht auch der intuitiven Einschätzung bzw. Erwartung, gemäß der gegebenen „Gleichung" für eine weibliche Entsprechung zu ‚Bruder'. Zu bemerken ist, dass zumindest in diesem Abschnitt der Wörter mit den höchsten Ähnlichkeitsscores die grundlegenden Gemeinsamkeiten erfolgreich modelliert sind – bei allen diesen Einträgen handelt es sich um Personenbezeichnungen. Ansonsten zeigt sich auch hier der bereits kommentierte Umstand, wo jeweils nur einige Aspekte im semantischen Gehalt (im Hinblick auf die Abfrage) als dominant erschlossen sind – hier etwa ‚weibliches Geschlecht' oder ‚Verwandtschaftsbeziehung'.

Zu bemerken ist, dass alternative Abfragen mit anderen Kombinationen der in Frage kommenden Ausgangswörter zu teilweise anderen Ergebnissen führen.

---

8 Vgl. MhdBDB – Textsuche: *tavelrundære* (URL 9); MhdBDB – Textsuche: *knabe* (URL 10).

9 Vgl. stellvertretend Fritz (1974:30–36); Ruipérez (1984); Löbner (2015:247–261).

Die nächstliegende „Paraphrase“ mit minimalen benötigten geschlechtsspezifischen Ausdrücken wäre wohl:

> +[bruoder, wîp] −[man]: [*sun: 0.6767, swester: 0.6584, vater: 0.6075, vetere: 0.6013, tohter: 0.5914, neve: 0.5863, oeheim: 0.5693, hûsvrouwe: 0.5562, niftel: 0.5279, muome: 0.5222, muoter: 0.5131, geselle: 0.4827, mâc: 0.4765, kone: 0.4685, ...*]

Dies führt beim vorliegenden Modell anscheinend zu einer Asymmetrie – wider Erwarten zeigen mehrere Bezeichnungen männlicher Verwandter höhere Scores.

Eine andere Möglichkeit stellt die vollständigere Auflistung der Ausgangswörter mit geschlechtsspezifischen allgemeinen Personenbezeichnungen dar, teilweise auch standesspezifisch; zum Ansatz vgl. Drozd/Gladkova/Matsuoka (2016:3520–3521), Viehhauser (2020:41–42):

> +[bruoder, wîp, vrouwe] −[man, herre]: [*swester: 0.6297, tohter: 0.6044, maget: 0.5972, amîe: 0.5747, Aglye: 0.5230, muoter: 0.5168, Isolde: 0.5018, Polixena: 0.4902, Alize: 0.4844, gespil: 0.4837, Hesiona: 0.4834, juncvrouwe: 0.4762, sun: 0.4755, kone: 0.4742, amîs: 0.4740, niftel: 0.4695, ...*]

Hier wird die Geschlechtsspezifität der Wörter viel klarer differenziert, auf der anderen Seite kommen sozusagen im Gefolge von ‚Frau‘ mehrere individualisierte Ausdrücke, nämlich die Eigennamen einzelner Frauenfiguren und ferner mehrere courtoise Personenbezeichnungen.

Die direkten „Wortnachbarn“ zu *Schwester* als Ausgangswort zeigen mit den oben angeführten analogischen Annäherungen manche Überschneidungen, jedoch auch Besonderheiten:

> +[swester] −[]: [*tohter: 0.8678, muome: 0.7653, sun: 0.7469, muoter: 0.7169, base: 0.6893, hûsvrouwe: 0.6848, bruoder: 0.6702, vater: 0.6696, kone: 0.6551, gemahelen: 0.6306, niftel: 0.6299, vetere: 0.6278, Elisabeth: 0.6199, oeheim: 0.6180, neve: 0.6128, Margarete: 0.6116, vriundinne: 0.6111, amîe: 0.6092, gemahel: 0.6030, ...*]

Die Wortanalogien, wo die gesuchte Relation aus einem Paar der „Musterwörter“ erschlossen und zur Vervollständigung des zweiten Paares von gegebenem Ausgangspunkt projiziert wird, sind generell sensitiv auf etwaige Asymmetrien, insbesondere Polysemie u. a. In der Regel sind die im Modell miterfassten Relationen komplexer, als für eine möglichst geradlinige Analogie-Zuordnung effektiv erschließbar ist. Je nach Anforderungen werden unterschiedliche Strategien angesetzt – eine naheliegende Möglichkeit ist dabei z. B.

mehrere „Musterpaare“ auszuwerten, wodurch eine Art verallgemeinerte Klassen konstruiert werden (vgl. Drozd/Gladkova/Matsuoka 2016:3520–3521). Ein solcher Ansatz kann auch bei der hier benutzten Gensim-Implementierung durch Eingabe von mehreren zu berechnenden Wortvektoren in die Abfrage genutzt werden, wie einige der angeführten Probebeispiele zeigen.

