**University of
Zurich^{UZH}**

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2020

Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings

Bubenhofer, Noah

Abstract: The present study focuses on serially occurring narrations of ‘everyday’ life, more specifically on birthing as narrated by mothers on online forums; the underlying idea being that these narrations happen against the background of cultural narratives. The present paper uses word embedding models to detect typical topics and actors in these narrations. The calculation of word embeddings automatically constructs semantic spaces, where semantic relations (synonymy in particular) can be modeled. This method offers a way to think of synonymy as ‘functional equivalence in discourse’. The present study relies on previous work with n-grams (Bubenhofer, 2018). N-grams are sequences of words that often appear together; their sequential order in different narrations gives insight in narrative patterns. A further step in the analysis is the construction of ‘narrative topoi’, which is achieved through clustering neighboring vectors. The emerging clusters can in turn be grouped into five narrative elements of ‘telling a birth story’: 1) disruption of daily life, 2) personnel, 3) body, 4) fear, 5) joy. While it seems obvious that certain themes ‘belong’ into the narration of a delivery, it is less obvious with what vocabulary these themes are expressed. The presented method of clustering word-embedding-profiles adds tremendously to the modelling of a narrative. Its advantages lie in its potential to show lexical variation, as it also includes rare, non-conformative orthographical variants. Furthermore, it allows for a discourse-specific (and usage-based) view on semantic relations. The same applies to relations between semantic clusters. Seen from a discourse linguistics or cultural analysis perspective, word embeddings renew our understanding of semantics. This shows particularly fruitful if used to analyze (discourse dependent) derivations between semantic spaces.

DOI: <https://doi.org/10.1515/zgl-2020-2014>

Other titles: Semantic equivalence in birth stories: application of word embeddings

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-195511>

Journal Article

Published Version

Originally published at:

Bubenhofer, Noah (2020). Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings. *Zeitschrift für germanistische Linguistik*, 48(3):562-589.

DOI: <https://doi.org/10.1515/zgl-2020-2014>

Noah Bubenhofer*

Semantische Äquivalenz in Geburts- erzählungen: Anwendung von Word Embeddings

Semantic Equivalence in Birth Stories: Application of Word Embeddings

<https://doi.org/10.1515/zgl-2020-2014>

Abstract: The present study focuses on serially occurring narrations of ‘everyday’ life, more specifically on birthing as narrated by mothers on online forums; the underlying idea being that these narrations happen against the background of cultural narratives.

The present paper uses word embedding models to detect typical topics and actors in these narrations. The calculation of word embeddings automatically constructs semantic spaces, where semantic relations (synonymy in particular) can be modeled. This method offers a way to think of synonymy as ‘functional equivalence in discourse’.

The present study relies on previous work with n-grams (Bubenhofer, 2018). N-grams are sequences of words that often appear together; their sequential order in different narrations gives insight in narrative patterns. A further step in the analysis is the construction of ‘narrative topoi’, which is achieved through clustering neighboring vectors. The emerging clusters can in turn be grouped into five narrative elements of ‘telling a birth story’: 1) disruption of daily life, 2) personnel, 3) body, 4) fear, 5) joy. While it seems obvious that certain themes ‘belong’ into the narration of a delivery, it is less obvious with what vocabulary these themes are expressed.

The presented method of clustering word-embedding-profiles adds tremendously to the modelling of a narrative. Its advantages lie in its potential to show lexical variation, as it also includes rare, non-conformative orthographical variants. Furthermore, it allows for a discourse-specific (and usage-based) view on semantic relations. The same applies to relations between semantic clusters. Seen from a discourse linguistics or cultural analysis perspective, word embeddings renew our understanding of semantics. This shows particularly fruitful if used to analyze (discourse dependent) derivations between semantic spaces.

*Kontaktperson: Prof. Dr. Noah Bubenhofer: Universität Zürich, Deutsches Seminar, Schönberggasse 9, CH-8001 Zürich, E-Mail: noah.bubenhofer@ds.uzh.ch

- 1 Sprachgebrauchsmuster
- 2 Funktionale Äquivalente: Word Embeddings
- 3 Geburtsberichte
 - 3.1 Korpus
 - 3.2 n-Gramme als Ausdruck von Sprachgebrauchsmustern
 - 3.3 Narrative Funktionen
 - 3.3.1 Berechnung des Models
 - 3.3.2 Clustering
 - 3.4 Ergebnisse
 - 3.4.1 Einbruch in den Alltag
 - 3.4.2 Personal
 - 3.4.3 Angst
 - 3.4.4 Körper
 - 3.4.5 Freude
- 4 Fazit
 - Literatur

1 Sprachgebrauchsmuster

Erzählen folgt Mustern. Diese Muster sind Gegenstand vieler Untersuchungen, die etwa rhetorischen oder literaturwissenschaftlichen Interessen entsprungen sind. In der Linguistik entwickelten etwa Labov und Waletzky (1973) eine Erzähltheorie, die insbesondere Alltagserzählungen im Blick hat und in der Folge vielfach verfeinert und erweitert worden ist (etwa Quasthoff 1980; Gülich 1980; Boueke 1995; Knapp 1997; Labov 2013).

Erzählmuster erfahren in der Korpuslinguistik keine größere Aufmerksamkeit, ausgenommen von eher literaturwissenschaftlich orientierten Ansätzen des „Distant Readings“ (Moretti 2009) und den Digital Humanities generell. Im Fokus der vorliegenden Studie liegen jedoch sogenannte Alltagserzählungen (Gülich 1980; Quasthoff 1980; Ehlich 1980), die seriell auftreten, nämlich Geburtsberichte, in denen Mütter die Geburten ihrer Kinder erzählen. Diese erscheinen in einschlägigen Webforen zu Geburt und Elternschaft in meist spezifisch mit „Geburtsberichte“ benannten Threads. Obwohl jede Geburt ein individuelles, emotionales Erlebnis ist, fällt bei der Lektüre solcher Geburtsberichte auf, wie diese oft bestimmten Mustern folgen und aus Versatzstücken bestehen, die in vielen der Geschichten auftreten. So findet man in Geburtsberichten auf Webforen wie swissmomforum.ch oder eltern.de Formulierungen wie „das Licht der Welt erblickt“, „unglaubliches Gefühl“ oder aber bestimmte Themen, etwa der Einbruch der Geburtsanzeichen in den Alltag:

Dann kam Weihnachten. Wir starteten gemütlich in den Tag, am Nachmittag testete dir Grosse ihr neues Laufrad bei einem Spaziergang an der Sonne. Zuhause angekommen nahm ich ein Erkältungsbad. Meine Lieben kochten Znacht und ich bekam die ersten Wehen. Nicht stark oder sonderlich anders als sonst, aber irgendwie regelmässig, alle 10min. (<https://www.swissmomforum.ch/viewtopic.php?f=21&t=199622>)

Zwar lassen sich stilistische und sachliche Unterschiede finden, da die Geburten natürlich unterschiedlich ausfallen und unterschiedlich erlebt werden. Trotzdem scheint es bestimmte narrative Muster zu geben, an denen sich die Erzählerinnen orientieren.

Die Fragestellung der vorliegenden Studie liegt aber nicht darin herauszufinden, wie Geburten tatsächlich stattfinden, auch nicht, wie sie von den Müttern erlebt werden, sondern wie sie von ihnen *erzählt* werden. Leitend ist dabei die Idee, dass diese Erzählungen vor der Folie kultureller Narrative, wie Geburten erzählt werden, geschehen. Unter „kulturellem Narrativ“ werden Werkzeuge der Bedeutungsstiftung und Mittel des Verstehens und Erklärens verstanden (Bubenhofer/Müller/Scharloth 2013: 419; Bubenhofer 2018). Da die Geschichten in großer Zahl und unter gleichen medialen Bedingungen (Webforum) und Textsorte vorliegen, bieten sich datengeleitete, korpuslinguistische Zugänge an, um zu versuchen, ein Narrativ zu rekonstruieren.¹

Ich habe bereits gezeigt, wie aus einem Korpus von gut 14.000 solcher Geschichten datengeleitet typische Sprachgebrauchsmuster berechnet werden können (Bubenhofer 2018). Im Grunde handelt es sich dabei um Mehrworteinheiten (Sequenzen von Wortformen), die in den Geschichten häufig vorkommen. Um aber auch zu berücksichtigen, an welchen Stellen der Geschichten diese Muster auftreten, berechnete ich nicht nur die Muster, sondern auch ihre typischen Positionen in den Geschichten. So lassen sich verschiedene Reihungen von typischen Mustern und damit essenzielle Typen von Geburtsberichten rekonstruieren, so wie im folgenden Beispiel:

An diesem Tag hatte ich ... → ..., dass es endlich losgeht → ich hatte das Gefühl, dass ... → Mein Mann und ich waren ... → war mir sicher, dass ... → auf den Weg in die ... → Ich sagte

¹ Bislang sind Geburtsberichte aus Internetforen kaum Gegenstand der Forschung zur Narration, vor allem nicht linguistischer Forschung. Eine große Ausnahme ist die Dissertation von Cecilia Colloseus (2016), die aus kulturanthropologischer Sicht Narrative des Gebärens untersucht und dabei Geburtsberichte einer qualitativen Analyse unterzieht. Eine weitere Arbeit, die Geburtsberichte von Müttern untersucht, ist jene von Barbieri et al. (2012), die mit psychoanalytischem Interesse italienischsprachige Forumsbeiträge als Datengrundlage verwendet. Dabei betrachten sie auch linguistische Phänomene, wählen aber ebenfalls einen qualitativen Zugang. Weiter sei auf Tienken (2015) verwiesen, die Webforen zu „Sternenkindern“, also Babys, die vor, während oder kurz nach der Geburt verstorben sind, untersucht.

ihr, dass ... → so heftig, dass ich ... → fühlte sich an, als → war ich fix und fertig → Ich hatte das Gefühl, → dass es nicht mehr lange ... → ich dachte, ich muss ... → , aber es ging nicht → , was das Zeug hielt → dann ging alles ganz schnell → Ich weiß nur noch ... → um [Uhrzeit] war es → das Licht der Welt erblickte → ich konnte es nicht glauben → ich war so froh ... → , dass es vorbei war → ich hätte nie gedacht, → und ich muss sagen, → Für mich war es eine ... → noch vor sich haben ...

