Die Generierung von Wortfeldern und ihre Nutzung als Findeheuristik. Ein Erfahrungsbericht zum Wortfeld „medizinisches Personal“

Adelmann, Benedikt
adelmann@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Franken, Lina
lina.franken@uni-hamburg.de
Universität Hamburg, Deutschland

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg, Deutschland

Krüger, Katharina
katharina.krueger@uni-hamburg.de
Universität Hamburg, Deutschland

Vauth, Michael
michael.vauth@tuhh.de
Technische Universität Hamburg, Deutschland

In vielen geistes- und sozialwissenschaftlichen Forschungsprojekten wird mit umfangreichen Textkorpora gearbeitet, die einerseits zu groß sind, um im Sinne eines *Close-Reading*-Ansatzes vollständig annotiert zu werden, in denen es andererseits aber einzelne Textpassagen von besonderer Forschungsrelevanz gibt, bei denen eine solche detaillierte Annotation wünschenswert ist. Das – nach Möglichkeit automatisierte – Auffinden solcher relevanten Textpassagen wird dadurch zu einem notwendigen Teilschritt des Arbeitsprozesses.

Eine simple, aber sehr effektive Möglichkeit der Suche ist die Nutzung von Wortfeldern. In der von Trier in den 1930ern entwickelten Wortfeldtheorie (Trier, 1973) wird das sprachliche Lexikon als – lexikalisch oder konzeptuell – strukturiert betrachtet.[1] Zwischen den Wörtern eines Wortfeldes bestehen Zusammenhänge, die es zu einer weitgehend stabilen semantischen Einheit machen, allerdings sind die Grenzen zwischen benachbarten Wortfeldern meist unscharf. Im digitalen Forschungskontext spielen Wortfelder vor#allem eine wichtige Rolle für die Taxonomieerstellung im Semantic Web und werden außerdem in der Lexikonforschung als Werkzeug genutzt (z.B. Bindi et al. 1994, Hamp und Feldweg 1997, Fellbaum 1998). Während diese Zugänge theoretisch fundiert sind, wird der Einsatz von Wortfeldern für die digitale Textanalyse bislang eher *ad hoc* genutzt, reflektierte Ansätze wie Heuser und Le-Khac (2011) sind die Ausnahme. Wir versuchen deshalb die Erstellung von Wortfeldern zu systematisieren und ihren Nutzen als Findeheuristik zu bewerten.

In diesem Beitrag berichten wir über die Erstellung von Wortfeldern unter Verwendung dreier Typen von Verfahren, die von der Erstellung aus bestehenden Ressourcen über die manuelle Generierung bis hin zu stark automatisierten Vorgehensweisen reichen, exemplarisch am Wortfeld „medizinisches Personal“. Dabei können Wortfelder u.#a. semantische Netze, standardisierte Vokabulare oder Konzepttaxonomien, unstrukturierte Wortlisten oder Kombinationen all dessen sein. Das Wortfeld bildet eine thematische Schnittstelle unserer Projekte im Forschungsverbund „Automatisierte Modellierung hermeneutischer Prozesse – Der Einsatz von Annotationen für sozial- und geisteswissenschaftliche Analysen im Gesundheitsbereich” (hermA, vgl. Gaidys et al., 2017): Das Verbundprojekt möchte anhand des Bereichs Gesundheit erarbeiten, wie die Automatisierung von Annotationen für hermeneutische Analyse- und Erkenntnisprozesse verbessert werden kann.

Verfahren zur Wortfeldgenerierung

Wortfelder aus bestehenden Ressourcen

Im Sinne der Wortfeldtheorie ist es naheliegend, bestehende lexikalische Ressourcen insbesondere strukturierter Daten zu verwenden, um daraus Wortfelder zu erstellen.

In dieser Hinsicht geeignet scheint der Online-Thesaurus *GermaNet* (Henrich und Hinrichs, 2010), der aktuell 164.814 lexikalische Einheiten umfasst, die in 128.100 sogenannte „Synsets“ semantisch strukturiert sind. Die hierarchisch-semantische Struktur des Wortnetzes ermöglicht es, schnell Subfelder zu identifizieren, die bei sehr spezifischem Erkenntnisinteresse nützlich sind. Wir haben für die Wortfelderstellung „medizinisches Personal“ alle 247 Hyponyme verwendet, die dem Begriff *Heilberufler* zugeordnet sind. Mit diesem Wortfeld wurden in einem Korpus aus 32 dystopischen Gegenwartsromanen 63 Begriffe (779 Erwähnungen) gefunden.