Bei der Einschätzung des Aussagewertes der Abfrageergebnisse auf dieser Grundlage sind ferner die Fälle mit sehr kleiner Entfernung der Ähnlichkeitsscores zu beachten, was die geradlinige Interpretation des nächstliegenden Eintrags als der einzig richtigen Antwort problematisch erscheinen lässt (vgl. Drozd/Gladkova/Matsuoka 2016:3527). In den vorliegenden Anwendungen wird vorgezogen, stets mehrere benachbarte Ergebnisse der Abfragen zu berücksichtigen, und nach Möglichkeit zu kommentieren (etwa in Fällen, wo zunächst überraschende, scheinbar wenig plausible Ergebnisse auf erkennbare Eigenschaften der Textgrundlage, den gattungsspezifischen Gebrauch, Mehrdeutigkeit u. a. zurückgeführt werden können).

2.4 Erfassung von semantisch benachbarten Ausdrücken anhand der Begriffskategorien im Vergleich zu Wortvektoren

Die semantische Annotation der Mittelhochdeutschen Begriffsdatenbank ist für verallgemeinerte Suchabfragen nach Bedeutungskategorien (einer Art Semen) bestimmt, man kann auf diese Weise auch inhaltlich benachbarte Lemmata bzw. Sememe erschließen, was dem Anliegen des angesprochenen, auf Wortvektoren basierenden Verfahrens nahekommt. Im Folgenden sollten beide Herangehensweisen einander gegenübergestellt werden. In einem geradlinigen Ansatz wird nach den Worteinträgen gesucht, in denen alle Begriffskategorien des Ausgangswortes enthalten sind (möglich wäre aber auch eine feinere Handhabung bzw. Gewichtung der einzelnen Kategorien).

Zur Veranschaulichung der Ergebnisse dieses Verfahrens soll ein elementares Beispiel mit dem Ausgangswort *bier* gezeigt werden, dessen Annotation mit zwei Begriffskategorien problemlos übersichtlich ist:

Ausgangswort: *bier*, Kategorien: *Getränke (21111307)*, *Genussmittel (21111400)*.[10]
Liste der Lemmata mit beiden enthaltenen Begriffskategorien:

afterbier, alantwîn, bier, biergelte, biersupper, briuhûs, briuwære, brüeje, ebrius, entrinken, entrinkez, erglaffen, griuzinc, heilwîn, honecmete, hovewîn, inebriare, kannenwirt, kipper, klârêt, lantwîn, lîtgebe, lîtgebinne, lîthus, lîtkouf, lûtertranc, malvasier, mete, minnetranc, molle, môraz, most, muglære, muscatel, ôsterwîn, pinôl, potare, reinval, rîtmâz, schavernac, sirop, slic, sûfære, sûfen, taberna, tabernieren, terran, tobetrunken, trenkære, tribian, trinkære, trinken, trinkens, trinkenz, trinkgeselle, trunken, trunkenbolt, trünkern, übertrinken, ungetrunken, vertrinken, vertrinks, verwepfen, vindeplan, vinum, Weindorf, Weinman, welschwîn, wîn, wîngerwe, wînglocke, wîngrabe, wîngülte, wînholz, wînhûs, wînkar, wînkezzel, wînkouf, wînköufel, wînrîche, wînschenke, wînslûch, wînstoc, wînviuhte, wînvuore, wînzelle, wînzürl, wînzürlgerihte, zetrinken, ziperwîn, zûberwîn.

Die aufgelisteten Lemmata lassen ihre Zugehörigkeit oder Nähe zum Ausgangswort in den meisten Fällen gut erkennen – es handelt sich um weitere Getränke, besonders differenziert beim Wein, ferner um das Bierbrauen (sowie Weinbau), das Gaststättenwesen und den Trinkgenuss. Auch bei der erwähnten eher sparsamen Annotation erscheinen die Zuordnungen hier generell plausibel. (Die einzelnen aufgelisteten Ausdrücke weisen neben den beiden Ausgangssemen von *bier* in vielen Fällen weitere spezifizierende „Seme“ auf.)