In Kombination mit weiteren Analysen zur Verteilung anderer Merkmale, etwa Wortarten oder Emotionsklassen, lassen sich sogenannte „narrative Topoi“ interpretativ herleiten und zu einem Modell eines „Geburtsberichte-Narrativs“ verdichten.

Die vorliegende Arbeit führt diese Analyse weiter, indem sogenannte „Word Embedding“-Modelle genutzt werden, um datengeleitet typische Themen und Akteure in den Geschichten zu finden. Die Berechnung von Word Embeddings hat sich in kurzer Zeit zu einer in der Computerlinguistik und dem Text Mining verbreiteten Methode entwickelt, um automatisch auf der Basis großer Textdaten semantische Räume zu konstruieren, in denen semantische Relationen, insbesondere Synonymie, modelliert werden können. In der kultur- und sozialwissenschaftlich orientierten Korpuslinguistik wird die Methode bislang kaum genutzt, birgt aber das Potenzial, Synonymie aus diskurslinguistischer Sicht erweitert als diskursfunktionale Äquivalenz zu denken. Die vorliegende Studie möchte von ersten Experimenten in diese Richtung anhand einer erweiterten Analyse der Geburtsberichte berichten und diese reflektieren.

2 Funktionale Äquivalente: Word Embeddings

Die Berechnung von Kollokationen ist in der Korpuslinguistik ein häufig angewandtes und erprobtes Verfahren, um gebrauchsemantisch typische Verwendungsweisen von Lexemen zu berechnen (Evert 2009). Ein Kollokationsprofil eines Lexems kann als semantisches Profil gelesen werden. Wenn zwei Lexeme ähnliche Kollokationsprofile aufweisen, dann handelt es sich um solche, die in den Daten ähnlich verwendet werden. Dies ist bei Synonymen der Fall.

Das Verfahren der Word Embeddings (Mikolov et al. 2013) systematisiert nun solche Kollokationsvergleiche. Das Assoziationsverhalten jeder Wortform oder jedes Lexems wird über die Daten generalisiert als Wortvektor repräsentiert. Der Wortvektor repräsentiert das Kovorkommen eines Wortes mit anderen Wörtern in den Daten als Zahlenreihe {0, 0, 122, 10, 204, 310, 2, ...}, wobei die Zahlen Häufigkeiten für ein Kovorkommen der beiden Wörter repräsentieren. Mittels unterschiedlicher statistischer Lernverfahren, etwa mit neuronalen Netzen, werden

über die gesamten Daten hinweg für jeden Type Vektoren berechnet, die das Kollokationsverhalten als distributionelles Modell repräsentieren (vgl. für eine Übersicht verschiedener Verfahren Lenci 2018). Mit den Vektoren kann die Position jedes Wortes in einem n -dimensionalen Vektorraum (wobei n der Anzahl Types entspricht, deren Kovorkommen bestimmt wird) repräsentiert werden. Zudem können vektorgeometrische Operationen verwendet werden: Ist der Winkel zwischen zwei Vektoren klein, verfügen die beiden Wörter über ähnliche Vektoren, werden also ähnlich verwendet und sind potenzielle Synonyme. Für jedes Wort im Vektorraum können die nächsten Nachbarn berechnet werden, so dass Gruppen von ähnlich verwendeten Wörtern datengeleitet berechnet werden können.

Word Embeddings werden erfolgreich für unterschiedliche Zwecke verwendet, bei denen es darum geht, semantische Ähnlichkeit auszunutzen. Normalerweise werden Word-Embedding-Modelle mit möglichst großen Datenmengen berechnet, um Referenzmodelle für eine Sprache zu erhalten.² Die Ausgabe von den nächsten zehn Nachbarn zu *weil* in einem umfangreichen Korpus deutschsprachiger Texte verschiedener Textsorten ergibt beispielsweise (in Klammern: Kosinus-Ähnlichkeit, je höher, desto ähnlicher):

weil: *da* (0.8665154576301575), *zwar* (0.8661791086196899), *denn* (0.8651647567749023), *wenn* (0.8639509677886963), *aber* (0.8609216213226318), *obwohl* (0.8489865064620972), *doch* (0.8444240093231201), *deshalb* (0.8283679485321045), *sondern* (0.8255864381790161), *selbst* (0.8252302408218384)

Die nächsten zehn Nachbarn von *Hebamme*, um eine zentrale Figur in den Geburtsberichten zu nennen, sind:

Hebamme: *Pflegefachmännern* (0.8411125540733337), *Arztgehilfinnen* (0.8256973624229431), *-therapeut* (0.8249690532684326), *Naturärzten* (0.8212056159973145), *Krankenwärter* (0.820869505405426), *Podologinnen* (0.8199600577354431), *-pfleger* (0.8197912573814392), *Pflegefachfrau* (0.8181213736534119), *Zahnärztin* (0.817986786365509), *Wartgeld* (0.8151979446411133)

Im Vergleich dazu ergeben die zehn nächsten Nachbarn zu *Arzt*:

Arzt: *Ärztin* (0.8921918272972107), *medizinisch* (0.8204370737075806), *Internist* (0.8186057806015015), *Schmerztherapeut* (0.8165487051010132), *Facharzt* (0.8064678907394409), *behandelnd* (0.8031118512153625), *Hausarzt* (0.8026583194732666), *Pneumologen* (0.8005311489105225), *Augenarzt* (0.7964136004447937), *Therapeut* (0.7958527207374573)

² Vgl. z. B. <https://devmount.github.io/GermanWordEmbeddings/> (01.04.2020).

Word-Embedding-Modelle können aber auch auf spezialisierten Korpora gerechnet werden. Ein Modell des Geburtsberichte-Korpus (Details dazu folgen weiter unten) zeigt dabei deutliche Differenzen zum Modell des allgemeinen Referenzkorpus von oben. Für das Beispiel *Hebamme*:

Hebamme: *Hebi* (0.9321193099021912), *Hebammenschülerin* (0.766809344291687), *Hebame* (0.7523912191390991), *Habamme* (0.7454160451889038), *Ärztin* (0.7398992776870728), *Hebis* (0.7237089276313782), *Nachthebamme* (0.7169431447982788), *Nachthebi* (0.6960358023643494), *hebi* (0.6955133080482483), *Claudine* (0.6929936408996582)

Im Vergleich mit dem allgemeinen Referenzkorpus sind hier die nächsten Nachbarn semantisch enger und quasisynonym. Darunter sind auch orthographisch abweichende Formen wie *Hebame* oder *Habamme* sowie dialektale Formen wie *Hebi*.

Je nach Datengrundlage zeigt sich Gebrauchsähnlichkeit unterschiedlich ausgeprägt. Diese Unterschiede scheinen aber aus computerlinguistischer Sicht nicht erwünscht zu sein, denn es existieren Testdatensätze, mit denen die Qualität eines Word-Embedding-Modells geprüft werden soll. So arbeiten Mikolov et al. (2013: 6) mit einem „Semantic-Syntactic Word Relationship test set“, mit dem z. B. Hauptstadt-Staat-Relationen, Paare von femininen und maskulinen Bezeichnungen wie Bruder / Schwester oder morphosyntaktische Beziehungen wie Tempusformen, Steigerungsformen, Derivationen etc. vorgegeben werden und überprüft wird, ob das Modell die richtigen Voraussagen macht. Wenn die Datengrundlage genug groß und bezüglich Themen heterogen und die Berechnung des Modells gut ist, können diese Testdatensätze zu großen Teilen „richtig“ beantwortet werden.

