

Carolin Müller-Spitzer / Sascha Wolfer (Mannheim)

Vernetzungsstrukturen digitaler Wörterbücher. Neue Ansätze zur Analyse

- | | | | |
|-----|---|-----|--|
| 1 | Zum Gegenstand dieses Beitrags | 4 | Vernetztheit – Korpusfrequenz – Nachschlagehäufigkeit |
| 2 | Eine neue Übersicht über die Vernetzungsstruktur: Die Vernetzung paradigmatischer Angaben als Gesamtgraph | 4.1 | Sind frequente Wörter paradigmatisch stärker vernetzt? |
| 3 | Detektion besonders stark vernetzter Stichwortgruppen mit Mitteln der Graphentheorie | 4.2 | Werden paradigmatisch stark vernetzte Stichwörter häufiger nachgeschlagen? |
| 3.1 | Synonym-Cliquen | 4.3 | Nachschlagehäufigkeit – Cliquen – Cluster |
| 3.2 | Synonym-Cluster | 5 | Ausblick |
| | | 6 | Literatur |

Abstract: In this contribution, we present a novel approach for the analysis of cross-reference structures in digital dictionaries on the basis of the complete dictionary database. Using paradigmatic items in the German Wiktionary as an example, we show how analyses based on graph theory can be fruitfully applied in this context, e. g. to gain an overview of paradigmatic references as a whole or to detect closely connected groups of headwords. Furthermore, we connect information about cross-reference structures with corpus frequencies and log file statistics. In this way, we can answer questions such as the following ones: Are frequent words paradigmatically linked more closely than others? Are closely linked headwords or headwords that stand more solitary in the dictionary visited significantly more often?

Keywords: Mediostrukturen, Vernetzungsstrukturen, Graphentheorie, Benutzungsforschung, Logfileanalyse / mediostructures, cross-reference structures, graph theory, research into dictionary use, log file analysis

„The scientific value of heuristic statistical methods may be illustrated by a metaphor. You will see different things if you walk, ride a bike, go by car, or look down from a plane. Statistics is a vehicle which can be used at arbitrary ‘velocity’ and arbitrary ‘height’, depending on how much overview you wish for and how detailed you want to look at a linguistic ‘landscape’.” (Köhler 2012: 16)

1 Zum Gegenstand dieses Beitrags

Gegenstand der Vernetzungs- oder Mediostrukturen digitaler oder gedruckter Wörterbücher sind allgemein gesagte lexikografische Verweisphänomene. Die Mediostruk-

tur gedruckter Wörterbücher wird in der Regel durch exemplarische Sichtung der Wortartikel eines Wörterbuchs oder mehrerer Wörterbücher analysiert, oder – wie Kammerer es ausdrückt – auf Basis einer „empirischen, nicht exhaustiven Untersuchung“ (Kammerer 1998: 325; für andere Beispiele solcher Untersuchungen s. z. B. Lindemann 1999 oder Müller 2002). Dabei werden verweisrelevante Textsegmente (i. S. v. Wiegand 2002: 175–177) nach verschiedenen Kriterien, beispielsweise nach den verweisinitiiierenden Angaben, den Verweisbeziehungen oder den Verweisadressen in eine Ordnung gebracht, sodass die Mediostruktur eines Wörterbuchs oder einer Klasse von Wörterbüchern möglichst genau beschrieben werden kann (zu allen relevanten Termini s. Wiegand 2002). Datengrundlage für diese Art der Analyse ist ein gedrucktes Buch, aus dem durch Lesen und kognitive Verarbeitungen Informationen gewonnen und klassifiziert werden.

Bei digitalen Wörterbüchern erfolgt die Analyse der Wörterbuchstrukturen oftmals ähnlich (vgl. z. B. Mann 2010: 28–29, 36–38). Sie kann aber auch ganz anders erfolgen, wenn man als Datengrundlage die gesamte digitale Datenbasis eines Wörterbuchs zugrunde legt und diese Datenbasis vollständig mithilfe statistischer Methoden auswertet. Genau dies wollen wir im Folgenden am Beispiel der paradigmatischen Angaben, d. h. den Angaben zu Synonymen, Antonymen, Sinn- und Sachverwandten und Ober- bzw. Unterbegriffen, des deutschen Wiktionary zeigen.

Vorab einige terminologische Klärungen: Das Konzept der *Mediostruktur* wurde für gedruckte Wörterbücher entwickelt und sollte daher nicht unreflektiert auf digitale Wörterbücher übertragen werden (vgl. u. a. Tarp 2008: 102 und Müller-Spitzer 2013). Außerdem muss bei digitalen Wörterbüchern immer die Ebene der Datenbasis von der der Präsentation getrennt werden (Blumenthal et al. 1988, Müller-Spitzer 2007: 170–180, Müller-Spitzer 2013: 368–372). Für die Ebene der Datenbasis wurde der Terminus *Vernetzungsstruktur* (Müller-Spitzer 2007: 169–180, Meyer 2014) vorgeschlagen, für die Ebene der Präsentation der Begriff *mediostrukturelle Einheiten*, da die Mediostruktur als Ganzes nur bei Sichtung der Präsentation in ihrem gesamten Umfang, besonders bei innovativen digitalen Wörterbüchern (vgl. Prinsloo et al. 2012, Tarp 2012), nicht sicher zu identifizieren ist (vgl. Müller-Spitzer 2013: 368–378). Verweise sollen nach Wiegand (2002: 180) als kognitive Entitäten, die aus Angaben zu erschließen sind, aufgefasst werden.

Ein einfaches Beispiel zur Verdeutlichung der notwendigen Trennung der Ebene der Datenbasis vs. der der Präsentation: In einer lexikographischen Datenbasis können flektierte Formen eines Verbs, z. B. die Form „ging“, dem zugehörigen Verb, hier „gehen“, zugeordnet sein, und zwar als Element der hierarchischen Baumstruktur der Artikelinhalte zum Stichwort „gehen“. Sollte ein digitales Wörterbuch aus dieser Datenbasis entwickelt werden, könnte aus diesen Knoten des Baumes ein Verweisartikel der Form „ging, s. *gehen*“ erstellt werden, wenn diese Art des Zugriffs nicht über die Suchfunktion abgedeckt würde. Auf der Ebene der Datenbasis würde „ging“ damit kein Element der Vernetzungsstruktur sein, da keine Vernetzung auf ein anderes Element angelegt wurde (analog zur Beschreibung in Müller-Spitzer 2013).

Auf der Ebene der Präsentation würden diese Angaben allerdings mediostrukturellen Einheiten entsprechen. Dementsprechend können Angaben auf der Ebene der Datenbasis vs. der Ebene der Präsentation ggf. verschiedenen textuellen Strukturen, nämlich einmal den Inhaltsstrukturen (i. S. v. Müller-Spitzer 2007: 152–169) und einmal den Mediostrukturen zugeordnet werden.

Wir haben in unserer Untersuchung zum Wiktionary allerdings den besonderen Fall, dass die Kodierung der Datenbasis eng an die Präsentation angelehnt ist, d. h. die sonst vorgetragenen Argumente zur Unterscheidung der Vernetzungsstruktur auf der Ebene der Datenbasis und den Mediostrukturen auf Ebene der Präsentation gelten hier nur begrenzt. Da es sich bei der Auszeichnungssprache des Mediawiki¹, die im Wiktionary zur Kodierung der Datenbasis eingesetzt wird, um eine layoutorientierte und keine inhaltsorientierte Datenauszeichnung handelt (i. S. v. Müller-Spitzer 2007: 100–103), finden sich auch auf der Ebene der Datenbasis – anders als bei einer konzeptuellen Inhaltsmodellierung – vorwiegend Kodierungen zur Präsentationsform. Dementsprechend werden alle Elemente, die in der Datenbasis als Links gekennzeichnet sind, und auch nur diese, auf der Ebene der Präsentation als mediostrukturelle Einheiten dargestellt.