Anhand der Wortvektoren werden im vorliegenden Modell (wie in oben angeführten Beispielen) die folgenden nächsten „Wortnachbarn“ zu *bier* identifiziert:

+[bier] −[]: [*môraz: 0.9131, klârêt: 0.8992, lûtertranc: 0.8831, mete: 0.8672, rou: 0.8608, buter: 0.8579, sirop: 0.8560, gebratener: 0.8509, rocke: 0.8412, most: 0.8374, quiten: 0.8361, lebezelte: 0.8345, gebranter: 0.8325, ezzich: 0.8314, kanne: 0.8285, wînbere: 0.8282, büte: 0.8279, wine: 0.8275, zwibolle: 0.8254, winzic: 0.8250, mandel: 0.8237, ingwer: 0.8230, salbei: 0.8225, sulze: 0.8220, gesotenez: 0.8218, wiltbrât: 0.8199, kochen: 0.8197, condimentum: 0.8182, siedez: 0.8171, gestôzener: 0.8170, gesalzen: 0.8166, gerîbener: 0.8162, senef: 0.8161, betroufen: 0.8146, einber: 0.8133, aschlouch: 0.8116, gevültez: 0.8106, kuoche: 0.8095, salse: 0.8094, semele: 0.8093, vîge: 0.8080, anîz: 0.8063, hûsenblâter:*

[10] Vgl. MhdBDB – Wortindex: *bier* (URL 11); MhdBDB – Wortindex: *21111307&21111400* (URL 12).

0.8061, vierdunc: 0.8056, morhe: 0.8022, gerste: 0.8020, tranc: 0.8018, kumpost: 0.8007, [… 94 weitere Einträge…]*, wîn: 0.7493 …*]

In einer relativ engen Nachbarschaft der Bezeichnung des traditionellen Genussmittels befinden sich anhand der Abfrage zahlreiche Ausdrücke für Getränke, Gerichte, Gewürze u. a. m. (vgl. die hohen Scores der Vektorähnlichkeit), hierzu gehören auch zusammengehörende verbale Bezeichnungen, Adjektive u. a. In diesem Fall ist die inhaltliche Abgrenzung breiter, als dies bei den Begriffskategorien möglich ist – hier *Getränke (21111307)*, *Genussmittel (21111400)*, was eine weiter gehende Spezifizierung mit sich bringt.

Selbst bei der vorgenommenen punktuellen Gegenüberstellung der beiden Verfahren zur Erfassung der semantischen Nähe lassen sich manche gleichartigen Ergebnisse beobachten, unverkennbar sind jedoch auch die Unterschiede. Viele von ihnen lassen sich auf die bearbeitungspragmatischen Aspekte (eine überwiegend manuelle Annotation einerseits bzw. eine kookkurrenzbasierte Extraktion anhand der Textdaten andererseits) zurückführen. Die Begriffskategorien – „Seme" werden auf der Ebene der Lemmata bzw. deren Sememe zugeordnet – z. B. ihre einzelnen Verwendungskontexte oder Vorkommenshäufigkeiten bleiben weitgehend außer Betracht (diese Aspekte können allerdings in die Annotation eingehen, z. B. durch Ansatz neuer Sememe oder bei der Disambiguierung). Ein auffälliger Unterschied besteht in der Fülle der Wortbildungskonstrukte, die in vielen Fällen die Seme der Ausgangslemmata übernehmen, so dass sie anhand der übereinstimmenden Seme identifiziert werden; im Hinblick auf den Textgebrauch handelt es sich aber oft um Randerscheinungen. Bei den durch Wort-Einbettungen erfassten benachbarten Ausdrücken sind hingegen die Verwendungskontexte geradezu bestimmend. Die Wortnachbarschaft ist in diesem Fall viel breiter, was einerseits durch die besonders reiche, differenzierte belegte Lexik in diesen Sachgebieten bedingt ist, andererseits durch sehr konkrete, spezifische Kategorien des Begriffssystems, die hier in der vorangehenden Probe angesetzt sind.

2.5 Formalisierter Vergleich der Texte mittels Wort-Einbettungen

Die konzeptuellen sowie rechentechnischen Grundlagen der Wort-Einbettungen lassen sich in verschiedenen Ausprägungen und Bereichen der Datenauswertung nutzen. Eine Möglichkeit ist auch die Vektorenberechnung für ganze Texteinheiten („Dokumente"), wodurch die gegenseitigen Ähnlichkeiten erfasst werden;

auch diese Funktionalität wird in der Softwarebibliothek Gensim bereitgestellt (vgl. URL 3). In diesem Rahmen wird dieselbe Datengrundlage wie bei den obigen Wortvektoren verwertet, nämlich die Lemma-Grundformen für alle Texte der MhdBDB, als Score der Textnähe dient hier wieder die Kosinus-Ähnlichkeit der Dokument-Vektoren.[11]

Als besonders effektiv erweist sich dieses Verfahren bei der Identifizierung der womöglich gleichen Texte aus verschiedenen Quellen, Bearbeitungen oder Editionen, die in einigen Fällen in der Datenbank parallel enthalten sind (z. B. bei Stricker) bzw. bei mehreren handschriftlichen Versionen u. a. (Ackermann, Nibelungenlied…), vgl. zum letztgenannten berühmten Heldenepos:[12]