Mit diesem Fokus auf ein rigides Evaluationsverfahren geht jedoch die Sprengkraft, die der Methode eigentlich innewohnt, größtenteils verloren (vgl. Bubenhofer/Calleri/Dreesen 2019). Denn die Evaluation ist einem strukturalistischen Modell von Semantik verhaftet. Modernere semantische Theorien legen den Fokus jedoch auf den Sprachgebrauch und erlangten paradoxerweise gerade mit korpuslinguistischen Methoden neue Aufmerksamkeit und neue Möglichkeiten der empirischen Fundierung (Storjohann 2015: 259). Gerade für die Modellierung sinnrelationaler Strukturen sind gebrauchsssemantische Theorien von Bedeutung:

Heute wissen wir, wie sehr die empirische Wende in der Linguistik die Forschung der Sinnrelationen revolutionierte und das Interesse auf die dynamische Natur und das flexible Verhalten von Beziehungsgeflechten in Text und Diskurs lenkte. (Storjohann 2015: 259)

Die empirische Grundlage und korpuslinguistische Methoden wie die Kollokationsanalyse sind die Basis für eine andere Perspektive auf semantische Sinnrelationen. Die Berechnung von Word-Embedding-Modellen auf der Basis bestimmter Korpuszusammenstellungen, die spezifische Diskurse repräsentieren, lässt diese

Perspektive noch weiter denken – und zwar dann, wenn man sich von den Interessen, wie sie etwa im Text Mining bestehen, löst. Aus diskurslinguistischer Sicht sind Word-Embedding-Modelle gerade auch dann interessant, wenn sie bei Evaluationstests, die sich an einer strukturalistischen Semantik orientieren, wenig erfolgreich sind. Denn die Beispiele oben für „nächste Nachbarn“ von *Hebamme* und anderen Wörtern im Vergleich in den beiden Modellen macht deutlich, dass semantischen Räume diskursiv geprägt sind. Eine strukturalistische Auffassung von Synonymie ist das falsche Konzept, um dieses Phänomen und vor allem die Relationalität zu unterschiedlichen Datengrundlagen und damit Diskursen, fassen zu können. Gleichzeitig ergeben sich mit dem Einbezug solcher Word Embeddings interessante neue Analyseperspektiven, wie im Folgenden gezeigt werden soll.

3 Geburtsberichte

3.1 Korpus

Die Daten für die vorliegende Studie wurden im August 2016 von sechs Webforen aus Deutschland, Österreich und der Schweiz heruntergeladen (vgl. Bubenhofer 2018: 360–362).³ In Tabelle 1 sind die sechs Foren und jeweils die Anzahl der Texte und Wörter aufgeführt. Es wurden jeweils alle beim jeweiligen Zeitpunkt des Downloads verfügbaren Texte heruntergeladen, die im Unterforum, das meist den Titel „Geburtsberichte“ trägt, erschienen sind.

Tab. 1: Geburtsberichte-Korpus.

Land	Forum	# Wörter	# Wörter ohne Kommentare	# Texte
D	http://www.urbia.de	7 364 108	7 255 550	8 808
D	http://www.babyforum.de	2 089 936	2 057 168	1 824
A	http://www.parents.at	1 199 174	1 183 046	1 647
CH	https://www.swissmomforum.ch	1 156 193	1 138 660	919
D	http://www.eltern.de	438 017	431 416	716
D	http://www.umstandsforum.de	289 807	285 453	568
Total		12 537 235	12 351 293	14 482

³ Die Korpusdaten können aus rechtlichen Gründen nicht öffentlich zur Verfügung gestellt werden. Derivate der Daten (n-Gram-Listen, Modelle etc.) sind aber unter <https://www.bubenhofer.com/publikationen/geburtsberichte/> verfügbar.

Die Texte aus den deutschen Foren sind mit knapp 12 000 Texten (gut 10 Mio. Wörter) im Vergleich zu den österreichischen (1600 Texte, 1,2 Mio. Wörter) und den schweizerischen (900 Texte, 1,2 Mio. Wörter) deutlich in der Überzahl. Als ‚Text‘ ist dabei jeweils ein Thread zu verstehen, also das Initialposting mit der Geburtsgeschichte inklusive möglicher Antworten bzw. Kommentare.⁴ Die Antworten sind jeweils sehr kurz. Für die weiteren Analysen wurden die Antworten entfernt und nur mit den Initialpostings weitergearbeitet, da dort normalerweise die eigentliche Geschichte erzählt wird. In wenigen Fällen mussten die Erzählerinnen die Geschichten auf mehrere Postings aufteilen, wenn sie die erlaubte Textlänge pro Posting überschritten. Solche Fälle wurden nicht automatisiert entfernt, machen jedoch nur einen sehr kleinen Teil der Geschichten aus. Nach Entfernen der Antworten schrumpfte das Gesamtkorpus auf 12 351 293 Wörter, also um etwa 2,5 %.

Abb. 1: Zeitliche Verteilung der Korpuszusammenetzung.

⁴ Die Reaktionen auf die Initialpostings haben manchmal den Charakter von Antworten, manchmal eher von Kommentaren. Technisch gesehen sind sie an das Initialposting angehängte Texte von anderen Nutzerinnen. Ich spreche im Folgenden vereinfachend von „Antworten“, ohne den genauen Charakter weiter zu spezifizieren.

Abbildung 1 zeigt die zeitliche Verteilung der Texte gegliedert nach Foren. Die meisten Geburtsberichte stammen aus den Jahren 2006 bis 2013. Während die tiefen Werte der Jahre 2003 bis 2005 damit zu erklären sind, dass nicht alle Foren bereits dann existierten bzw. die Texte nicht greifbar sind, ist die kontinuierliche Abnahme bis 2016 unklar. Sie geht zwar hauptsächlich auf das Forum *urbia.de* zurück, das den Hauptteil des Korpus ausmacht, der Trend ist jedoch auch bei den anderen Foren sichtbar. Anscheinend werden die analysierten Foren immer weniger für Geburtsberichte verwendet. Ob stattdessen andere Foren oder andere Kommunikationsmittel verwendet werden, müsste noch untersucht werden.

Die im HTML-Format heruntergeladenen Daten wurden korpuslinguistisch aufbereitet, indem die verfügbaren Metadaten (wie Autorschaft und Datum) und der eigentliche Text extrahiert, Wortartklassen maschinell annotiert (unter Verwendung des *TreeTaggers*, Schmid 1994) und in die Korpusdatenbank CWB überführt worden ist (vgl. Bubenhofer 2018: 362).

3.2 n-Gramme als Ausdruck von Sprachgebrauchsmustern

Kern der in Bubenhofer (2018) präsentierten Analysen sind sogenannte Sprachgebrauchsmuster und ihre typischen Positionen in den Geschichten. Diese Analysen sollen nicht Thema der vorliegenden Studie sein, doch hier kurz referiert werden, um die Vorarbeiten erkenntlich zu machen.⁵

Um Sprachgebrauchsmuster in Korpora zu identifizieren, eignen sich sog. Wort-*n*-Gramme, also Mehrworteinheiten oder Sequenzen von Wörtern, die häufig zusammen auftreten (Bubenhofer 2015). Für die Geburtsberichte wurden sogenannte typische „komplexe *n*-Gramme“ berechnet (Bubenhofer/Scharloth 2011, 2013, 2015):

- Sequenzen von unmittelbar nacheinander stehenden Wörtern, *n*-Gramme, wobei die Sequenz aus einer Anzahl *n* Wörtern besteht. Im vorliegenden Fall $n \geq 5$.
- Die *n*-Gramme können aus Wortformen, lemmatisierten Formen (Grundformen) oder Wortartklassen bestehen (und einer Mischung davon). Im vorliegenden Fall bestehen die berechneten *n*-Gramme auf den Grundformen, wobei z. B. Zahlenwerte (Datums-, Zeit-, Gewichtsangaben etc.) mit einem Platzhalter *CARD* ersetzt werden.
- Die *n*-Gramme müssen statistisch auffällig für das Untersuchungskorpus der Geburtsberichte sein. Um aus der riesigen Menge aller *n*-Gramme im Korpus

⁵ Die folgenden Ausführungen sind teilweise direkt Bubenhofer (2018: 363–382) entnommen, jedoch stark gekürzt.

die statistisch signifikanten zu berechnen, wird ein Referenzkorpus eingesetzt, das aus allen im Zeitraum von 2010 bis 2016 erschienenen Zeitungsartikeln in den gedruckten Publikationen *Die Zeit* und *Der Spiegel* besteht. Auch im Referenzkorpus wurden alle n-Gramme berechnet und danach die Frequenzen jedes n-Gramms in den beiden Korpora mit einem statistischen Signifikanztest (Log-Likelihood Ratio) verglichen (vgl. Bubenhofer 2015 für Details). So ergibt sich eine Liste von 6941 n-Grammen, die statistisch signifikant für das Geburtsberichte-Korpus sind.⁶

Das häufigste unter den signifikanten n-Grammen dieser Liste ist: *an CARD um CARD Uhr*. Es handelt sich also um ein lemmatisiertes n-Gramm, das beispielsweise folgendermaßen realisiert wird:

- am 29.5.2010 um 02.23 Uhr
- am 24.09.09 um 0:46 Uhr
- am 31.07.05 um 5:40 Uhr
- am 01.07.2014 um 22.03 Uhr
- am 13.08.2014 um 18.09 Uhr

Weitere Beispiele sind *das Licht der Welt erblicken; auf den Weg machen; ich habe das Gefühl* (jeweils lemmatisierte Form) etc. Die Liste zeigt bereits, welche Sprachgebrauchsmuster typisch für die Geburtsberichte sind, aber um narrative Muster zu finden, ist deren sequenzielle Anordnung in den Texten relevant. Um dies zu erreichen wurde für jedes n-Gramm die typische relative Position in den Geschichten berechnet. Eine Position nahe 0 bedeutet am Anfang der Geschichte, nahe 1 am Ende, 0,5 in der Mitte. Da die Positionen eines n-Gramms in den Geschichten unterschiedlich sind, wurde neben dem Mittelwert auch die Standardabweichung davon berechnet, um zu sehen, ob das n-Gramm relativ stabil immer an der gleichen Position vorkommt oder stark streut.