Gehören die Daten, die wir im Folgenden analysieren, also nun zur Trägermenge der Mediostruktur der deutschen Edition des Wiktionary oder zur Vernetzungsstruktur? Unserer Meinung nach analysieren wir im Folgenden einen Teil der Vernetzungsstruktur des Wiktionary, da wir die Datenbasis nicht durch Screen Scraping² der online verfügbaren Wortartikel gewonnen haben, sondern durch Herunterladen des Wiktionary-Dumps, d. h. der Datenbasis des Wiktionary.³

Gegenstand unserer Analyse waren also Daten wie in Abbildung 1 gezeigt und nicht der online verfügbare Wortartikel (vgl. Abb. 2). Wir denken trotzdem, dass wir mit unseren Analysen Impulse für die Beschreibung von Vernetzungs- und Mediostrukturen oder mediostrukturellen Einheiten digitaler Wörterbücher geben können. Gleichzeitig werden wir aufzeigen, wie mediostrukturelle Analysen sinnvoll in Beziehung zu anderen (Meta-)Daten wie Korpusfrequenzen und Nachschlagehäufigkeiten gesetzt werden können.

¹ <https://www.mediawiki.org/wiki/MediaWiki> (zuletzt eingesehen am 1. April 2015)

² http://de.wikipedia.org/wiki/Screen_Scraping (zuletzt eingesehen am 1.4.2015).

³ Datengrundlage für unsere Untersuchung ist die Datenbasis des deutschen Wiktionary vom 24. Dezember 2014, d. h. alle zu der Zeit verfügbaren Wortartikel, deren Stichwörter mit dem Zusatz „(deutsch“) gekennzeichnet sind. Eine Seite im Wiktionary kann mehrere Wortartikel umfassen. In diesem Falle wurde immer der erste deutsche Wortartikel verwendet. Dumps aller Wikimedia-Produkte sind verfügbar unter <https://dumps.wikimedia.org/> (zuletzt eingesehen am 1.4.2015).

```

{{Wort der Woche|3|2006}}
== kalt ({{Sprache|Deutsch}}) ==
=== {{Wortart|Adjektiv|Deutsch}} ===
{{Deutsch Adjektiv Übersicht
|Positiv=kalt
|Komparativ=kälter
|Superlativ=am kältesten
}}
{{Worttrennung}}
:kalt, {{Komp.}} käl·ter, {{Sup.}} am käl·tes·ten
[...]

{{Synonyme}}
:[2] [[abgehärtet]], [[abgestumpft]], [[abweisend]], [[alexithymisch]],
[[cool]], [[eisig]], [[empfindungslos]], [[emotionslos]], [...]
:[3] [[nüchtern]], [[spartanisch]], [[ungemütlich]]

{{Gegenwörter}}
:[1] [[warm]], [[heiß]], [[tropisch]]
:[2] [[barmherzig]], [[empathisch]], [[freundlich]], [...]
:[3] [[behaglich]], [[einladend]], [[gemütlich]]
:[4] [[warm]]

{{Oberbegriffe}}
:[1] [[kühl]]

{{Unterbegriffe}}
:[1] [[winterlich]], [[arktisch]], [[antarktisch]], [[polar]]
:[1] [[arschkalt]], [[bitterkalt]], [[eisig]], [[eisigkalt]], [...]

```

Abb. 1: . Datenbasis zum Artikel „kalt“ im deutschen Wiktionary (Stand: 26.3.2015).




kalt (Deutsch) [Bearbeiten]		
Adjektiv [Bearbeiten]		
Worttrennung: kalt, Komparativ: käl-ter, Superlativ: am käl-tes-ten	Positiv	Komparativ
	kalt	kälter
		am kältesten
Aussprache: IPA: [kalt], Komparativ: [ˈkɛltɐ], Superlativ: [ˈkɛltɛstn̩] Hörbeispiele:  kalt ^(Info) , Komparativ:  kälter ^(Info) , Superlativ:  am kältesten ^(Info) Reime: -alt		
Bedeutungen: [1] eine niedrige Temperatur habend; von/mit niedriger Temperatur [2] kein Mitgefühl habend oder zeigend [3] ungemütlich [4] von <i>Mietpreisen</i> : ohne Nebenkosten, Betriebskosten		
Herkunft: Die älteren Entsprechungen (<i>althochdeutsch</i> : <i>kalt</i> → <i>goh</i> , <i>mittelhochdeutsch</i> : <i>kalt</i> → <i>gmh</i> , gotisch: <i>kalds</i> → <i>got</i> , altnordisch: <i>kaldr</i> → <i>non</i>) gehen auf ein Partizip zum germanischen Verb *kala- „frieren, kalt werden“ zurück; verwandte Wörter in nicht germanischen Sprachen sind etwa <i>lateinisch</i> : <i>gelu</i> → <i>la</i> Frost, Kälte ^[1]		
Synonyme: [2] abgehärtet, abgestumpft, abweisend, alexithymisch, cool, eisig, empfindungslos, emotionslos, erbarmungslos, frigid, frigide, frostig, gefühllos, gefühlsarm, gefühlskalt, gleichgültig, grausam, hart, hartherzig, herzlos, indifferent, kalblütig, kaltherzig, kühl, lieblos, mitleidslos, unbarmherzig, unfreundlich, skrupellos, steinern, ungerührt [3] nüchtern, spartanisch, ungemütlich		
Gegenwörter: [1] warm, heiß, tropisch [2] barmherzig, empathisch, freundlich, gefühlvoll, herzlich, liebevoll, mitfühlend, mitleidsvoll, warm, warmherzig [3] behaglich, einladend, gemütlich [4] warm		
Oberbegriffe: [1] kühl		
Unterbegriffe: [1] winterlich, arktisch, antarktisch, polar [1] arschkalt, bitterkalt, eisig, eisigkalt, eiskalt, eiseskalt, frostig, gekühlt, hundekalt, lausekalt, saukalt, scheidkalt		

Abb. 2. Artikel „kalt“ im deutschen Wiktionary (Stand: 26.3.2015).

Auch wenn wir es bei unseren Daten mit Elementen der Trägermenge der Vernetzungsstrukturen zu tun haben, kann es einer genaueren Beschreibung dienlich sein, die Terminologie der Mediostrukturen gedruckter Wörterbücher auf solche Vernetzungsstrukturen zu übertragen, da nur dieser Bereich terminologisch weit durchdrungen ist (vgl. u. a. Wiegand/Smit 2013 und die auf S. 214 zitierte Literatur, Wiegand 2002, Tarp 1999, Meyer 2014). Ausgehend von dieser Terminologie könnte man unseren Untersuchungsgegenstand folgendermaßen beschreiben: Unsere Datengrundlage ist ein Teil der konkreten Mediostruktur der deutschen Edition des Wiktionary, d. h. eine bestimmte partielle konkrete Mediostruktur. Trägermenge dieser partiellen Mediostruktur sind alle Angaben mit Verweiskennzeichnung (i. S. v. Wiegand 2002: 236), die in den Angabeklassen „Synonyme“ „Sinnverwandte Wörter“, „Gegenwörter“, „Oberbegriffe“ und „Unterbegriffe“ vorkommen und die an eine bestehende lemmatische Verweisaußenadresse adressiert sind. Alle Angaben innerhalb dieser Angabegruppen sind als Links dargestellt, sodass es sich eindeutig um Angaben mit Verweiskennzeichnung handelt (zu einem Beispiel von Synonymangaben ohne Verweiskennzeichnung s. Kammerer 1998: 323). Diejenigen verweisvermittelnden Angaben, die eine nicht bestehende lemmatische Verweisaußenadresse adressieren (im Wiktionary rot

gekennzeichnet), haben wir von der Analyse ausgeschlossen. Auf dieser Datengrundlage beruhen alle hier dargestellten Analysen. Allgemeiner gesagt analysieren wir im Folgenden alle Verweisangaben aus den fünf Angabegruppen „Synonyme“, „Sinnverwandte Wörter“, „Gegenwörter“, „Oberbegriffe“ und „Unterbegriffe“ des jeweiligen Stichworts, die einen vorhandenen Wortartikel als Link-Ziel haben.