Eine Alternative sind kontrollierte Vokabulare, die mit Begriffen operieren, welche durch eine Redaktion definiert werden. Die Begriffe werden mit Metadaten, etwa Synonymen oder Übersetzungen, angereichert und hierarchisiert. Für medizinisches Personal wurde exemplarisch am größten deutschsprachigen Vokabular gearbeitet, der gemeinsamen Normdatei (GND[2]) der Deutschen Nationalbibliothek (DNB). Wir haben die

in der GND enthaltenen Begriffe zum Themenfeld „medizinisches Personal“ recherchiert, indem wir die hierarchische Struktur der Begriffe ausgehend vom allgemeinen Begriff „Arzt“ durchgegangen sind. In der Folge wurden dann alle relevanten Teilbäume aus der XML-Datei des gesamten Vokabulars extrahiert. Der Versuch, im Vokabular als verwandt markierte Begriffe einzubeziehen, war nicht zielführend, da sich inhaltlich disparate Felder ergaben. Schließlich konnte ein Wortfeld „medizinisches Personal“ mit 127 Begriffen aus der GND generiert werden. Die Suche wurde auf ein Korpus von 195 Bundestagsprotokollen angewendet und erbrachte 5.686 Fundstellen, mindestens ein Begriff des Wortfeldes war in fast jedem der vorher manuell als relevant recherchierten Protokolle vorhanden. Allein 4.508 Treffer entfallen dabei auf den Begriff „Arzt“ oder flektierte Formen davon, zahlreiche Begriffe tauchten im speziellen Korpus nicht auf.

Die Erstellung von Wortfeldern aus bestehenden Ressourcen ist verhältnismäßig unaufwändig. Allerdings muss ein besonderes Augenmerk auf eventuell fehlende, fehlerhafte oder unausgewogene Wörter bzw. Zusammenhänge gelegt werden, um diese Defizite nicht in die generierten Wortfelder zu übernehmen.

Manuell generierte Wortfelder

Eine manuelle *Ad-hoc*-Zusammenstellung von Wörtern als Wortfeld ist für literarische Texte meist nicht geeignet, etwa wenn diese historisches Vokabular enthalten. Um auch Textstellen finden zu können, die zeitgenössische Termini für medizinisches Personal nutzen, haben wir prototypische literarische Texte um 1900 mit dem Annotationstool *CATMA* (Meister et al., 2016) ausgezeichnet. In den sieben annotierten Romanen wurden insgesamt 21 zeitspezifische Bezeichnungen identifiziert, die dem Wortfeld „medizinisches Personal“ zugeordnet wurden.

Ein alternatives manuelles Verfahren ist die Auswertung historischer Lexika, um weitere historische Bezeichnungen zu ermitteln. In anderen Fällen eignen sich Sachregister, etwa aus wissenschaftlichen Publikationen.

Solche manuellen Verfahren bieten sich insbesondere für Texte an, die vom üblichen Sprachgebrauch abweichen. Das *Close Reading* der Texte und die manuelle Annotation relevanter Termini orientieren sich an herkömmlichen geisteswissenschaftlichen Arbeitsweisen. Gleichzeitig können sie Ausgangspunkt für (halb-)automatische Verfahren sein.

Automatische Verfahren

Ausgehend von bereits identifizierten Begriffen haben wir mit *Word Embeddings* gearbeitet. Verfahren der Wahl war *word2vec* (Mikolov et al., 2013) in der Implementation „gensim“ (#eh##ek & Sojka, 2010), das aus einer großen Menge an Sätzen unüberwacht Vektoren vorgegebener Dimension zu jedem vorkommenden Wort erzeugt, sodass in ihrer Bedeutung ähnliche Wörter möglichst ähnliche Vektoren erhalten und umgekehrt. Für die Vektordimension, die Größe des zu berücksichtigenden Kontextes und alle anderen Parameter des Verfahrens verwendeten wir die voreingestellten Standardwerte. Als Trainingsdaten dienten Volltexte von über 2.500 Erzähltexten um 1900, wobei die Aufteilung in Wörter und Sätze nach einer einfachen Heuristik erfolgte. Zur Erstellung von Wortfeldern bestimmten wir anschließend die Kosinus-Ähnlichkeit vorher bekannter Schlüsselwörter wie „Arzt“ oder „Doktor“ zu allen anderen Wörtern im word2vec-Modell. Wir sortierten die Wörter in absteigender Reihenfolge der so bestimmten Vektorähnlichkeit, verwarfen alle mit Ähnlichkeit unter 50#% und erhielten auf diese Weise zu jedem Schlüsselwort je eine Liste im Korpus potenziell semantisch ähnlich verwendeter anderer Wörter.