Nach diesen Berechnungen zeigt sich dann beispielsweise, dass *Hallo ihr Lieben, ich* immer genau am Anfang der Geschichte vorkommt (mittlere Position: 0, Standardabweichung s: 0), *der eigentliche ET⁷ war der* ebenfalls eher am Anfang (mittlere Position: 0,06, Standardabweichung s: 0,1).

Abbildung 2 zeigt nun für die häufigsten 1609 n-Gramme die Korrelation von durchschnittlicher Position und Standardabweichung. Auf der x-Achse ist die Position, auf der y-Achse die Standardabweichung abgetragen. Jeder Punkt symbolisiert ein n-Gramm; je höher auf der y-Achse dieser liegt, desto unklarer

⁶ Die Liste kann unter www.bubenhofer.com/publikationen/geburtsberichte/ (01.04.2020) abgerufen werden.

⁷ ET steht für ‚errechneter Termin‘.

ist dessen Position in den Geschichten (höhere Standardabweichung). Eine Reihe von bezüglich Position besonders stabilen und instabilen n-Grammen ist beschriftet. Auf der begleitenden Website zur Publikation ist eine interaktive Version der Darstellung einsehbar, so dass alle n-Gramme angesehen werden können.⁸

Abb. 2: Geburtsgeschichten: Typische Positionen der n-Gramme.

Es zeigt sich nun ein generelles Bild, wonach besonders am Anfang und Ende der Geschichten die n-Gramme eher stabile Positionen einnehmen, in der Mitte jedoch nicht. Das ist plausibel, da der Anfang und das Ende von Geschichten

⁸ Zudem sind weitere Visualisierungen verfügbar, die eigene Analysen erlauben: www.bubenhofer.com/publikationen/geburtsberichte/ (01.04.2020).

oft formalisierter sind als der Rest. Interessant ist jedoch, dass n-Gramme, die den Akt des Nabelschnur-Durchschneidens (*Mein Mann schnitt die Nabelschnur; durfte die Nabelschnur durchschneiden* etc.) beschreiben, deutlich (und stabil) am Ende der Geschichten vorkommen. Das Skript einer Geburt ist ja aber mit dem Nabelschnur-Durchschneiden noch längst nicht zu Ende; die Nachgeburt beispielsweise könnte ebenfalls wiederkehrendes Thema sein. Das ist in vielen Geschichten jedoch nicht der Fall.

In der weiteren Analyse ergeben sich aufgrund der n-Gramme eine Reihe von „narrativen Topoi“, die ich im Anschluss an den diskurslinguistischen Topos-Begriff anbinde (Kienpointner 2017; Wengeler 2003), allerdings in erweiterter Form im Sinne eines Denkmusters oder Stereotyps (Kienpointner 2017: 187). Im Kontext von Erzählungen meint „narrativ“ dann aber ein diskursiv entstandener und damit stereotyp wirkender „Erzählort“. Beispiele für solche narrativen Topoi sind etwa in der Orientierungsphase der Erzählungen „Ansprache, Motivierung“ und „temporale Situierung“. Sodann in der Phase der Komplikation „Selbstbeobachtung und Selbstempfindung“ oder „Ungeduld“. Gegen Ende ist dann oft ein narrativer Topos „Geschwindigkeit“ vorhanden, der die Resolution einleitet und durch Sprachgebrauchsmuster wie *ging alles sehr schnell, dann ging alles sehr schnell* u. ä. konstituiert wird.

Ergebnis der Analysen ist ein Modell eines Geburtsberichte-Narrativs, das die narrativen Topoi den Phasen von Labov und Waletzky zuordnet und weitere Beobachtungen zu sprachlichen Auffälligkeiten nennt (Bubenhofer 2018: 382).

3.3 Narrative Funktionen

Im zweiten Teil der Analyse soll ein Word-Embedding-Modell genutzt werden, um den Wortschatz aus einer Perspektive funktionaler Äquivalenz betrachten zu können. Im Detail geht es um die folgenden zwei Ziele:

1. Das Word-Embedding-Modell ist eine Zusammenfassung eines semantischen Raumes, der weiteren interpretativen Schritten unterzogen werden kann. Hypothesengeleitet können bestimmte zentrale Wörter darin situiert und deren „nächste Nachbarn“, also gebrauchssemantisch gesehen ähnlich verwendete Wörter gefunden werden.
2. In einem zweiten Schritt wird das Modell verwendet, um Cluster von ähnlich verwendeten Wörtern zu identifizieren und somit für die Textsorte und das dahinterliegende narrative Muster zentrale Themen, Akteure und Topoi zu finden. Die Grundidee liegt dabei darin, mit einem strukturentdeckenden Clustering-Verfahren Gruppen von nahe beieinanderliegenden Punkten im Vektorraum zu finden und diese automatisch mit einem Label zu versehen.

3.3.1 Berechnung des Modells

Für die Berechnung des Modells wurden alle Grundformen (Lemma) der Wortformen im Korpus verwendet. Es gibt verschiedene Verfahren zur Berechnung von Word Embeddings. Im vorliegenden Fall verwendeten wir die von Mikolov et al. (2013) vorgeschlagene Methode, den sogenannten „Word2Vec“-Algorithmus. Dabei sind zwei grundsätzlich verschiedene Architekturen möglich, das „Continuous Bag-of-Words“- und das „Continuous Skip-gram“-Modell, mit „CBOW“ und „Skip-gram“ bezeichnet. Die beiden Architekturen unterscheiden sich darin, welche Wörter während des Lernprozesses berücksichtigt werden: In der CBOW-Architektur wird ein größerer Kontext ohne Berücksichtigung der Sequenz verwendet, um die Assoziationen zu modellieren. In der Skip-gram-Architektur wird ein Kontext von C Wörtern (in unserem Fall: 10) verwendet.

Mikolov et al. (2013: 7) zeigten, dass die Skip-gram-Modellierung insbesondere für semantische Klassifikationen besser funktioniert als für syntaktische Aufgaben. Daher entschieden wir uns bei der vorliegenden Aufgabe für die Skip-gram-Modellierung.

Zur Berechnung des Modells wurde die Implementierung im gensim-Paket (Řehůřek/Sojka 2010) verwendet.⁹ Mit der Script-Sprache python können die Modelle unkompliziert berechnet werden. Als Parameter wurde wie erwähnt die Skip-gram-Modellierung ausgewählt, ein Kontext von C=10 („window“) und eine Mindestfrequenz des Lemmas, damit es überhaupt einbezogen wird, von 10 gewählt. Das Modell ist auf der begleitenden Website verfügbar.¹⁰

Die folgenden Beispiele geben einen Eindruck zu Ausgaben des Modells:

Wehe: anrollen (0.7714179754257202), Hammerwehe (0.7110541462898254), wehe. (0.6944172978401184), Wehe. (0.6870291829109192), Wehenpause (0.6566441059112549), Nächste (0.6404309868812561), Welle (0.6262998580932617), Preßwehe (0.6261609792709351), Ausatmen (0.617641270160675), weglegen (0.6134337186813354)

Schmerz: Wehenschmerz (0.8821748495101929), schmerzen (0.746356725692749), Schmerzen. (0.7165833115577698), Dauerschmerz (0.6974124908447266), Bewegungsdrang (0.6873125433921814), wehenschmerz (0.6761729717254639), schmerzen. (0.6695544123649597), Geburtsschmerz (0.6639039516448975), Wehenspitzen (0.6634974479675293), Rückenwehen (0.6496166586875916)

Kind: Baby (0.8563555479049683), Tochter (0.7126445770263672), Sohn (0.6942390203475952), Kleine (0.6648626327514648), Zwilling (0.6449035406112671), Kindlein (0.6330938935279846), Zwerg (0.6289287805557251), Maus (0.6128379106521606), Butzi (0.6042927503585815), Babies (0.60008704662323)

⁹ Vgl. <https://radimrehurek.com/gensim/models/word2vec.html> (01.04.2020).