Noch eine letzte Vorbemerkung dazu, warum wir gerade das Wiktionary ausgewählt haben, obwohl dessen lexikographische Qualität zum Teil kritisch betrachtet wird (u. a. Fuertes-Olivera 2009, Hanks 2012: 77–82). Unser Ziel ist es, die Möglichkeiten der Analyse von Vernetzungsbeziehungen mit der gesamten Datenbasis des Wörterbuchs als Datengrundlage zu explorieren. Voraussetzung für die Analysen, wie wir sie in diesem Beitrag zeigen wollen, ist demnach keine nach aktuellem wissenschaftlichen Stand reflektierte lexikographische Praxis, sondern schlicht, dass wir eine große lexikographische Datenbasis haben, die uns frei zur Verfügung steht, und dass wir diese Datenbasis mit anderen uns frei zugänglichen Daten, beispielsweise mit aus Logfiles ermittelbaren Nachschlagehäufigkeiten, in Verbindung bringen können. Diese beiden Voraussetzungen gelten unseres Wissens nur für das Wiktionary. Wir könnten theoretisch auch Wörterbücher unseres Arbeitsgebers (IDS Mannheim, für die Onlinewörterbücher s. www.owid.de) als Datenbasis verwenden, diese wären allerdings wesentlich kleiner, sodass die Analysen mit dem hier vorgestellten Schwerpunkt nicht durchgeführt werden könnten. Trotzdem wäre es natürlich interessant, diese Analysen auf andere lexikographische Daten, z. B. auf von lexikographischen Expertinnen und Experten erarbeitete Wörterbücher, zu übertragen.

Dieser Beitrag ist folgendermaßen gegliedert: In Abschnitt 2 zeigen wir eine Übersicht über die Vernetzung paradigmatischer Angaben im Wiktionary. Dabei werden alle Vernetzungen in einem Gesamtgraphen dargestellt und erste Grundanalysen vorgestellt. Ein Graph ist – allgemein formuliert – eine Struktur, die aus Knoten und Verbindungen zwischen Knoten, sog. Kanten, besteht. In unserem Fall werden die Wortartikel die Knoten sein. Jede Verweisangabe auf einen anderen Wortartikel ist der Ausgangspunkt einer Kante, die diesen Knoten mit einem zweiten verbindet. Der Schwerpunkt liegt dabei weniger auf der Interpretation der Einzelergebnisse, als auf der Darstellung der Analysemöglichkeiten. Wie man aus dieser Gesamtmenge mithilfe von Methoden aus dem Bereich der Graphentheorie besonders stark vernetzte Stichwortgruppen detektieren kann, wird in Abschnitt 3 gezeigt. Im vierten Abschnitt wird die Vernetztheit der analysierten Stichwörter zu Daten über ihre Nachschlagehäufigkeit und Korpusfrequenz in Bezug gesetzt. Der Beitrag endet mit einem Ausblick.

2 Eine neue Übersicht über die Vernetzungsstruktur: Die Vernetzung paradigmatischer Angaben als Gesamtgraph

Als eine erste Annäherung an die Angaben zur paradigmatischen Vernetzung im Wiktionary sollen zunächst alle vernetzungsrelevanten Angaben in einer Gesamtdarstellung visualisiert werden, sozusagen eine Weltkarte der paradigmatischen Angaben im Wiktionary gezeichnet werden. Einen solchen Gesamtüberblick kann man eben nur dann zeigen, wenn die gesamten vernetzungsrelevanten Daten digital zur Verfügung stehen. Eine dementsprechende erste Übersicht über die paradigmatischen Vernetzungen im Wiktionary erhält man, wenn man alle aus- und eingehenden Kanten aus allen fünf relevanten Angabeklassen in einem Gesamtgraphen darstellt (vgl. Abb. 3). In diesem Graphen werden zur besseren Übersichtlichkeit nur die Knoten (= Stichwörter) und nicht die Verbindungen unter ihnen (= Kanten) dargestellt. Dabei bilden sich deutlich drei Gruppen heraus: die der Verben, Substantive und Adjektive. Die größte Gruppe ist dabei diejenige der Substantive. Die Visualisierungsroutine, die zur Darstellung des Graphen verwendet wird, ordnet jene Stichwörter räumlich nah nebeneinander an, zwischen denen viele Verbindungen bestehen. Aus der Darstellung des Gesamtgraphen lässt sich somit ablesen, dass die paradigmatischen Vernetzungen, wie man es erwarten würde, v. a. zwischen Stichwörtern der gleichen Wortart angelegt sind. In diesem Gesamtbild lässt sich außerdem erkennen, dass eine große Gruppe von Stichwörtern in der Peripherie des Graphen angeordnet wird. Dies sind Stichwörter, die nur sehr schwach mit anderen Stichwörtern vernetzt sind. Das ist beispielsweise dann der Fall, wenn zwei Stichwörter miteinander verbunden sind, aber keine Verbindung in den Restgraphen vorliegt. Die gedruckte Darstellung hier ist in der Größe wie auch in der schwarz/weiß-Darstellung nur unzulänglich; eine digitale Version dieses Graphen, die eine vergrößerte Darstellung von Ausschnitten durch „Hineinzoomen“ erlaubt, haben wir daher online zur Verfügung gestellt.⁴ Wollte man den Gesamtgraphen wirklich gut explorierbar gestalten, müsste auch diese digitale Darstellung allerdings noch deutlich überarbeitet werden, beispielsweise durch interaktive Elemente wie das Ein- und Ausklappen von Knoten.

Innerhalb unserer gesamten Datengrundlage sind 179.541 von insgesamt 240.402 Stichwörtern, also etwa 75% des gesamten Stichwortbestandes, mit keinem anderen Stichwort verknüpft und daher nicht in den Graphen in Abbildung 1 integriert. Die größte Gruppe mit 150.000 Elementen darin sind die flektierten Formen, die im Wiktionary eigene Wortartikeleinträge haben. Da diese flektierten Formen nicht paradigmatisch vernetzt sind, haben wir sie aus allen folgenden Analysen ausgeschlossen; die anderen unverbundenen Einträge aber in unserer Datengrundlage belassen.

⁴ <<http://www.ids-mannheim.de/fileadmin/lexik/bilder/all.links.pdf>> (zuletzt eingesehen am 1.4.2015).

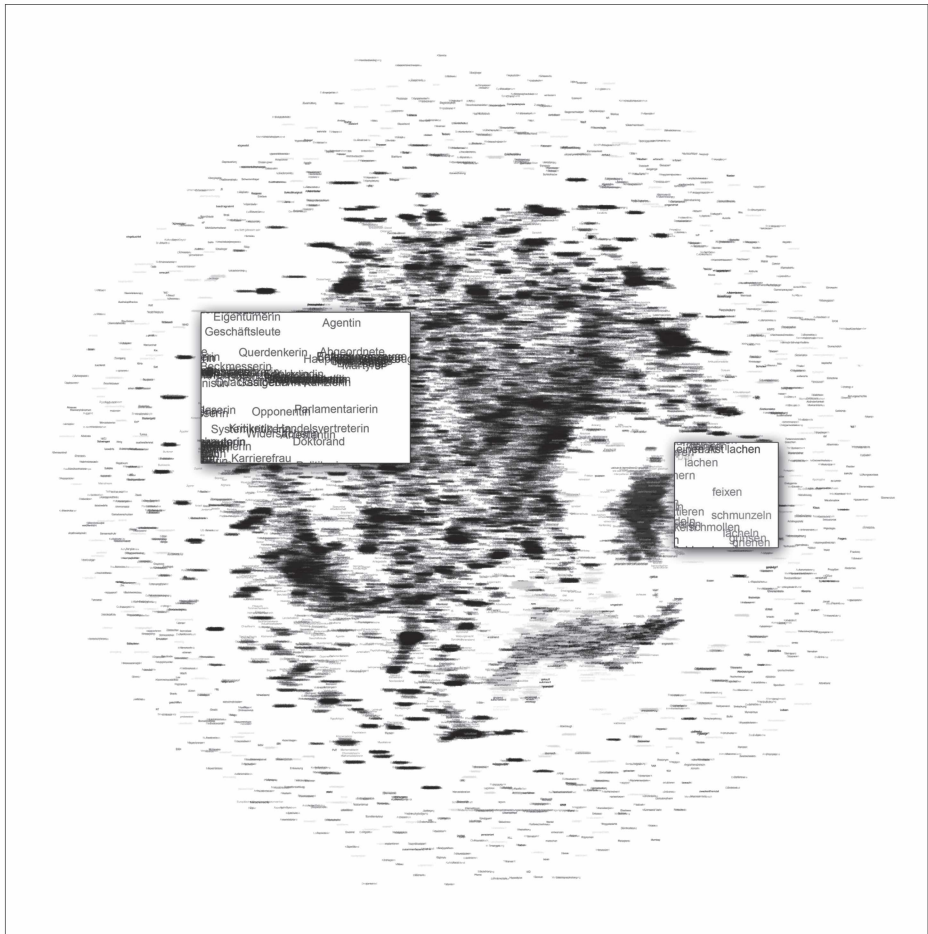


Abb. 3: Die paradigmatische Vernetzung im Wiktionary als Gesamtgraph.