Die daraus hervorgehenden Listen wurden manuell hinsichtlich der Begriffe durchsucht, die tatsächlich medizinisches Personal bezeichnen. Die gefundenen 131 Begriffe (z.#B. Physikus, Wundarzt, Hebmutter) erweiterten das manuell erstellte Wortfeld umfassend. Ohne die halbautomatische Unterstützung wäre es kaum denkbar gewesen, die Vielzahl an Berufsbezeichnungen in den literarischen Texten zu ermitteln.

Zum Einsatz von Wortfeldern

Die vorgestellten Verfahren zur Generierung von Wortfeldern nutzen wir als Findeheuristiken für die Bearbeitung der drei Korpora, die wir für unsere differenten Forschungsfragen untersuchen.

Ausgehend von der Suche mit dem *GermaNet*-Wortfeld wurden Texte in dem Korpus dystopischer Romane mit dem Annotationstool *CATMA*[3] identifiziert, in denen medizinisches Personal besonders häufig genannt wird. In diesen Texten konnten nun wiederum unter Einbezug der Annotationen der Kapitelüberschriften Kapitel identifiziert werden, in denen diese Figuren präsent sind. Erste Stichproben haben gezeigt, dass so Textpassagen gefunden werden, in denen Figuren in medizinischer Hinsicht handelnd auftreten. In literaturwissenschaftlicher Hinsicht hat die wortfeldbasierte Suche damit das Potenzial, bestimmte Motive, Themen und Figurentypen auffindbar zu machen.

Mit dem aus dem GND-Vokabular erstellten Wortfeld haben wir ein Korpus von Bundestags- und Bundesratsprotokollen in *MaxQDA*[4] automatisch annotiert. Dieses Korpus war im Vorfeld auf Grundlage manueller Recherchen als thematisch relevant markiert worden, um Diskurse zu Akzeptanzproblematik von Telemedizin zu analysieren. In dieser Diskursarena spielt medizinisches Personal eine zentrale Rolle. Die Suchergebnisse strukturieren die weiteren Textanalysen,

denn sie zeigten relevante Textpassagen zur Rolle des medizinischen Personals sowohl im Diskurs über als auch im Realisieren der Telemedizin auf.

Die manuell zusammengestellten Wortfelder aus literarischen Texten um 1900 sowie einem historischen Lexikon haben wir ebenso wie die automatisch mit *word2vec* generierten und manuell bereinigten Listen auf unser Textkorpus von über 2.500 Prosatexten angewendet. Die Begriffe für „medizinisches Personal" aus den literarischen Texten wurden dabei 42.569-mal in 1.574 Dokumenten gefunden; die aus dem Lexikon ergaben 16.713 Treffer in 1.418 Dokumenten; die der automatisch generierten Listen 57.968 Treffer in 1.704 Dokumenten. Die identifizierten Textstellen können nun zielgerichtet im Zusammenhang mit der Fragestellung zu Krankheit und Geschlecht um 1900 analysiert werden. Unter anderem lässt sich die von Berufsbezeichnungen häufig implizierte Geschlechtszugehörigkeit betrachten (z.#B. „Krankenschwester", „Hebamme", „Sanitäter", „College" bzw. „Krankenpfleger" vs. „Krankenpflegerin").

Fazit: Wortfelder als Findeheuristik

Die vorgestellten Methoden der Wortfeldgenerierung können in manuelle, halbautomatisierte und automatisierte Generierungsstrategien unterteilt werden. Außerdem unterscheiden sie sich hinsichtlich der genutzten Ressourcen: Es gibt Verfahren, in denen die Wörter aus den zu analysierenden Texten – und damit direkt aus dem Forschungsobjekt – stammen, und solche, in denen von konkreten textuellen Kontexten bzw. Forschungsgegenständen unabhängige Wörter genutzt werden.