¹⁰ Vgl. <http://www.bubenhofer.com/publikationen/geburtsberichte/> (01.04.2020).

weil: *deswegen* (0.7363709211349487), *sodass* (0.6999390125274658), *da* (0.6969295740127563), *außerdem* (0.6948705911636353), *deshalb* (0.6797378063201904), *zumal* (0.6710671782493591), *obwohl* (0.6692008376121521), *nämlich* (0.6597287654876709), *aber* (0.6589906811714172), *daher* (0.6349456310272217)

extrem: *immens* (0.7714534997940063), *enorm* (0.7656272053718567), *ziemlich* (0.7187795639038086), *wahnsinnig* (0.707932710647583), *derb* (0.7021247744560242), *sehr* (0.6977962255477905), *Rippenschmerzen* (0.696819543838501), *paaren* (0.694392204284668), *mörderisch* (0.6901798248291016), *massiv* (0.6896220445632935)

3.3.2 Clustering

Um das Modell besser handhabbar zu machen, wird nun ein Clustering der Vektoren durchgeführt. Wie die Beispiele im vorhergehenden Abschnitt zeigen, kann jedes Wort im Vektorraum zusammen mit seinen nächsten Nachbarn als Cluster aufgefasst werden. Es handelt sich dabei um Ausdrücke, die in den Daten ähnlich verwendet werden, also in einer semantischen Relation, nämlich in semantischer Ähnlichkeit, zueinander stehen.

Lenci (2018: 161) beschreibt die Berechnung semantischer Ähnlichkeit als die Erfolgsgeschichte der distributionellen Modelle, mahnt aber gleichzeitig an, dass selten zwischen semantischer Ähnlichkeit und semantischer Abhängigkeit unterschieden würde. Ersteres meint eine mehr oder weniger synonymische Beziehung, letzteres eine Reihe von unterschiedlichen semantischen Relationen wie Hyper- oder Hyponyme, Antonymie, Meronymie etc. Die oben genannten Beispiele von Ausgaben des Word-Embedding-Modells zeigen dies: Die nächsten Nachbarn von *Kind* wie *Baby* oder *Kindlein* gehören eher in die erste Kategorie der semantischen Ähnlichkeit, während *Ausatmen* zu *Wehe* eher in einer semantischen, hier: thematischen, Relation zueinander stehen.

Bei der Entwicklung der Berechnungsmodelle für distributionelle Semantik spielt eine genauere Diskriminierung der semantischen Beziehungen eine immer stärkere Rolle (ebd.: 162; vgl. auch Boleda/Herbelot 2016). Für unsere Nutzung spielen diese Unterschiede jedoch eine untergeordnete Rolle. Zunächst ist es das Ziel, semantische Teilräume zu identifizieren, die in den Geburtsberichten eine wichtige Rolle einnehmen. Dafür müssen im Modell sinnvolle Cluster Grenzen gebildet werden. Darauf zielt ein Clustering-Algorithmus ab, wobei es wiederum zahlreiche Verfahren dafür gibt.

Ein bekanntes Verfahren ist der K-means-Algorithmus, der versucht optimale Clusterzentren zu bilden, die in minimal möglicher Distanz (quadrierte Distanz) zu den zum Cluster gehörenden Punkten stehen (Arthur/Vassilvitskii 2006). Vorgegeben wird die gewünschte Anzahl von Clustern (n).

Das Word-Embedding-Modell umfasst insgesamt 16.218 Types. Die Clustergrößen sind unterschiedlich, aber leitend bei der Bestimmung der Anzahl gewünschter Cluster ist das Verhältnis zwischen Anzahl Types und angestrebter durchschnittlicher Granularität der Cluster: Bei gut 16.000 Types und 2000 Clustern sind also durchschnittlich acht Wörter pro Cluster zu erwarten. Die Beispiele oben, bei denen jeweils die zehn nächsten Wörter angezeigt werden, zeigen, dass dies oft eine sinnvolle Anzahl ist. Wir experimentierten jedoch mit unterschiedlichen n : 500, 1000, 2000 und 5000.

Der Clustering-Algorithmus weist nach erfolgter Berechnung jedem Wort eine Cluster-ID zu. Um eine sprechende Beschreibung des Clusters zu erhalten, berechneten wir für jeden Cluster den sogenannten Zentroiden, also das Wort im Modell, das am nächsten der geometrischen Mitte des Clusters steht. Um dieses Label noch sprechender zu machen, wählten wir die drei nächsten Nachbarn zum Zentroiden, so dass sich folgende Liste ergibt (im Ausschnitt):

Tab. 2: Ausschnitt aus den geclusterten Word Embeddings (Version 2000 Cluster).

Cluster 1	Label 1	Distanz zu Zentroid	Label 2	Distanz zu Zentroid	Label 3	Distanz zu Zentroid
	<i>Oberarzt</i>	0.96	<i>Ärztin</i>	0.95	<i>Oberärztin</i>	0.94
Wörter	<i>Aerztin</i>		<i>Chefärztin</i>		<i>Oberärztin</i>	
	<i>Arzt</i>		<i>Doc</i>		<i>OÄ</i>	
	<i>Assistenzarzt</i>		<i>OA</i>		<i>Ärtzin</i>	
	<i>Assistenzärztin</i>		<i>Oberarzt</i>		<i>Ärztin</i>	
	<i>Chefarzt</i>					
Cluster 2	Label 1	Distanz zu Zentroid	Label 2	Distanz zu Zentroid	Label 3	Distanz zu Zentroid
	<i>Kreuzstich</i>	0.84	<i>Notfall</i>	0.83	<i>Anesthesist</i>	0.78
Wörter	<i>Kreuzstich</i>		<i>Notfall</i>			
Cluster 3	Label 1	Distanz zu Zentroid	Label 2	Distanz zu Zentroid	Label 3	Distanz zu Zentroid
	<i>nuckeln</i>	1.00	<i>genuckelt</i>	0.87	<i>schmatzen</i>	0.82
Wörter	<i>nuckeln</i>					
Cluster 4	Label 1	Distanz zu Zentroid	Label 2	Distanz zu Zentroid	Label 3	Distanz zu Zentroid
	<i>Eisbeutel</i>	0.94	<i>Akkupunktur-nadeln</i>	0.88	<i>einflößen</i>	0.88

Tab. 2: (fortgesetzt)

Wörter	Akkupunkturna- deln		Röhrchen		Wärmekissen	
	Akupunkturnadel		Sandsack		Zunge	
	Band		Sauerstoff- schlauch		venösen	
	Eisbeutel					
Cluster 5	Label 1	Distanz zu Zentroid	Label 2	Distanz zu Zentroid	Label 3	Distanz zu Zentroid
	wegdösen	0.91	weg- gedämmert	0.91	Sekunden- schlaf	0.88
Wörter	Durchatmen		hindurch		wegdösen	
	Dösen		immerwieder		weggedämmert	
	Einschlafen		manchmal		weggedöst	
	Hyperventilieren		rauben		weggenickt	
	Ruhepause		sogar		zeitweise	
	Sekundenschlaf		totmüde		zwischen	

Es ist also ersichtlich, dass sich ein Cluster (1) ergeben hat, dem verschiedene Bezeichnungen für medizinisches Personal (*Arzt, Ärztin, Oberärztin* etc.) angehören und dessen Label sich aus den Wörtern *Oberarzt, Ärztin* und *Oberärztin* ergibt, die dem Zentroiden am nächsten stehen. Bemerkenswerterweise stehen auch die Akronyme und Abkürzungen *OA, Doc* und *OÄ* im Cluster, was den Vorteil eines datengeleiteten Verfahrens ersichtlich macht, da solche Bezeichnungen z.B. in manuell erstellten Ontologien, Thesauri o. ä. nicht genannt würden.

Weiter zeigt sich, dass die Cluster sehr unterschiedlich groß sind; so umfasst der Cluster 3 nur gerade ein Wort, *nuckeln*. Als Label wurden noch *genuckelt* und *schmatzen* berücksichtigt, da immer die drei dem Zentroid nächsten Nachbarn ausgewählt werden. In diesem Fall hat sich das Clustering als zu fein erwiesen. Beim Clustering mit $n = 1000$ erscheint *nuckeln* nämlich im folgenden Cluster: *Brustwarze, Busen, andocken, anlegen, brust, docken, genuckelt, gesaugt, gierig, nuckeln, satt, saugen, schmatzen*.

Die Cluster umfassen Wörter unterschiedlicher Wortarten, was Cluster 5 (*wegdösen, weggedämmert, Sekundenschlaf*) zeigt. Aus enger semantischer Sicht mag dies unbefriedigend sein. Aus einer gebrauchssemantisch und diskurslinguistisch interessierten Perspektive sind diese Äquivalente jedoch interessant.

3.4 Ergebnisse

Sind die Wortformen im Korpus mit den Clusterzugehörigkeiten annotiert, können anschließend Dichteverteilungen bestimmter Cluster berechnet werden. Es wird also die Häufigkeitsverteilung über den Verlauf der Geschichten berechnet, indem für jedes Auftreten einer Wortform eines bestimmten Clusters die relative Position zwischen 0 (Anfang der Geschichte) und 1 (Ende) bestimmt wird. So ergeben sich Histogramme wie in 3.