Auch quantitative Verteilungen lassen sich aus diesem Gesamtgraphen ablesen, wie die Verteilung der einzelnen Angabegruppen und die Menge an Stichwörtern verteilt auf die verschiedenen Wortartengruppen (vgl. Tab. 1 und 2). Die Angabeklasse mit der höchsten Anzahl an Kanten innerhalb der paradigmatischen Vernetzungen ist die der Synonyme, gefolgt von den Oberbegriffen. Nur etwa halb so oft sind Vernetzungen in den Angabeklassen der Unterbegriffe und Gegenwörter im Wiktionary angesetzt. Bei der überwiegenden Zahl von Stichwörtern, zu denen paradigmatische Relationen im Wiktionary angegeben wurden, handelt es sich um Substantive.

Tab. 1: Menge und Anteil der Verbindungen (Kanten), aufgeteilt nach Angabeklassen.

<i>Angabeklasse</i>	<i>Anzahl Verbindungen(=Kanten)/ Angabeklasse</i>	<i>Anteil Verbindungen/Angabeklasse an Gesamtvernetzungen</i>
Gegenwörter	31.398	15,4%
Oberbegriffe	57.622	28,2%
Sinnverwandte Wörter	20.021	9,81%
Synonyme	60.021	29,4%
Unterbegriffe	34.991	17,1%

Tab. 2: Menge und Anteil der Stichwörter nach Wortarten, zu denen paradigmatische Relationen im Wiktionary angegeben sind.

<i>Wortart</i>	<i>Anzahl Artikel/Wortart mit paradigmatischen Angaben</i>	<i>Anteil Artikel/Wortart mit paradigmatischen Angaben</i>
Substantiv	42.022	75,30%
Adjektiv	5.258	9,42%
Verb	4.254	7,63%
Sonstige	3.344	5,99%
Abkürzung	886	1,59%
Partizip	24	0,04%

Der Zusammenhang zwischen den Wortarten und dem jeweiligen Anteil an Wortartikeln, die über eine bestimmte paradigmatische Angabeklasse verfügen, wird genauer in Abbildung 4 gezeigt. Hier ist zu erkennen, dass z. B. der Anteil an Synonymen sowohl bei Substantiven (40,6 %), Verben (56,9 %), Adjektiven (44,6 %) und Sonstigen (40,2 %) ⁵ hoch ist, wohingegen der Anteil an Einträgen, in denen Oberbegriffe angegeben werden, bei Substantiven mehr als doppelt so groß ist (60,5 %) wie in den anderen Wortarten. Dies ist sprachwissenschaftlich gesehen keine Überraschung, aber in der quantitativen Verteilung ohne Analyse der gesamten Datenbasis sonst nicht so explizit zu spezifizieren.

⁵ In der Gruppe „Sonstiges“ sind v. a. Adverbien, Interjektionen, Konjunktionen, Präpositionen, Pronomen, Redewendungen und Wortverbindungen zusammengefasst. Insgesamt handelt es sich um 52 ‚Wortarten‘, in denen aber auch eher merkwürdige Kategorien wie ‚Merksspruch‘, ‚Formativ‘ oder ‚Geflügeltes Wort‘ fallen; Kategorien, deren Vielfalt sicher in der kollaborativen Erstellungsweise des Wiktionary begründet liegt.

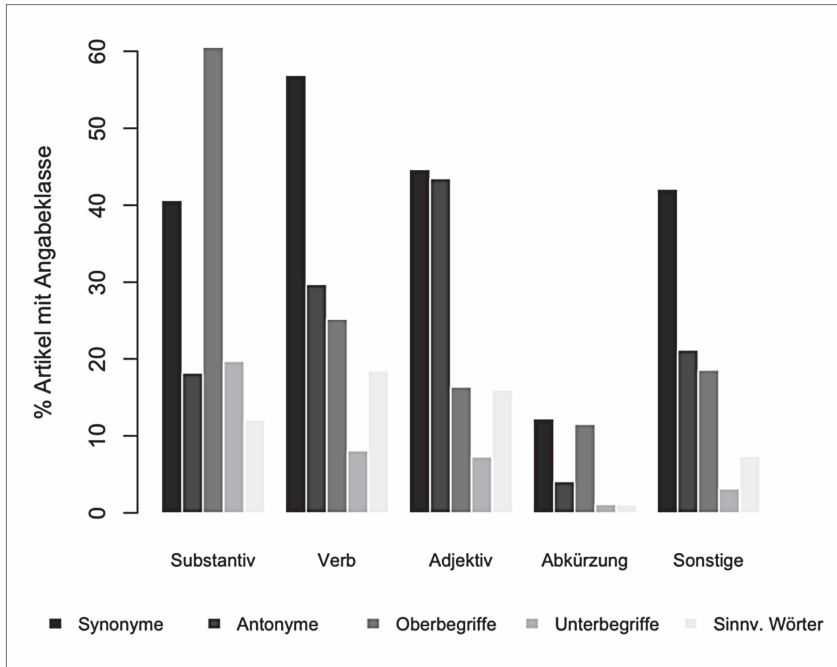


Abb. 4: Anteil der Wortartikel mit paradigmatischen Angabeklassen geordnet nach Wortarten.

Auf diese Weise lassen sich – anders als bei der exemplarischen Analysen der Wortartikel eines gedruckten Wörterbuchs – genaue, nachmessbare Aussagen zur Vernetzung in einem Wörterbuch treffen. Dies betrifft nicht nur den prozentualen Anteil an Artikeln mit Angabeklassen, sondern beispielsweise auch die Anzahl der in allen fünf relevanten Angabeklassen aufgeführten Relationspartner. An der Aufstellung in Tabelle 3 lässt sich beispielsweise erkennen, dass zwar die entsprechenden Angabeklassen bei Substantiven im Schnitt häufiger vertreten sind, diese aber anscheinend bei Verben mit mehr Angaben befüllt sind.

Tab. 3: Durchschnittliche Anzahl der im Wortartikel angegebenen paradigmatischen Relationspartner, getrennt nach Wortarten.

Abkürzung	Adjektiv	Partizip	Sonstige	Substantiv	Verb
0.48	3.00	0.046	1.43	2.70	3.28

Zusammenfassend lässt sich festhalten, dass ca. 25% des gesamten Stichwortbestandes im deutschen Wiktionary paradigmatisch vernetzt ist, dass diese Vernetzungen v. a. unter Stichwörtern gleicher Wortart angelegt sind, dass in der überwiegenden Mehrheit der Fälle paradigmatische Angaben in Wortartikeln zu Substantiven

gegeben werden, wobei jedoch bei Verben die durchschnittliche Anzahl der angegebenen Relationspartner höher ist als bei Substantiven. Nach diesem Gesamtüberblick sollen im folgenden Abschnitt Analysemethoden vorgestellt werden, die dazu dienen, besonders stark vernetzte Stichwortgruppen innerhalb dieser Gesamtmenge vernetzungsrelevanter Daten zu detektieren.

3 Detektion besonders stark vernetzter Stichwortgruppen mit Mitteln der Graphentheorie

Bei paradigmatischen Vernetzungen im Wörterbuch (oder auch anderer Vernetzungsarten) ist es interessant zu untersuchen, ob bestimmte Stichwortgruppen besonders stark vernetzt sind und wie eine Verteilung der ‚Vernetzungsgrade‘ insgesamt aussieht. Mit einem exemplarischen Zugang können solche Gruppen allerdings nur mehr oder weniger zufällig entdeckt werden. Auf Basis einer Analyse aller vernetzungsrelevanten Angaben kann man sich diesen Fragen dagegen mit Mitteln aus dem Bereich der Graphentheorie nähern.