> +[NLC – Nibelungenlied (C) (Hs. C) (ANONYM); Heldenepik] −[]: [*(NLB – Nibelungenlied (B/C) (Bartsch/deBoor) (ANONYM); Heldenepik, 0.9251), (NBB – Nibelungenlied (Hs. B) (nach Batts) (ANONYM); Heldenepik, 0.9059), (NLA – Nibelungenlied (Hs. A) (Batts) (ANONYM); Heldenepik, 0.7975), (BRF – Biterolf und Dietleib (ANONYM); Heldenepik, 0.7149), (KLA – Die Klage (ANONYM); Heldenepik, 0.7036), (KU – Kudrun (ANONYM); Heldenepik, 0.6617), (ERB – Herzog Ernst (Hs. B) (ANONYM); Spielmannsepik, 0.5109), ...*]

In mehreren Fällen werden Gattungsgruppen erkennbar, die bereits in den obigen Beispielen belegt sind, seltener auch Autorenœuvres (ggf. gebunden an Gattungen – Konrad von Würzburg, die Epen Hartmanns von Aue…):

> +[AX – Alexius (KONRAD VON WÜRZBURG); Religiöse Versdichtung] −[]: [*(SL – Silvester (KONRAD VON WÜRZBURG); Religiöse Versdichtung, 0.8008), (PRT – Partonopier und Meliur (KONRAD VON WÜRZBURG); Aventiurenroman, 0.7152), (PT – Pantaleon (KONRAD VON WÜRZBURG); Religiöse Versdichtung, 0.7070), (TRO – Der Trojanische Krieg (KONRAD VON WÜRZBURG); Klassische Epen, 0.6255), (HZ – Herzmaere (KONRAD VON WÜRZBURG); Kleinere Erzählungen, Fabeln und Lehrgedichte,*

[11] Bei der Berechnung der Dokument-Vektoren wurden weitgehend die impliziten Werte der Gensim-Bibliothek übernommen (vgl. gensim.models.doc2vec.Doc2Vec, URL 3), es gilt für die Zahl der Vektor-Dimensionen vector_size=100 sowie das Textfenster für die berücksichtigte Wortumgebung window=5; erhöht wurde die Anzahl der Berechnungsdurchgängen des Modells epochs=40 (gegenüber dem impliziten Wert 10).

[12] Die Angaben zu den einzelnen Werken (Autoren, Gattungen, ggf. Sammlungen, Editionen, Textquellen u. a.) sind der Datenbank entnommen, vgl.: MhdBDB – Text List (URL 13) – in den einzelnen Teilseiten, verlinkt von den Textsiglen und Autorennamen.

0.6174), (ENG – Engelhard (KONRAD VON WÜRZBURG); Religiöse Versdichtung, 0.6142), (SW – Der Schwanritter (KONRAD VON WÜRZBURG); Kleinere Erzählungen, Fabeln und Lehrgedichte, 0.6092), (NT – Das Turnier von Nantes (KONRAD VON WÜRZBURG); Höfischer Roman, 0.5229), ...]

+[ER – Erec (HARTMANN VON AUE); Artusdichtung] −[]: [*(IW – Iwein (HARTMANN VON AUE); Artusdichtung, 0.7949), (WGL – Wigalois, der Ritter mit dem Rade (WIRNT VON GRAVENBERC); Artusdichtung, 0.7315), (GR – Gregorius (HARTMANN VON AUE); Religiöse Versdichtung, 0.7117), (GL – Gauriel von Muntabel (KONRAD VON STOFFELN); Artusdichtung, 0.7040), (LZ – Lanzelet (ULRICH VON ZATZIKHOVEN); Artusdichtung, 0.6601), (TAN – Tandareis und Flordibel (DER PLEIER); Artusdichtung, 0.5469), (MEL – Meleranz (DER PLEIER); Artusdichtung, 0.5348), ...*]

Neben dem oben illustrierten Vergleich von ganzen Texten ergibt sich eine interessante Möglichkeit ferner bei gemeinsamen Abfragen zu Dokumentvektoren und den bei ihnen intern verfügbaren Wortvektoren. Bei geeigneten Bedingungen können so auch gewisse benachbarte Wortschatzbereiche im Hinblick auf konkrete Texte erfasst werden, z. B. bei einer thematisch sowie gattungsbedingt relativ deutlich ausgeprägten und abgegrenzten Gruppen der Sachtexte, etwa Kochbücher:

+[ABS – Ein alemannisches Büchlein von guter Speise (ANONYM); Kochbücher] −[]: [*KBL4 – Das Reichenauer Kochbuch aus der Badischen Landesbibliothek (Ka1) (ANONYM); Kochbücher: 0.8641, KBL3 – Inntalkochbuch (ANONYM); Kochbücher: 0.7941, KDO – Das Kochbuch aus dem Deutschen Orden (ANONYM); Kochbücher: 0.7881, HUB1 – Cpg. 551, 186r–196v (H2) (ANONYM); Kochbücher: 0.7720, HUB3 – Kochrezeptsammlung des cpg 583 (ANONYM); Kochbücher: 0.7711, MBS1 – Münchner Kochbuchhandschriften aus dem 15. Jahrhundert, Cgm 811 (ANONYM); Kochbücher: 0.7500, MBS2 – Münchner Kochbuchhandschriften aus dem 15. Jahrhundert, Clm 15632 (ANONYM); Kochbücher: 0.7500, HUB2 – Cpg. 551, 197r–204r (H2) (ANONYM); Kochbücher: 0.7287, MBS5 – Münchner Kochbuchhandschriften aus dem 15. Jahrhundert, Cgm 725 (ANONYM); Kochbücher: 0.7276, DES2 – Daz buoch von guoter spise (Adamson) (ANONYM); Kochbücher: 0.7113, GSP – Daz buoch von guoter spîse (Gloning) (ANONYM); Kochbücher: 0.7067, ...*];

[*speck: 0.6935, wînbere: 0.6865, pêterlîn: 0.6710, værbs: 0.6667, galgan: 0.6595, areweiȝ: 0.6508, salbei: 0.6464, zwibolle: 0.6420, würzen: 0.6413, bachen: 0.6350, eiertoter: 0.6277, samenære: 0.6250, bachs: 0.6220, wëcholter: 0.6205, smalz: 0.6170, phanne: 0.5948, safrân: 0.5913,*

kalpvleisch: 0.5901, ingwer: 0.5889, lb: 0.5881]; [*Thadaeus: 0.5846, Plat: 0.5593, Daemon: 0.5572, Füllensac: 0.5485, Telestes: 0.5435, Alteclere: 0.5390, Nagengast: 0.5363, Mistelbach: 0.5326, Mathan: 0.5239, Ruofach: 0.5227, Herburc: 0.5186, Frankenlant: 0.5173, Schetzin: 0.5147, Innsbruck: 0.5134, Fell: 0.5105, Glockenitz: 0.5099, Nimmervol: 0.5080, Anzman: 0.4995, Slangenzagel: 0.4989, Lepp: 0.4966*]

Die Identifizierung der gleichartigen Texte derselben oder ähnlicher Gattungen kann in diesem Fall als überzeugend bezeichnet werden, auch die vektormäßig benachbarten Einzelwörter erscheinen hier inhaltlich weitgehend plausibel (Lebensmittel, Gewürze, Getränke, Kochgeschirr u. a.); bei den eigens aufgelisteten Eigennamen ist jedoch trotz hoher Ähnlichkeitsscores ihr Aussagewert kaum auszumachen – es handelt sich oft um Personen- oder Ortsnamen mit niedriger Frequenz, bei denen die kookkurrenzbasierte Erfassung stark von den Einzelbelegen abhängt. Zu bemerken ist ferner, dass diese Art der Zuordnung von Lexemen zu Texten in manchen Fällen viel weniger nachvollziehbare Ergebnisse liefert, so v. a. bei umfangreichen, thematisch vielfältigen Werken. Zu beachten ist auch der Umstand, dass die Erfassung des textspezifischen oder (wie auch immer aufzufassenden) charakteristischen Wortschatzes mit weniger komplexen, z. B. textstatistischen Ansätzen erreichbar ist.

3. Abschließende Überlegungen – Einsatzmöglichkeiten der Wort-Einbettungen in der Infrastruktur der MhdBDB

Die *word embeddings* erscheinen als eine wertvolle Erweiterung der traditionelleren Verfahren der Textverarbeitung und -analyse. Während manche Bereiche nach wie vor Domäne von höchst spezialisierten Forschungen der Computer- oder Informationswissenschaft u. a. darstellen, ist durch die freie Verfügbarkeit der Softwarewerkzeuge eine breitere individuelle Nutzung in verschiedensten Gebieten möglich geworden. Einige der bisher veröffentlichten online Anwendungen haben teilweise den Charakter einer Technologie-Demonstration,[13] es gibt aber auch komplexere Infrastrukturen.[14]

[13] Vgl. z. B.: Word embedding demo (URL 6) – für Finnisch und Englisch.
[14] Vgl. WebVectors: word embeddings online (URL 5) – für Englisch und Norwegisch.