Im Folgenden werden 43 Cluster präsentiert, die in der qualitativen Analyse wiederum zu fünf zentralen Elementen der Geburtsberichte zusammengefasst sind: a) „Einbruch in den Alltag“, b) „Personal“, c) „Körper“, d) „Angst“ und e) „Freude“. Es gibt weitere zentrale Cluster, die aus Platzgründen nicht besprochen werden können. Die 43 ausgewählten Cluster sollen exemplarisch die Nützlichkeit der Methode für die Modellierung eines Geburtsberichte-Narrativs zeigen.

3.4.1 Einbruch in den Alltag

Abbildung 3 zeigt fünf verschiedene Cluster, die mit den drei Wörtern etikettiert sind, die den jeweiligen Zentroiden am nächsten stehen. Der Cluster *Kakao / Salat / Müsli* umfasst insgesamt die folgenden Wörter: *Banane, Cola, Getränk,*

Abb. 3: Dichteverteilung Geburtsberichte: Cluster „Alltag“.

Ingwertee, Joghurt, Kaffee, Kaffee, Kakao, Kiosk, Käse, Milchreis, Nudel, Obst, Pommes, Salat, Schokolade, Stärkung, Suppe, Teller, Traubenzucker, Zeitschrift, kochen, leck, lecker, schlürfen. Die Verteilung über den Verlauf der Geschichten zeigt nun deutlich, dass Ausdrücke dieses Clusters hauptsächlich im ersten Drittel der Geschichten benutzt werden. Bei den anderen abgebildeten Clustern ist das ähnlich, wobei die Cluster *Müsli / Currywurst / Käsebrot* und *Rotwein / Gläschen / Wein* auch gegen Ende der Geschichten wieder etwas zunehmen.

Diese Cluster und ihre Dichteverteilung lassen vermuten, dass die Erzählungen oft damit beginnen, den Einbruch der eigentlichen Geschichte, also der Geburt, in den Alltag zu beschreiben. Im Anschluss an die Struktur von Labov und Waletzky (1973: 112) könnte dies als Übergang von der Orientierungs- zur Komplikationsphase beschrieben werden.

3.4.2 Personal

Abbildung 4 zeigt eine Reihe von Clustern, die verschiedene Personen- und Rollenbezeichnungen umfassen. Dazu gehören die Hebamme, Ärztinnen und Ärzte, Patientinnen sowie der Partner oder Freunde. Auffallend ist zunächst, dass das Word-Embedding-Modell ziemlich vollständig die verschiedenen medizinischen Funktionen auseinanderhält: *Arzt/Ärztin, Frauenarzt/Frauenärztin, Anästhesist/in, Kinderarzt/-ärztin* und Hebammen. Die Hebammen sind in den Geschichten zentrale Figuren und auffallend ist im Modell, dass der Ausdruck *Hebamme* folgendem Cluster zugeordnet ist:

Artzt, C., Charo, Christl, Claudine, Dagmar, Dok, Dorli, Doro, Dorothea, Ellen, Erika, Eva-Maria, Fachärztin, Friederike, Gaby, Gundula, HEbi, Habamme, Hebamen, Hebamme, Hebamme., Hebbi, Hebi, Hebischülerin, Helferin, I., Ingrid, Ivana, KIA, Lernschwester, Lidia, Ludmilla, Michela, Nachthebi, Nachtschwester, Nasrin, Oberarzt, Pressversuch, Renat, Riona, Sie|sie, Sigrid, Stationsschwester, Susanna, Ulla, Uta, Ute, Vera, Wehenveratmen, Y., Z., bimmeln, diensthabene, dies., habamme, herrein, kuckte, nochmal., reingerannt, reingestürmt, reinschaute, schauend, sie, u., vorsichtig, Ärztin.

Interessant ist, dass besonders viele Vornamen ähnlich verwendet werden wie die Bezeichnungen *Hebamme, Hebi* (und normabweichende Varianten), also normalerweise die Namen der Hebammen darstellen. Bezeichnungen für *Arzt*-personen sind alles normabweichende und deshalb niedrigfrequente Ausdrücke, bei denen die Zuordnung zum Cluster fragwürdig ist.

Die Benennung der Hebamme mit dem Vornamen zeigt das offensichtlich enge Verhältnis zwischen Gebärender und Hebamme, das auch in der Erzählung so dargestellt wird. Bei den anderen Funktionsbezeichnungen des medizinischen Personals werden keine Namen genannt.

Abb. 4: Dichteverteilung Geburtsberichte: Cluster „Personen“.

Die Hebamme selber unterscheidet sich in den Erzählungen wiederum von den Ausdrücken *Hebammenschülerin*, *Geburtshelferin* und *Beleghebamme*, mit denen wiederum deutlich wird, dass die so benannten Personen in einem anderen Verhältnis zur Erzählerin stehen. Zusammen mit den Informationen zur Dichteverteilung der Benennungen über den Verlauf der Geschichten offenbart sich das Hierarchiegefüge des medizinischen Personals und der weiteren Personen während einer Geburt: Beim Cluster *Claudine / HEbi / Erika* (der für die persönlich betreuende Hebamme steht) ist eine relativ gleichmäßige, wenn auch stetige, leichte Steigung der Nennungshäufigkeiten ersichtlich. Der/die Frauenarzt/-ärztin spielt nur am Anfang der Geschichten eine Rolle, der/die Anästhesist/in ziemlich genau in der Mitte, ähnlich wie die Bezeichnungen *Oberarzt/-ärztin* (allerdings mit weniger ausgeprägter Steigerung) und gegen Ende spielt nur noch der/die *Kinderarzt/-ärztin* eine bedeutende Rolle, in Kombination mit dem Verb *wegbringen*, was auf Komplikationen beim Neugeborenen verweisen

dürfte, weswegen das Baby in kinderärztliche Obhut gebracht werden muss. Auch die Bezeichnung *Beleghebamme* und *Geburtshelferin*, die am Anfang der Geschichten häufiger genannt werden, verweist auf den Funktionswandel einer Hebamme im Verlauf der Geschichte: Am Anfang der Geburt ist sie die Gatekeeperin, die darüber entscheidet, ob die werdende Mutter überhaupt schon bereit für die Geburt ist – semantisch-funktionales Äquivalent zur Bezeichnung *Beleghebamme* sind deshalb nicht wie bei der behandelnden Hebamme Vornamen, was auf ein distanziertes Verhältnis schließen lässt.

3.4.3 Körper

Die Geburt ist eine körperliche Angelegenheit par excellence. So sind auch mehrere Cluster auffallend, die Bezeichnungen für Körperteile oder Körperfunktionen enthalten.

Abb. 5: Dichteverteilung Geburtsberichte: Cluster „Körper“.

Abbildung 5 zeigt einige dieser Cluster und ihre Dichteverteilung. Die verschiedenen Körperzonen spielen naturgemäß zu unterschiedlichen Zeitpunkten in der Geschichte eine Rolle. Der Cluster *Bauchschmerz / Bauchkrämpfe / Unterleibsschmerzen* und *kugelrund / Murrel / Riesenbauch* bilden Startpunkte der

Erzählungen. Auf der anderen Seite ist mit dem Cluster *Brust / nuckeln / Brust*, die (nach Labov und Waletzky) Resolution der Geschichte eingetreten. Dazu gehört aber auch der Cluster *Schamlippenriss / Klitorisriss / II*. (II. verweist auf den Schweregrad der Verletzung), mit dem auch die negativen Begleiterscheinungen einer Geburt thematisiert werden. Auch die Cluster *Leistengegend / Leiste / Leisten* und *Untersuchung / Muttermunduntersuchung / Muttermund* erklären sich aus dem Geburtsablauf heraus als zentral in der ersten Hälfte der Erzählung, bei der die Wehen und das Messen der Öffnung des Muttermundes zentrale Ereignisse sind.

3.4.4 Angst

Ängste sind immer wieder Gegenstand der Geburtsberichte. 6 zeigt sechs Cluster, die Formen von Angst ausdrücken. Aufgrund der Verteilung über die Geschichten lassen sich unterschiedliche Typen von Angst unterscheiden.

Abb. 6: Dichteverteilung Geburtsberichte: Cluster „Angst“

Der Cluster *Angst / bammel / angst* vereint die folgenden Ausdrücke: *Angst, Bammel, Panik, Schiß, angst, bammel, panik, panisch*. Die Verteilung zeigt zwei kleinere Spitzen – gleich zu Beginn und vor dem letzten Drittel und der Cluster scheint eher Ausdrücke unspezifischer Angst vor der Geburt zu repräsentieren.

Etwas anders gelagert ist der Cluster *Heidenangst / Schisser / Horrorszenario*, der relativ groß ist und 57 verschiedene Wörter umfasst. Die Spitze am Ende der Geschichten lässt vermuten, dass es sich auch um Ausdrücke handelt, die in evaluativen Passagen verwendet werden und retrospektiv die begründete oder unbegründete Angst thematisieren. Mit *Alptraum, Desaster, Drama, Horrorszenario, Katastrophe, Todesangst* oder *Tortur* versammelt der Cluster besonders starke Angstausdrücke; genau so dazu gehören aber auch *Hit, Kinderspiel, Klacks, grundlos* und *unbeschadet*, die Ängste relativieren.