Interessant sind in diesem Zusammenhang die Konzepte von Cliques und Clustern, da es diese Konzepte ermöglichen, aus der Menge der gesamten Vernetzungen Gruppen von Stichwörtern zu entdecken, die besonders stark vernetzt sind, ohne ein bestimmtes Stichwort als Ausgangspunkt auszuwählen, z. B. Gruppen von Synonymen. Dieser Analyseansatz ermöglicht es damit – um die ‚Weltkarte‘ des Gesamtgraphen wieder als Bild aufzugreifen – aus dieser Welt datengetrieben kleinere Ausschnitte mit bestimmten vernetzungsrelevanten Eigenschaften zu identifizieren. Eine Clique bezeichnet dabei in einem Graphen eine Gruppe von Knoten, innerhalb derer jeder Knoten mit jedem anderen Knoten verbunden ist. Die Richtung der Verbindung spielt dabei keine Rolle.⁶ Cliques sind somit ein recht strenges Konzept, da schon beim Wegfall einer Verbindung zu einem Wort aus der Clique ein potentielles weiteres Mitglied ausgeschlossen würde. Cluster innerhalb von Graphen sind ein weniger strenges Konzept. Auch schwächer verbundene Untergruppen von Graphen können als Cluster identifiziert werden, wenn einzelne Mitglieds-knoten (Wortartikel) über Verbindungen (Vernetzungen) erreichbar sind. Dementsprechend ist ein Wortartikel nicht Mitglied eines bestimmten Clusters, wenn es nicht erreicht werden kann. Ein Cluster ist also eine Gruppe von Knoten, die untereinander gut erreicht werden können. Im Folgenden sollen Übersichten über die Anzahl und Größe von Cliques und Clustern im Wiktionary, sowie einzelne Beispiele der größten Cliques und Clustern gezeigt werden. An dieser Stelle beschränken wir uns auf den Synonymgraphen, um die prinzipielle Verwendbarkeit der graphentheoretischen Konzepte in lexikogra-

⁶ Vgl. <http://de.wikipedia.org/wiki/Clique_%28Graphentheorie%29> (zuletzt eingesehen am 1.4.2015).

fischem Rahmen zu testen, d. h. wir zeigen Gruppen von Stichwörtern, die im Wiktionary über eine Synonymangabe als ein- oder ausgehende Vernetzung verbunden sind.

3.1 Synonym-Cliquen

Abbildung 5 zeigt einen Gesamtüberblick über die ermittelten Cliques mit vier oder mehr Mitgliedern. Die größte Synonym-Clique im deutschen Wiktionary enthält elf Stichwörter, d. h. in allen diesen elf Stichwörtern sind alle anderen zehn Stichwörter als Synonyme verzeichnet oder als synonymisches Verweisziel angegeben. Von diesen großen Cliques gibt es allerdings nur wenige: nur neun Cliques verfügen über acht oder mehr Mitglieder.⁷

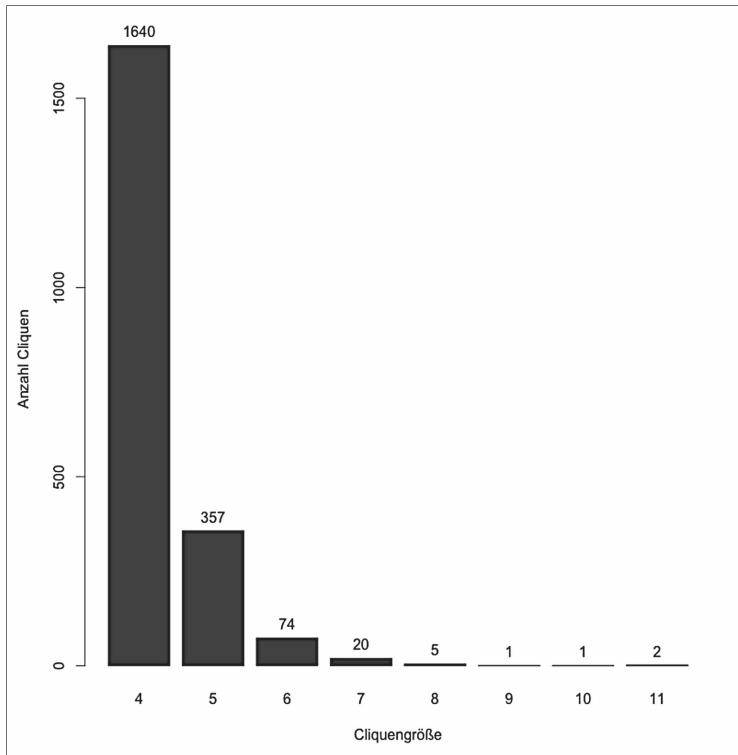


Abb. 5: Übersicht über Synonym-Cliquen im deutschen Wiktionary

⁷ Weitere Cliques und Cluster stehen unter <http://www.ids-mannheim.de/fileadmin/lexik/bilder/cliques.zip> und <http://www.ids-mannheim.de/fileadmin/lexik/bilder/cluster.zip> (zuletzt eingesehen am 1.4.2015) zur Verfügung.

Bei den beiden größten Cliques handelt es sich um eine Gruppe von Kausalkonnektoren (Abb. 6) und um eine Clique rund um das Stichwort „Toilette“ (Abb. 7). Ein Blick in die Versionsgeschichte der Einträge zeigt, dass viele der Synonymangaben bei den Kausalkonnektoren von einem Benutzer namens ‚Tekla‘ am 30.1.2006 bearbeitet wurden (z. B. „ergo“, „mithin“, „deshalb“ und „demnach“). Ähnlich wurden bei der Clique rund um „Toilette“ viele Synonymangaben am 1.4.2013 durch ‚Elleff Groom‘ ergänzt (z. B. bei „Pott“, „Häusl“, „Scheißhaus“, „Null-Null“, „Klosett“), wohingegen Einträge wie „Tö“ und „Örtchen“ nicht von ‚Elleff Groom‘ bearbeitet wurden und vielleicht aus diesem Grund nur als Verweisziel und nicht als Ausgangspunkt eines Synonymverweises angegeben sind. Man kann also vermuten, dass eine solche einheitliche synonymische Cliquenbildung in einem kollaborativen Wörterbuch wie dem Wiktionary v. a. dann gelingt, wenn ein Bearbeiter oder eine Bearbeiterin viele der in Frage kommenden Stichwörter bearbeitet.

Ähnliches gilt auch für die kleinere, aber besonders schöne „Memme“-Clique (Abb. 8), deren Synonymangaben auch v. a. durch ‚Elleff Groom‘ am 16.7.2011 ergänzt wurden.