Auch in einem philologisch angelegten Informationssystem wie der Mittelhochdeutschen Begriffsdatenbank zeichnen sich dabei manche Nutzungsmöglichkeiten ab, zumal die für sie zentralen lexikalisch-semantischen Fragestellungen den naheliegenden Einsatzbereich der *word embeddings* darstellen. Die hier präsentierten Untersuchungen nutzen die intern verfügbaren Abfragemöglichkeiten der MhdBDB und erfordern anschließende Verarbeitung der Daten zum word2vec- bzw. doc2vec-Modell, an dem die Abfragen erst möglich sind. Eine allgemein nutzbare Funktionalität aufgrund dieser Technologie wäre als eine weitere Option zu den Abfragemodi ggf. einzubauen, vielleicht im Anschluss an den aktuellen „Relaunch" der MhdBDB, bei dem eine technologische Erneuerung und benutzerseitige Funktionserweiterung im Vordergrund stehen (vgl. Hinkelmanns/Zeppezauer-Wachauer 2020:74, 79 ff.; Brom 2019:175 f.).

Eine Suche mittels Wortvektoren wäre wohl eher als eine eigenständige Abfrageebene bzw. Zugriffsmodus auf die MhdBDB-Daten einzusetzen. Eine solche autonome Teilfunktion könnte z. B. eine Ermittlung von „bedeutungsnahen" Ausdrücken zum gegebenen Eintrag besorgen – anhand von *word embeddings* oder anhand der vorhandenen Begriffskategorien. Dabei ist mit unterschiedlichen Stark- sowie Schwachstellen zu rechnen. Bei vollständiger individueller Verarbeitung von Texten (Lemmatisierung und Disambiguierung, ggf. mit Erweiterung der Wörterbuch-Datenbank um neu belegte Einträge) liegen auch bei seltenen bzw. auch einmaligen Worteinheiten „menschlich" bestätigte Annotationen vor. Für automatische Berechnung von Wortvektoren sind hingegen niedrigfrequente Wörter generell problematisch (in den praktischen Anwendungen werden die seltenen Einträge sogar oft ausgeklammert).[15] Auf der anderen Seite könnten die *word embeddings* automatisch auch anhand der Rohtexte berechnet werden, sodass sie auch für Wörter bzw. Textstellen vorliegen, die nicht manuell annotiert sind (in den hier angeführten Proben werden zur besseren Übersichtlichkeit nur lemmatisierte Wortformen benutzt, es könnte sich aber genauso um Texte ohne solche Aufbereitung handeln). Individuell unterschiedlich fällt letztlich auch die Differenziertheit der begrifflichen Annotation (Anzahl der zugeordneten Begriffskategorien, ggf. mit Unterscheidung von Sememen und ihre Disambiguierung für einzelne Textstellen). Die berechneten Wortvektoren zeigen manchmal feinere Differenzierungsmöglichkeiten (z. B. bei der Identifizierung von semantisch am nächsten benachbarten weiteren Wörtern). Durch

[15] Die implizite Mindestfrequenz wäre in der entsprechenden Gensim-Funktion min_count=5 (vgl. URL 4); für die vorliegende Auswertung der MhdBDB wurde diese Schwelle im Hinblick auf den Gesamtumfang der Textdaten auf 2 gesenkt.

Satzkookkurrenzen kommen auch syntaktisch-semantische Aspekte mit ins Spiel, sodass manchmal Zusammenhänge „erkannt“ werden, die anhand der aktuellen Markierung durch Begriffskategorien nicht fassbar sind. In Sachgebieten, für die das zu Grunde liegende Begriffssystem reich differenziert ist, ist die semantische Annotation auf dieser Basis hingegen genauer.

In den vorliegenden eher einführenden Verarbeitungsproben mittels Gensim bleiben weite Teile der MhdBDB-Daten unberücksichtigt, insbesondere die Unterscheidung der Sememe innerhalb der Lemmata; gerade diese Ebene ließe sich potenziell einbeziehen (für entsprechend annotierte Texte bzw. Textabschnitte). Gut vorstellbar ist auch der Einsatz von *word embeddings* zur (partiellen) Überprüfung der manuellen Annotationen (z. B. Lemmatisierung bzw. Disambiguierung – zumindest bei häufigeren Einträgen, wo auf diese Weise vielleicht auffällige Abweichungen oder etwaige Verwechslungen ermittelt werden könnten).

Eine direkte Kombinierbarkeit mit Datenbank-Abfragen wäre ggf. auch denkbar, es würde allerdings eine Erweiterung der Abfrage-Syntax erfordern (z. B. Suche nach gegebenen, den üblichen MhdBDB-Suchkriterien genügenden Ausdrücken „samt word2vec-Nachbarn“ – bis zu einer bestimmten Vektor-Entfernung). Dabei müsste die Kombinierbarkeit der beiden Suchebenen geregelt werden, was ggf. bei komplexeren Suchanfragen nicht trivial wäre.