Den Clustern *ausgeliefert / Versagerin / ausliefern* sowie *diffus / hammerhart / Bewegungsdrang* und *ohnmächtig / bewußtlos / ersticken* ist gemeinsam, dass ihre Spitzen vor dem Ende ungefähr im dritten Viertel der Geschichten zu finden sind. Sie stammen also oft aus den Erzählpassagen, in denen die Austreibungsphase und Geburt selber erzählt werden. Der Cluster *ausgeliefert / Versagerin / ausliefern* umfasst die Wörter *Versager, Versagerin, ausgeliefert, ausliefern, ausreißen, erschlagen, fuehlte, fühlen, geborgen, hilflos, unwohl, wohler*. Diese Wörter könnten verwendet werden, um einen Moment völligen Kontrollverlusts im Moment der höchsten Schmerzen auszudrücken, für den auch spezifische Mehrworteinheiten wie *ich konnte einfach nicht mehr* oder *ich fing an zu weinen* typisch sind, wie ich in den n-Gramm-Analysen in Bubenhofer (2018: 376) bereits feststellen konnte.

Wie bereits der Cluster *diffus / hammerhart / Bewegungsdrang* zeigt auch *kraß / grausam / schrecklich* (mit: *Hammer, Horror, Hölle, Qual, Schlimmste, fies, gemein, grausam, hammer, höllisch, kraß, schlimm, schrecklich, ätzend*) zwei Spitzen und ist vor allem im Kontext der Beschreibung der Wehen zu finden.

3.4.5 Freude

Schließlich möchte ich semantische Cluster zu Ausdrucksweisen von Freude diskutieren (vgl. 7). Dabei wird deutlich, dass die Erzählungen meistens in einer Coda enden, in der fast ausschließlich äußerst positive Emotionen ausgedrückt werden. Der Cluster *verliebt / verlieben / verzaubern* ist deutlichstes Beispiel dafür. Ähnlich zentral ist der Cluster *unvergleichlich / ergreifend / atemberaubend*, der die ganze Palette an Ausdrucksweisen von Freude aber auch Magie umfasst, also das Erlebnis als ganz besonderen, individuellen, magischen Moment rekonstruiert: *Glücksmoment, Grenzerfahrung, Kraftakt, Naturgewalt, Traum, Urgewalt*,

atemberaubend, beeindruckend, beschreiben, betrügen, bewegend, bezeichnen, durchleben, einmalig, einschneidend, emotional, empfunden, ergreifend, faszinierend, grauenhaft, kraftvoll, magisch, präsent, schmerzlich, unvergleichlich, unvorstellbar, wahnsinns, wahrlich, wertvoll, zurückdenken.

Abb. 7: Dichteverteilung Geburtsberichte: Cluster „Freude“.

Die Cluster *komplikationslos / unkompliziert / verlaufen* und *komplikationslos / unbeschwert / unproblematisch* sind sowohl am Anfang als auch am Ende der Geschichten ausgeprägt. Das rührt daher, dass sie nicht nur die Geburt charakterisieren können, sondern insbesondere auch die Schwangerschaft, deren Verlauf oft Thema am Beginn der Geschichten ist. Die folgenden dem zweiten der beiden Cluster zugeordneten Wörter deuten dies an: *Anfangszeit, Bilderbuch, ICSI, Schnelligkeit, Traumhaft, Traumschwangerschaft, ansich, beschwerdefreie, beschwerlich, chaotisch, harmonisch, komplikationsfrei, komplikationslos, langwierig, nervenaufreibend, problematisch, prägen, schmerzarm, turbulent, unbeschwert, unerwartet, ungeplant, unproblematisch, unschön, unspektakulär, überraschend, überschatten.*

Durch das Clustering der Word Embeddings konnten Erkenntnisse zu fünf zentralen Elementen der Geburtsberichte gewonnen werden, wobei diese Elemente unterschiedlicher Art sind:

- **Erzählfunktionen:** Insbesondere die Cluster der Gruppe „Einbruch in den Alltag“ sind Beispiele, um typische Strukturen und Gliederungen von

Geschichten zu rekonstruieren, also (im Sinne von Labov/Waletzky 1973) Erzählfunktionen zu bestimmen.

- **Rollen:** Die Cluster der Gruppe „Personal“ helfen dabei, die unterschiedlichen Rollen in den Geschichten und ihre Stellung zueinander zu erfassen.
- **Themen:** Die Cluster der drei Bereiche „Körper“, „Angst“ und „Freude“ tragen zum Verständnis zentraler Themen in den Geschichten bei.

Es versteht sich von selbst, dass sich diese Erkenntnisse nicht automatisch aus den Daten heraus ergeben, sondern Resultat eines interpretativen Prozesses sind, der vor dem Hintergrund einer bestimmten Erzähltheorie stattfindet. Der Mehrwert liegt jedoch in der empirischen Fundierung und dem datengeleiteten Zugang: Dass die gefundenen Elemente in Geburtsberichten zentral sind, ist nicht unbedingt überraschend. Welches Vokabular dahinter steckt, zu großen Teilen jedoch schon.¹¹

4 Fazit

Die dargestellte Methode des Clusterings von Word-Embedding-Profilen erweist sich als äußerst hilfreich, um zur Modellierung eines Narrativs, hier am Beispiel von Geburtsberichten, beizutragen. Die Methode kann aber nur eine unter mehreren korpuslinguistischen Verfahren sein und besteht zudem aus wichtigen hermeneutischen Schritten auf der Basis einer bestimmten – jeweils den Forschungsinteressen angepassten – Erzähltheorie. Die Methode leistet m. E. Folgendes:

- **Lexikalische Variation:** Die Cluster erfassen automatisch eine erstaunlich breite lexikalische Variation, selbst seltene, normferne orthographische Varianten, deren Breite nur datengeleitet abgedeckt werden kann.
- **Diskursspezifische semantische Äquivalenz:** Die strukturalistisch geprägte Vorstellung von semantischen Relationen kann durch die Idee einer diskursspezifischen semantischen Äquivalenz abgelöst werden, was konsequent einem gebrauchsbasierten Paradigma folgt.
- **Textdynamische Abhängigkeit:** Wenn die Cluster zusätzlich auf ihre Distribution über den Verlauf der Texte hin analysiert werden, erweitert diese zusätzliche Analysedimension die Interpretationsmöglichkeiten: In den Blick gelangt die Abhängigkeit semantischer Cluster von bestimmten Erzählfunktionen.

¹¹ Die vollständigen Datensätze und Dichteberechnungen ausgewählter Cluster sind aber online einsehbar (vgl. www.bubenhofer.com/publikationen/geburtsberichte/; 01.04.2020).

- **Relationen zwischen semantischen Clustern:** Wenn alle drei Dimensionen, Variation, semantische Äquivalenz und textdynamische Abhängigkeit, analytisch kombiniert werden, spiegeln sich darin komplexe Relationen zwischen semantischen Clustern diskursspezifisch wider. So werden beispielsweise die in den Geschichten dargestellten Hierarchieverhältnisse des Personals deutlich. Die Körpercluster wiederum zeigten die Semantisierung des Körpers – welche Körperteile werden überhaupt benannt, wie werden sie benannt und welche Rolle spielen sie in den Geschichten? Die Angst- und Freudecluster bringen Licht in die Versprachlichung von Gefühlen, wiederum in Abhängigkeit spezifischer Erzählfunktionen.

Aus einer diskurslinguistischen oder kulturanalytischen Sicht aktualisieren Word Embeddings eine alte distributionalistische Idee von Semantik, zurückgehend auf Wittgenstein (1995: PU 45) und Firth (1957: 190). Besonders fruchtbar für diskurs- und kulturanalytische Zwecke werden Word Embeddings aber dann, wenn sie zur Modellierung diskursspezifischer semantischer Räume genutzt werden und so gerade die Abweichungen zwischen verschiedenen semantischen Räumen systematisch analysiert werden können.

Mit einer Erzähltheorie im Hintergrund, wie etwa jener von Labov und Waletzky (Labov/Waletzky 1973), lassen sich die datengeleitet berechneten semantischen Räume neu deuten – und umgekehrt die Erzähltheorie um diskurslinguistische Aspekte ergänzen. Labov und Waletzky arbeiten beispielsweise mit den Kategorien „Orientierung“, „Komplikation“, „Evaluation“, „Resolution“ und „Coda“, um die wichtigen Phasen von Alltagserzählungen zu beschreiben (Labov/Waletzky 1973: 112ff.). Die berechneten semantischen Cluster und ihre Distribution über den Verlauf der Geschichten lassen sich vor der Folie dieser Kategorien interpretieren, und dies auf der Basis von mehr als 14.000 Geschichten. Die Kategorien lassen sich so differenzierter verstehen und auf einen bestimmten Diskurs hin – Geburten – adaptieren. So lässt sich etwa der Übergang von der Orientierungsphase (Situierung der Geschichte in Bezug auf Person, Ort, Zeit und Handlungssituation) zur Komplikation (Beginn der Ereignisse) genauer beschreiben, wie die semantischen Cluster unter der Überschrift „Einbruch in den Alltag“ zeigen.