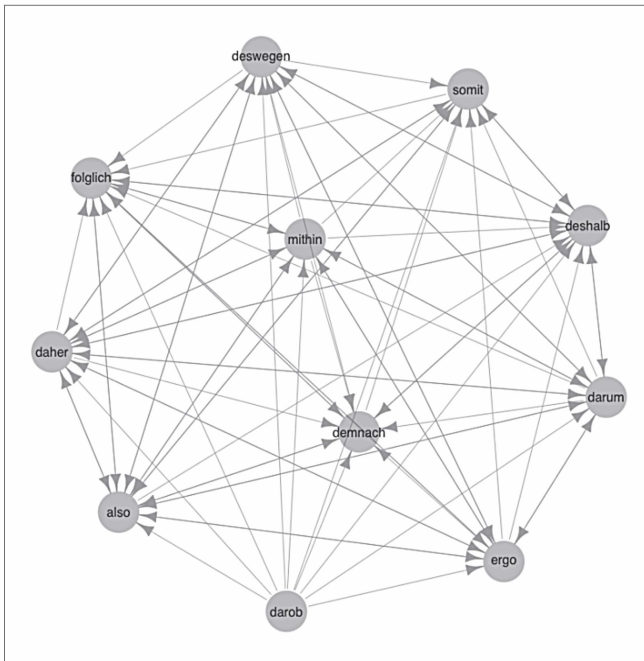


Abb. 6: Clique „deswegen“

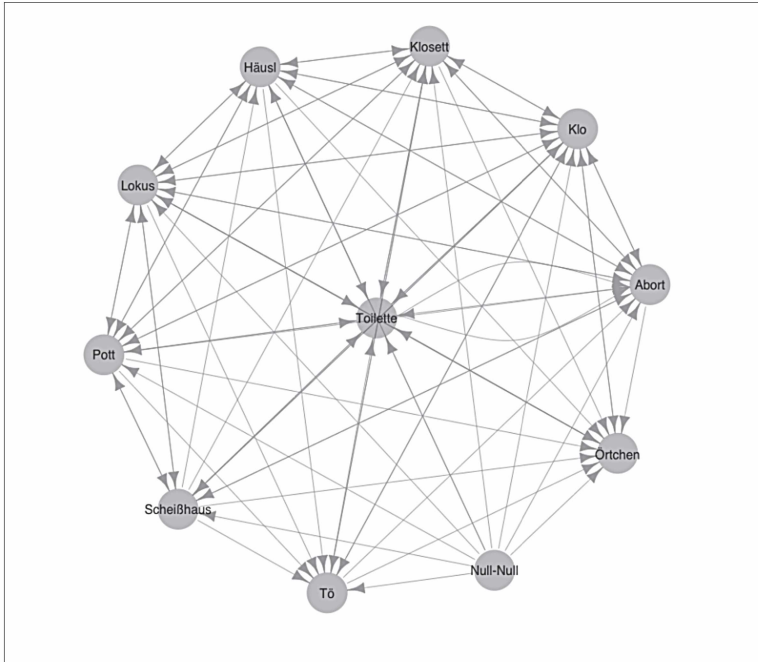


Abb. 7: Clique „Toilette“

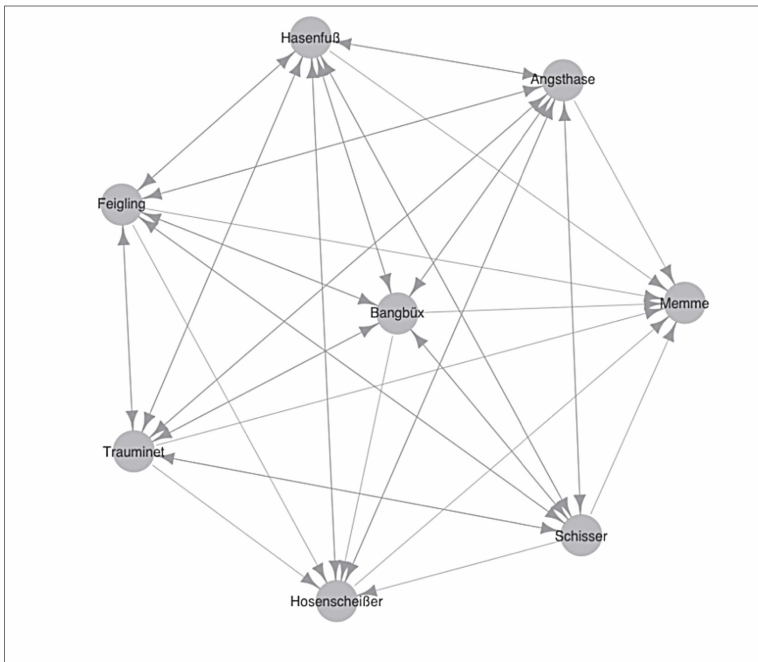


Abb. 8: Clique „Memme“

Das Konzept der Clique ermöglicht es also, sehr systematisch vernetzte Wortartikel zu detektieren. Zum Entdecken synonymisch eng vernetzter Stichwortgruppen kann außerdem das Konzept der Cluster ergänzend hinzugezogen werden, da die dort ermittelten Gruppen nach etwas weniger strengen Kriterien gewonnen werden und so weitere synonymische Gruppen in den Blick kommen können.

3.2 Synonym-Cluster

Abbildung 9 zeigt einen Gesamtüberblick über die von uns ermittelten Synonym-Cluster im deutschen Wiktionary. Es ist auffällig, dass die Anzahl der Cluster im Bereich von vier bis sechs Mitgliedern weniger stark abfällt als dies für die Cliques der Fall ist. Dies ist höchstwahrscheinlich dem weniger strengen Konzept des Clusters im Vergleich zur Clique geschuldet. Allerdings sind sehr große Cluster ebenfalls selten; nur fünf ermittelte Cluster haben zehn oder mehr Mitglieder. Das größte ermittelte Cluster gruppiert 14 Stichwörter rund um die „Besenwirtschaft“ (Abb. 10), das zweite hier gezeigte Beispiel ist das „Ressentiment“-Cluster mit elf Mitgliedern (Abb. 11).