Insgesamt erweisen sich die hier in einigen Stichproben präsentierten, auf Wortvektoren basierenden Analyseverfahren als vielseitig nutzbare, wertvolle Hilfsmittel für eine philologische Auswertung sprachlicher Daten, wobei auch historische Sprachstufen keine Ausnahme darstellen. Der notgedrungen eingeschränkte Umfang des Quellenmaterials im Vergleich zu gegenwartssprachlichen Textdaten sowie weitere Spezifika dieser Texte sind bei der Auswertung zu berücksichtigen, stellen jedoch, wie es scheint, keine grundsätzlichen Hindernisse für den Einsatz dieser Methoden dar. Als eines der Desiderate in diesem Bereich könnte vielleicht eine direkt verfügbare öffentliche Benutzerschnittstelle für diese Daten angeführt werden, am ehesten als Bestandteil vorhandener Informationssysteme. Im Fall der MhdBDB, könnten die laufenden Innovationsarbeiten als eine Gelegenheit angesehen werden, bei dem inzwischen traditionsreichen altgermanistischen Online-Werkzeug auch diese aktuellen Errungenschaften der „künstlichen Intelligenz“ in der Sprachverarbeitung den interessierten Benutzerkreisen bereitzustellen.

Literaturverzeichnis

Primärliteratur:

Die Textquellen und die zugehörigen Annotationsdaten anhand der Mittelhochdeutschen Begriffsdatenbank (URL 1); vgl. die Übersicht und die verlinkten Einzeleinträge: MhdBDB – Text List (URL 13); bei Bedarf werden konkrete Datenbank-Abfragen in den Fußnoten spezifiziert.

Sekundärliteratur:

BROM, Vlastimil (2019): Die Mittelhochdeutsche Begriffsdatenbank als ein vielseitiges Arbeitsinstrument zur Analyse älterer deutschsprachiger Texte. In: *Brünner Beiträge zur Germanistik und Nordistik*, Nr. 33, Supplementum, Brno, S. 173–184. Zugänglich unter: https://digilib.phil.muni.cz/handle/11222.digilib/142280 [18.11.2021].

BUBENHOFER, Noah (2020): Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings. In: *Zeitschrift für germanistische Linguistik*, Nr. 48, 3, S. 562–589.

BUBENHOFER, Noah (2018): Serialität der Singularität. Korpusanalyse narrativer Muster in Geburtsberichten. In: *Zeitschrift für Literaturwissenschaft und Linguistik*, Nr. 48, 2, S. 357–388.

DORNSEIFF, Franz (Begr.) (2020): *Der deutsche Wortschatz nach Sachgruppen*, 9., überarbeitete und erweiterte Auflage, bearbeitet von QUASTHOFF, Uwe. Berlin; Boston.

DROZD, Aleksandr / GLADKOVA, Anna / MATSUOKA, Satoshi (2016): Word Embeddings, Analogies, and Machine Learning: Beyond king − man + woman = queen. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* Osaka, S. 3519–3530. Zugänglich unter: https://www.semanticscholar.org/paper/Word-Embeddings%2C-Analogies%2C-and-Machine-Learning%3A-%2B-Drozd-Rogers/686b52953471a9d7a515215ba54ad0350c6b0472 [18.11.2021].

GÖRZ, Günther / SCHMID, Ute / BRAUN, Tanya (Hrsg.) (2021): *Handbuch der Künstlichen Intelligenz.* 6. Aufl. Berlin; Boston.

HINKELMANNS, Peter / ZEPPEZAUER-WACHAUER, Katharina (2020): ez ist ein wârheit, niht ein spel, daz netze was sinewel: Die MHDBDB im Semantic Web. In: FISCHER, Martin (Hrsg.): *Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen: Akten der Tagung Bamberg, 08.–10. November 2018 (Bamberger interdisziplinäre Mittelalterstudien, Bd. 15).* Bamberg, S. 73–86. Zugänglich unter: https://fis.uni-bamberg.de/handle/uniba/48993 [18.11.2021].

FARES, Murhaf / KUTUZOV, Andrei / OEPEN, Stephan / VELLDAL, Erik (2017): Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: TIEDEMANN, Jörg (Hrsg.): *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017*. Linköping, S. 271–276.

FRITZ, Gerd (1974): *Bedeutungswandel im Deutschen. Neuere Methoden der diachronen Semantik.* Tübingen.

LÖBNER, Sebastian (2015): *Semantik. Eine Einführung.* Berlin; Boston.