Gleichzeitig ist eine Anreicherung der narrativen Phasen möglich, indem etwa die Figuren in den Geschichten und ihre Beziehungen zueinander und Rollen in der Erzählung herausgearbeitet werden können. Dies zeigten die Personen-Cluster, bei denen eben auch sichtbar wird, wie die jeweiligen Figuren benannt werden: Mit Berufsbezeichnungen, Abkürzungen, Kosenamen oder eben Vornamen und Vornamengruppen.

Die Experimente zeigen aber auch das Potenzial für eine gebrauchssemantisch, diskurslinguistisch orientierte Analyse: Sie ermöglichen die datengeleitete Identifikation von abstrakter fassbaren Clustern von ähnlich verwendeten Wörtern, die verdeutlichen, dass „Ähnlichkeit“ hier so etwas wie eine funktionale Äquivalenz in einem bestimmten Diskurs ausdrückt. Um etwa das „Personal“ zu identifizieren, das in Geburtserzählungen wichtig ist, muss nicht in onomasiologischer Richtung von Personen ausgegangen und erst definiert werden, wie diese genannt werden könnten, sondern es ergeben sich datengeleitete Cluster von Ausdrucksformen mit ähnlicher Verwendung. Die breite Variation der Benennungen wird dabei sofort ersichtlich und die spezifischen Funktionen der einzelnen semantischen Cluster können rekonstruiert werden.

Literatur

- Arthur, David / Vassilvitskii, Sergei (2006): k-means++: The Advantages of Careful Seeding. Stanford InfoLab. Online unter: <http://ilpubs.stanford.edu:8090/778/>.
- Barbieri, Gian Luca et al. (2012): Looking forward to the birth of a child: tales of motherhood in forums. In: *International Journal of Psychoanalysis and Education IJPE* 4. URL: <http://www.psych.edu.org/index.php/IJPE/article/view/61/48> (01.04.2020).
- Boleda, Gemma/Herbelot, Aurélie (2016): Formal Distributional Semantics: Introduction to the Special Issue. In: *Computational Linguistics* 42, S. 619–635.
- Boueke, Dietrich (1995): *Wie Kinder erzählen: Untersuchungen zur Erzähltheorie und zur Entwicklung narrativer Fähigkeiten*. München: Fink.
- Bubenhof, Noah (2015): Kollokationen, n-Gramme, Mehrworteinheiten. In: Roth, Kersten / Wengeler, Martin / Ziem, Alexander (Hg.): *Handbuch Sprache in Politik und Gesellschaft*. Berlin / New York: De Gruyter (= Handbücher Sprachwissen 19), S. 69–93.
- Bubenhof, Noah (2018): Serialität der Singularität: Korpusanalyse narrativer Muster in Geburtsberichten. In: *Zeitschrift für Literaturwissenschaft und Linguistik*, S. 1–32.
- Bubenhof, Noah / Calleri, Selena / Dreesen, Philipp (2019): Politisierung in rechtspopulistischen Medien: Wortschatzanalyse und Word Embeddings. In: *OBST. Osnabrücker Beiträge zur Sprachtheorie*, S. 211–242.
- Bubenhof, Noah / Müller, Nicole / Scharloth, Joachim (2013): Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz. In: *Zeitschrift für Semiotik, Methoden der Diskursanalyse* 35, S. 419–444.
- Bubenhof, Noah / Scharloth, Joachim (2011): Korpuspragmatische Analysen alpinistischer Literatur. In: Elmiger, Daniel / Kamber, Alain (Hg.): *La linguistique de corpus – de l'analyse quantitative à l'interprétation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretation*. Neuchâtel: Institut des sciences du langage et de la communication (= Travaux neuchâtelais de linguistique 55). S. 241–259.
- Bubenhof, Noah / Scharloth, Joachim (2013): Korpuspragmatische Methoden für kulturanalytische Fragestellungen. In: Benitt, Nora et al. (Hg.): *Kommunikation Korpus Kultur: Ansätze und Konzepte einer kulturwissenschaftlichen Linguistik*. Trier: WVT (= Giessen Contributions to the Study of Culture). S. 47–66.

- Bubenhofer, Noah / Scharloth, Joachim (2015): Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate. In: *Zeitschrift für Germanistische Linguistik* 43, S. 1–26.
- Colloseus, Cecilia (2016): Gebären – Erzählen. Kulturalanthropologische und interdisziplinäre Perspektiven auf die Geburt als leibkörperliche Grenzerfahrung. Mainz: Johannes Gutenberg-Universität Mainz.
- Ehlich, Konrad (Hg.) (1980): *Erzählen im Alltag*. Frankfurt am Main: Suhrkamp. (= Suhrkamp-Taschenbuch Wissenschaft).
- Evert, Stefan (2009): 58. Corpora and collocations. In: Lüdeling, Anke / Kytö, Merja (Hg.): *Corpus Linguistics*. Berlin / New York: De Gruyter (= Handbücher zur Sprach- und Kommunikationswissenschaft 29). S. 1212–1248.
- Firth, John Rupert (1957): *Modes of Meaning*. In: *Papers in Linguistics 1934–1951*. London: Oxford University Press. S. 190–215.
- Gülich, Elisabeth (1980): Konventionelle Muster und kommunikative Funktionen von Alltagserzählungen. In: *Erzählen im Alltag*. Frankfurt am Main: Suhrkamp (= Suhrkamp-Taschenbuch Wissenschaft). S. 335–384.
- Kienpointner, Manfred (2017): 9. Topoi. In: Roth, Kersten Sven / Wengeler, Martin / Ziem, Alexander (Hg.): *Handbuch Sprache in Politik und Gesellschaft*. Berlin / Boston: De Gruyter. S. 187–211.
- Knapp, Werner (1997): *Schriftliches Erzählen in der Zweitsprache*. Berlin / Boston: De Gruyter.
- Labov, William (2013): *The Language of Life and Death: The Transformation of Experience in Oral Narrative*. Cambridge: Cambridge University Press.
- Labov, William / Waletzky, Joshua (1973): Erzählanalyse. Mündliche Versionen persönlicher Erfahrung. In: Ihwe, Jens (Hg.): *Literaturwissenschaft und Linguistik*. Frankfurt am Main: Athenäum Fischer (= Eine Auswahl Texte zur Theorie der Literaturwissenschaft). S. 78–126.
- Lenci, Alessandro (2018): Distributional Models of Word Meaning. In: *Annual Review of Linguistics* 4, S. 151–171.
- Mikolov, Tomas et al. (2013): Efficient Estimation of Word Representations in Vector Space. In: arXiv:1301.3781 [cs].
- Moretti, Franco (2009): *Kurven, Karten, Stammbäume. Abstrakte Modelle für die Literaturgeschichte*. Frankfurt am Main: Suhrkamp (= edition suhrkamp 2564).
- Quasthoff, Uta (1980): *Erzählen in Gesprächen: Linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags*. Tübingen: Narr (= Kommunikation und Institution).
- Řehůřek, 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: ELRA. S. 45–50.
- Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK. URL: <https://pdfs.semanticscholar.org/bd0b/ab6fc8cd43c0ce170ad2f4cb34181b31277d.pdf> (01.04.2020).
- Storjohann, Petra (2015): 11. Sinnrelationale Wortschatzstrukturen: Synonymie und Antonymie im Sprachgebrauch. In: Haß, Ulrike / Storjohann, Petra (Hg.): *Handbuch Wort und Wortschatz*. Berlin / Boston: De Gruyter. S. 248–273.
- Tienken, Susanne (2015): Von der Fehlgeburt zum Sternenkind. Ein Neologismus und seine kulturelle Bedeutung. In: Ångsal, Magnus P. / Grub, Frank-Thomas (Hg.): *Visionen und Illusionen: Beiträge zur 11. Arbeitstagung schwedischer Germanistinnen und Germanisten*

Text im Kontext in Göteborg am 4./5. April 2014. Frankfurt am Main: Peter Lang Edition (= Nordeuropäische Arbeiten zur Literatur, Sprache und Kultur). S. 129–150.

Wengeler, Martin (2003): Argumentationstopos als sprachwissenschaftlicher Gegenstand.

Für eine Erweiterung linguistischer Methoden bei der Analyse öffentlicher Diskurse.

In: Geideck, Susan / Liebert, Wolf-Andreas (Hg.): Sinnformeln. Linguistische und soziologische Analysen von Leitbildern, Metaphern und anderen kollektiven Orientierungsmustern. Berlin / New York: De Gruyter (= Linguistik – Impulse & Tendenzen).

S. 59–82.

Wittgenstein, Ludwig (1995): *Tractatus logico-philosophicus*. Tagebücher 1914–1916. Philosophische Untersuchungen. 2. Frankfurt am Main: Suhrkamp. (= Werkausgabe 1).