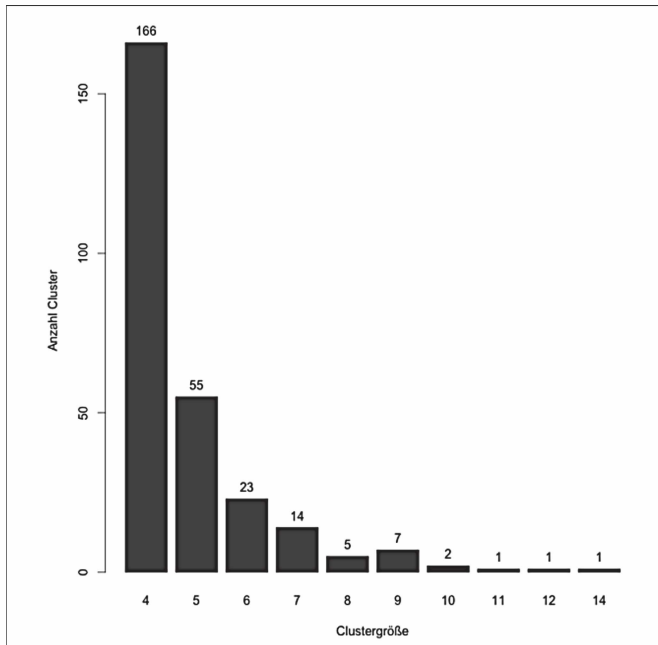


Abb. 9: Übersicht über Synonym-Cluster im deutschen Wiktionary

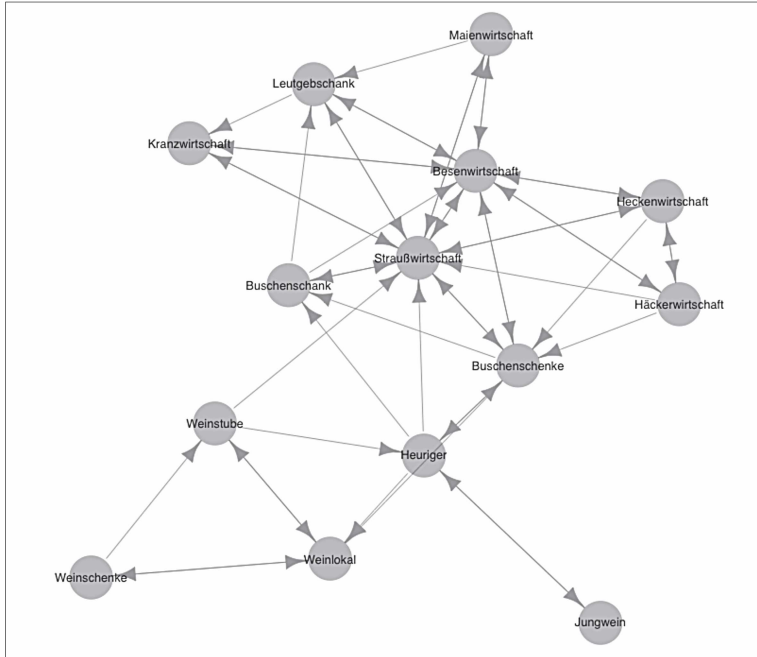


Abb. 10: Cluster „Besenwirtschaft“

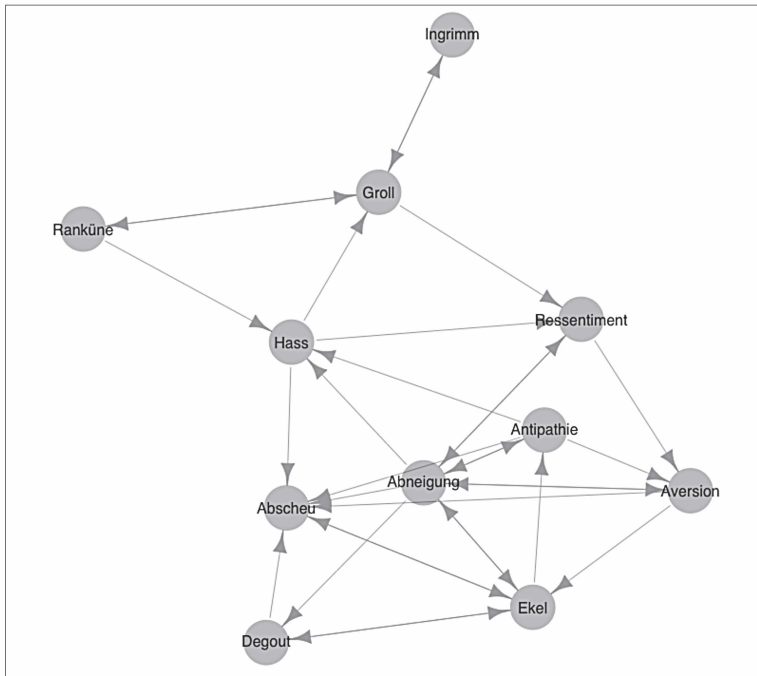


Abb. 11: Cluster „Ressentiment“

Auch bei den Clustern zeigt sich, dass zumindest mehrere Mitglieder einer Gruppe von einem einzelnen Wiktionary-Autor bearbeitet wurden, z. B. im Cluster „Besenwirtschaft von ‚L.S.‘ (‚Buschenschenke“, „Heuriger“, „Häckerwirtschaft“, alles am 19.9.2005) oder von ‚Nasobema Lyricum‘ (‚Straußwirtschaft“, „Maienwirtschaft“, „Heckenwirtschaft“, am 16.3.2013 bzw. 29.3.2014). Da die Zugehörigkeit zu einem Cluster aber eben nicht die vollständige Vernetzung aller Mitglieder der Gruppe voraussetzt, ist dieser Effekt nicht so stark ausgeprägt wie bei den Cliques.

Die Beispiele zeigen unserer Meinung nach insgesamt sehr deutlich, dass es sowohl mit dem Konzept der Clique, aber auch mit dem etwas weicheren Konzept der Cluster gelingt, große bedeutungsverwandte Gruppen von Stichwörtern sinnvoll zu detektieren. Voraussetzung ist wie anfangs gesagt, die gesamte Datenbasis mit allen vernetzungsrelevanten Angaben zur Verfügung zu haben. Neben der reinen Analyse der Vernetzungen im Wörterbuch könnte dieser Analyseansatz auch bei der Überarbeitung eines Wörterbuchs verwendet werden, um Gruppen von Stichwörtern zu entdecken, deren Einbindung in den Gesamtgraphen verbessert werden könnte.

In diesem Abschnitt haben wir gezeigt, wie man mit dem Konzept der Cliques und Cluster sinnvoll eng bedeutungsverwandte Gruppen innerhalb eines Stichwortbestandes entdecken kann. Ähnliche Verfahren könnten beispielsweise auch bei Angaben zur Wortbildung Anwendung finden, um (morphologische) Wortfamilien zu detektieren. Im letzten Teil dieses Beitrags soll der Blick nun über Vernetzung im engeren Sinne hinausgehen, indem weitere Daten in die Analysen einbezogen werden.

4 Vernetztheit – Korpusfrequenz – Nachschlagehäufigkeit

In diesem Abschnitt zeigen wir, wie man Analysen zur Vernetztheit von Stichwörtern in einem digitalen Wörterbuch zu anderen (Meta-)Daten wie Korpusfrequenzen und Nachschlagehäufigkeiten in Beziehung setzen kann. Drei Fragen wollen wir dabei beantworten: Sind frequente Wörter im Korpus im Wörterbuch paradigmatisch stärker vernetzt? (4.1) Werden paradigmatisch stark vernetzte Stichwörter häufiger nachgeschlagen? (4.2) Können Daten zur Nachschlagehäufigkeit sinnvoll in das Konzept der paradigmatischen Cliques und Cluster eingebracht werden? (4.3) Die Korpusfrequenzen, die wir verwenden, beruhen auf den Frequenzen der DeReKo-Wortformenliste (Kupietz et al. 2010). Die Nachschlagehäufigkeiten sind berechnet aus den Logfiles zum deutschsprachigen Wiktionary des gesamten Jahres 2014⁸. Die von der Wikimedia Foundation stündlich zur Verfügung gestellten Logfiles wurden auf das gesamte

⁸ Diese Daten sind verfügbar unter <https://dumps.wikimedia.org/other/pagecounts-raw/2014/> (zuletzt eingesehen am 1.4.2015).

Jahr 2014 aggregiert. Wir berichten normierte Zugriffswerte, d. h. die Anzahl an Zugriffen auf eine bestimmte Seite in einer Million Zugriffe.

4.1 Sind frequente Wörter paradigmatisch stärker vernetzt?

Betrachtet man alle Stichwörter des deutschen Wiktionary, sind, wie anfangs herausgestellt, die überwiegende Mehrzahl der Knoten (75%) gar nicht über eine paradigmatische Vernetzung in den Gesamtgraphen eingebunden. Übrig bleiben damit als Grundlage für die Analyse etwa 61.000 Einträge, die hier wieder nach den sechs Wortartenkategorien (Substantive, Verben, Adjektive, Partizipien, Sonstige, Abkürzungen) getrennt analysiert werden. Um die anfangs gestellte Frage beantworten zu können, muss nun die Korpusfrequenz einer Stichwortgruppe zu ihrer Vernetztheit in Beziehung gesetzt werden. Diese beiden Variablen werden folgendermaßen berechnet: Für die Bestimmung des Vernetzungsgrades werden zunächst alle Stichwörter beachtet, die mindestens eine ein- oder ausgehende Verbindung im paradigmatischen Vernetzungsgraphen aufweisen. Gezählt werden nun alle ein- und ausgehende Verbindungen. Eine eingehende Verbindung besteht dann, wenn ein Wort bei einem anderen Wort als Verweisziel genannt wird. Einen Vernetzungsgrad von „5“ erhält man für ein Stichwort demnach beispielsweise dann, wenn in einem Artikel unter der Angabegruppe „Synonyme“ zwei Synonymangaben stehen, und das infrage kommende Stichwort in drei anderen Artikeln als Synonym, und damit als Verweisziel, aufgeführt ist. Die Korpusfrequenzen werden hier in Frequenzdezilen angegeben, die auf logarithmierten Frequenzen beruhen (vgl. auch Müller-Spitzer et al. 2015, 14–19). Wir verwenden Frequenzdezile, weil auf diese Weise das komplette Frequenzspektrum so in zehn Gruppen aufgeteilt wird, dass in jede Gruppe gleich viele Wortartikel fallen. Die Frequenzgruppen, die sich so ergeben, sind somit gleich groß und können einfacher untereinander verglichen werden.

Abbildung 12 zeigt eine Übersicht zum Zusammenhang von Vernetzungsgrad und Korpusfrequenz. Der Plot zeigt vor allem bei Substantiven, Adjektiven, Verben und Sonstigen, dass die häufigeren Wörter besser in den Restgraphen integriert sind, d. h. je häufiger ein Wort ist, desto stärker ist der Vernetzungsgrad des Stichworts im paradigmatischen Vernetzungsgraphen des Wiktionary. Bei dieser Analyse zeigt sich außerdem, dass Substantive im höchsten Frequenzdezil nicht so hoch steigen wie Verben und Adjektive. Das liegt u. a. an vielen Eigennamen, die im Wiktionary häufig (auch) als Substantive bezeichnet werden. Es gibt 667 Substantive aus Frequenzdezil 10, die nicht mit dem Restgraphen verbunden sind. Beispiele (aus den ersten zehn im Alphabet) sind: „Aaron“, „Abel“, „Abbas“, „Abraham“, „Adelheid“, „Adenauer“, „Adi“ und „Adolf“. Es gibt dagegen nur 24 Verben und nur 37 Adjektive aus dem 10er-Frequenzdezil ohne Verbindung zum Restgraphen. Man kann davon ausgehen, dass eine solche Analyse für ein von lexikographischen Expertinnen und Experten erarbeitetes Wörterbuch anders aussehen würde, da in der Regel Eigenna-