Durch eine Kombination der unterschiedlichen Methoden können textinhärente und textunabhängige Aspekte – folglich induktive und deduktive Ansätze – berücksichtigt werden, was bessere Ergebnisse ermöglicht. Um die Nutzung von Wortfeldern als Findeheuristiken weiter voranzutreiben, sollte deshalb die geeignete Kombination der Verfahren vor dem Hintergrund der Wortfeldtheorie weiter ausgearbeitet und evaluiert werden.

Fußnoten

1. Für eine Übersicht zur linguistischen Wortfeldtheorie vgl. Vassilyev (1974) und Lehrer (1974, S. 15-45).
2. http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html und http://gnd.eurospider.com/s , jeweils zuletzt abgerufen am 27.09.2018.
3. Je nach Größe des Wortfelds können mithilfe der Suchfunktion des Tools auch größere Korpora auf die Begriffe des Wortfelds durchsucht werden. Gefundene Begriffe können annotiert und somit beliebig viele Wortfeldsuchen miteinander kombiniert werden. Beim Annotationsvorgang ist es darüber hinaus möglich, die Ergebnisse der Wortfeldsuche manuell gegebenenfalls auch unter Berücksichtigung des Kontextes zu selektieren.
4. In MaxQDA können einzelne Begriffe gesucht werden.Eine programminterne Lemmatisierung findet statt, ist in ihren Regeln innerhalb des proprietären Programms aber nicht transparent. Für das Wortfeld mussten die 127 Begriffe sowie 88 zugehörige Synonyme der Begriffe eingetragen werden, um ein Wortfeld innerhalb der Programmoberfläche wiederum manuell zu erstellen. Offensichtlich falsche Synonyme wurden dabei aus Gründen der Suchgenauigkeit manuell entfernt, fehlende Synonyme wurden nicht hinzugefügt.

Bibliographie

Bindi, Remo, Calzolari, Nicoletta / Monachini, Monica / Pirelli, Vito / Zampolli, Antonio (1994): *Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition*. In: Literary and Linguistic Computing 9, S. 29–46.

Fellbaum, Christiane (Hg.) (1998): *WordNet: an electronic lexical database*. Cambridge, Mass: MIT Press.

Gaidys, Uta / Gius, Evelyn / Jarchow, Margarete / Koch, Gertraud / Menzel, Wolfgang / Orth, Dominik / Zinsmeister, Heike (2017): *Project Description. HermA: Automated Modelling of Hermeneutic Processes.* In: Hamburger Journal für Kulturanthropologie 7 (2017), S. 119–123.

Hamp, Birgit / Helmut Feldweg (1997): *GermaNet – a Lexical-Semantic Net for German*. In: Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, 1997.

Henrich, Verena / Hinrichs, Erhard (2010): *GernEdiT – The GermaNet Editing Tool*. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, S. 2228–2235.

Heuser, Ryan / Le-Khac, Long (2011): *Learning to Read Data: Bringing out the Humanistic in the Digital Humanities*. In: Victorian Studies 54, S. 79–86.

Lehrer, Adrienne (1974): *Semantic fields and lexical structure*. Amsterdam: North-Holland Publ. Co. [u.a.].

Meister, Jan Christoph / Petris, Marco / Gius, Evelyn / Jacke, Janina (2016): *CATMA 5.0 [Software für Textannotation und -analyse]*: http://www.catma.de (Zugriff: 24.09.2018).

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): *Efficient Estimation of Word Representations in Vector Space*. ArXiv: https://arxiv.org/abs/1301.3781 (Zugriff: 25.09.2018).

#eh##ek, Radim / Sojka, Petr (2010): *Software Framework for Topic Modelling with Large Corpora*. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, S. 45–50.

Trier, Jost (1973): *Über Wort- und Begriffsfelder*. Darmstadt: Wissenschaftliche Buchgesellschaft. [Zuerst in: Trier, Jost (1931): Der deutsche Wortschatz im Sinnbezirk des Verstandes. Heidelberg, S. 1 - 26 und 310 - 322])

Vassilyev, Leonid M. (1974): *The Theory of Semantic Fields: A Survey*. In: Linguistics 12, S. 79–94.