MIKOLOV, Tomas / YIH, Wen-tau / ZWEIG, Geoffre (2013): Linguistic Regularities in Continuous Space Word Representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta: Association for Computational Linguistics*, S. 746–751. Zugänglich unter: https://aclanthology.org/N13-1090 [18.11.2021].

MIKOLOV, Tomas / CHEN, Kai / CORRADO, Greg / DEAN, Jeffrey (2013): Efficient Estimation of Word Representations in Vector Space. In: *arXiv.org*. Zugänglich unter: https://arxiv.org/abs/1301.3781 [18.11.2021].

RUIPÉREZ, Germán (1984): *Die strukturelle Umschichtung der Verwandtschaftsbezeichnungen im Deutschen. Ein Beitrag zur historischen Lexikologie, diachronen Semantik und Ethnolinguistik.* Marburg.

ŘEHŮŘEK, Radim / SOJKA, Petr (2010): Software Framework for Topic Modelling with Large Corpora. *In: Proceedings of LREC 2010, workshop New Challenges for NLP Frameworks.* Valletta, S. 46–50. Zugänglich unter: https://is.muni.cz/publication/884893/en/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka [18.11.2021].

VIEHHAUSER, Gabriel (2020): Mittelalterliche Texte als Modellierungsaufgabe. In: FISCHER, Martin (Hrsg.): *Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen: Akten der Tagung Bamberg, 08.–10. November 2018 (Bamberger interdisziplinäre Mittelalterstudien, Bd. 15)*. Bamberg, S. 15–50. Zugänglich unter: https://fis.uni-bamberg.de/handle/uniba/48993 [18.11.2021].

Internetquellen:

URL 1: *MhdBDB – Mittelhochdeutsche Begriffsdatenbank.* http://mhdbdb.sbg.ac.at/ [18.11.2021] (Einzelne Teilseiten bzw. Datenbankabfragen werden jeweils in den Fußnoten spezifiziert.).

URL 2: *Gensim – Topic modelling for humans.* https://radimrehurek.com/gensim/ [18.11.2021].

URL 3: models.doc2vec – Doc2vec paragraph embeddings. *Gensim*, https://radimrehurek.com/gensim/models/doc2vec.html [18.11.2021].

URL 4: models.word2vec – Word2vec embeddings. *Gensim*.

https://radimrehurek.com/gensim/models/word2vec.html#gensim.models.word2vec.Word2Vec [18.11.2021].

URL 5: WebVectors: word embeddings online. *Nordic Language Processing Laboratory.* http://vectors.nlpl.eu/explore/embeddings/en/ [18.11.2021] – für Englisch und Norwegisch.

URL 6: Word embedding demo. *Turku NLP group.* http://bionlp-www.utu.fi/wv_demo/ [18.11.2021].

URL 7: BUBENHOFER, Noah (2019): Word Embeddings: Funktionale Äquivalenz statt Synonymie. In: *Sprechtakel* (publiziert 02.03.2019). https://www.bubenhofer.com/sprechtakel/2019/03/02/word-embeddings-funktionale-aequivalenz-statt-synonymie/ [18.11.2021].

URL 8: *MhdBDB – Statistik*. http://mhdbdb.sbg.ac.at/LastStatistics.de.html [18.11.2021]; (generiert am 12.11.2021).

URL 9: *MhdBDB – Textsuche: tavelrundære.* http://mhdbdb.sbg.ac.at/mhdbdb/App?action=TextQueryModule&string=%40tavelrund%C3%A6re&filter=&texts=%21&startButton=Suche+starten&contextSelectListSize=1&contextUnit=1&verticalDetail=3&maxTableSize=100&horizontalDetail=3&nrTextLines=3 [18.11.2021].

URL 10: *MhdBDB – Textsuche: knabe.* http://mhdbdb.sbg.ac.at/mhdbdb/App?action=TextQueryModule&string=%40knabe&filter=&texts=%21&startButton=Suche+starten&contextSelectListSize=1&contextUnit=1&verticalDetail=3&maxTableSize=100&horizontalDetail=3&nrTextLines=9 [18.11.2021].

URL 11: *MhdBDB – Wortindex: bier.* http://mhdbdb.sbg.ac.at/mhdbdb/App?action=Dic&lid=702&mode=00 [18.11.2021].

URL 12: *MhdBDB – Wortindex: 21111307&21111400*. http://mhdbdb.sbg.ac.at/mhdbdb/App?action=DicSelect&LemmaSelectAction=Dic&mode=00&LemmaSelectPattern=21111307%2621111400 [18. 11. 2021].

URL 13: *MhdBDB – Text List.* http://mhdbdb.sbg.ac.at/mhdbdb/App?action=TextList [18.11.2021].