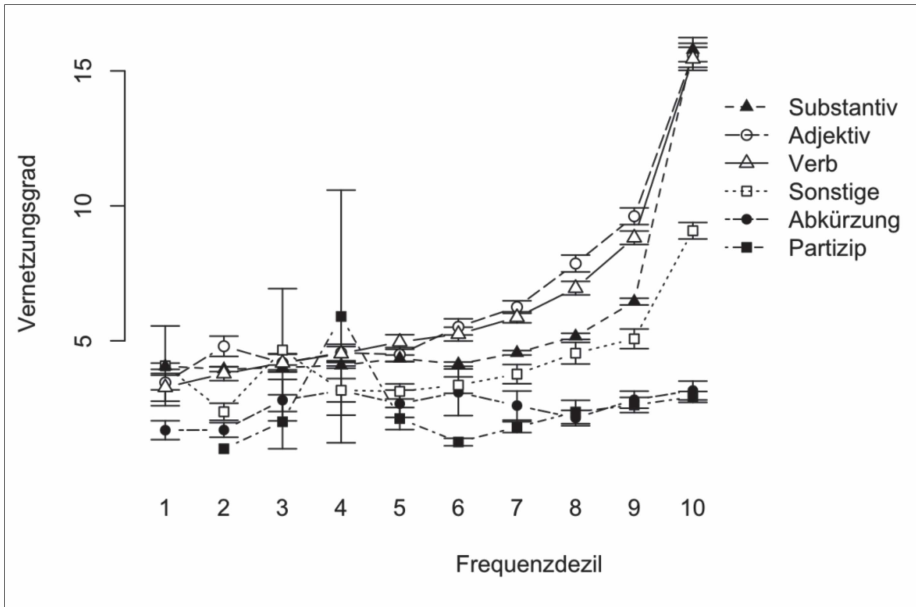


Abb. 12: Zusammenhang von paradigmatischem Vernetzungsgrad und Korpusfrequenz Stichwortgruppen/Wortart.

men in einem allgemeinsprachlichen Wörterbuch nur lemmatisiert werden, wenn sie als Gattungsbezeichnungen fungieren. Die Kurve zu Substantiven würde damit v. a. in den hohen Frequenzdezilen vermutlich anders aussehen, nämlich stärker ansteigen.

Zusammenfassend lässt sich festhalten: Häufige Wörter im Korpus sind im deutschsprachigen Wiktionary paradigmatisch stärker vernetzt. Im Folgenden wollen wir noch Informationen über die Nachschlagehäufigkeiten in die Analysen einbeziehen und die Frage beantworten, ob paradigmatisch stark vernetzte Stichwörter auch häufiger nachgeschlagen werden.

4.2 Werden paradigmatisch stark vernetzte Stichwörter häufiger nachgeschlagen?

Die Frage, ob paradigmatisch stark vernetzte Stichwörter häufiger nachgeschlagen werden, ist nicht einfach zu beantworten, denn die im vorigen Abschnitt gezeigten Analysen haben gezeigt, dass Wörter mit einer höheren Korpusfrequenz besser vernetzt sind. Wir wissen zusätzlich aus anderen Untersuchungen (vgl. Kopenig et al. 2014), dass Wörter mit höherer Korpusfrequenz häufiger nachgeschlagen werden. Daher können wir zur Beantwortung dieser Frage nicht einfach den paradigmatischen Vernetzungsgrad und die Nachschlagehäufigkeit eines Stichworts miteinander in Beziehung setzen, da dann eigentlich nur der Korpusfrequenzeffekt wiederholt gemessen werden würde. Um dies zu umgehen, setzen wir in der folgenden

Analyse die Variablen Korpusfrequenz, Vernetzungsgrad und Nachschlagehäufigkeit gruppenbasiert in Beziehung. Auf diese Weise verschränken wir den Vernetzungsgrad und die Korpusfrequenz von Wörtern ineinander, sodass der Einfluss der Korpusfrequenz auch gesondert für weniger oder stärker vernetzte Wörter exploriert werden kann. Andersherum kann auch der Einfluss des Vernetzungsgrads gesondert für seltenere und häufigere Wörter analysiert werden.

Die Stichwörter sind nach ihrem Vernetzungsgrad jeweils so zusammengefasst, dass die entstehenden Gruppen hinsichtlich ihrer Größe nicht zu stark divergieren (vgl. Tab. 4) Wichtig ist wiederum zu beachten, dass die flektierten Formen, die im Wiktionary in großer Zahl als eigene Stichwörter angesetzt sind, von der Analyse ausgeschlossen wurden, d. h. in der ‚unverbunden‘-Kategorie nicht auftauchen.

Tab. 4: Anzahl von Stichwörtern je Vernetzungsgrad.

nicht verbunden:	24.860
1 Kante	11.191
2 Kanten	11.862
3–6 Kanten	21.484
> 6 Kanten	16.769

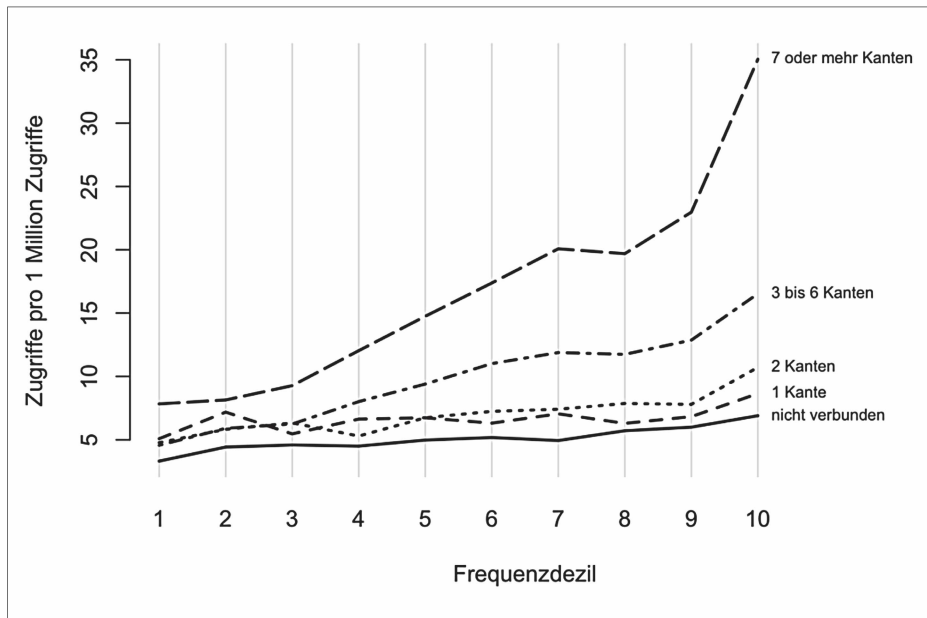


Abb. 13: Zusammenhang zwischen Nachschlagehäufigkeit, Korpusfrequenz und paradigmatischer Vernetztheit.

Abbildung 13 zeigt den Zusammenhang von Nachschlagehäufigkeit, Korpusfrequenz und paradigmatischer Vernetztheit. Es ist deutlich zu erkennen, dass insbesondere paradigmatisch stark vernetzte Stichwörter, die *auch* häufig im Korpus vorkommen, besonders häufig nachgeschlagen werden. Eine starke paradigmatische Vernetzung allein führt, wie an den geringen Unterschieden zwischen den Linien in den Frequenzdeziolen eins und zwei deutlich zu erkennen, nicht zu einer erhöhten Nachschlagehäufigkeit. Sehr klar ist aber auch abzulesen, dass Wörter, die solitär im Wörterbuch stehen, zumindest im Durchschnitt nicht häufig nachgeschlagen werden, egal ob sie im Korpus frequent sind oder nicht. Dies ist in Abbildung 13 an dem vergleichsweise schwachen Anstieg der durchgezogenen Linie abzulesen. Der Anstieg der obersten gestrichelten Linie hingegen (stark vernetzte Wörter mit 7 oder mehr Kanten) ist sehr deutlich. Der folgende Abschnitt greift noch einmal das in Abschnitt 3 dargestellte Konzept der Cliques und Cluster auf und verbindet diese Analysen mit Daten über die Nachschlagehäufigkeit der in ihnen enthaltenen Stichwörter.

4.3 Nachschlagehäufigkeit – Cliques – Cluster

In diesem letzten Abschnitt wollen wir zeigen, wie man ein weiteres Maß auf die in Abschnitt drei gezeigten Cliques und Cluster anwenden kann. Dabei interessiert uns die Frage, welche Knoten in den Cliques und Clustern besonders häufig nachgeschlagen werden. Anhand zweier Einzelbeispiele, der „ergo“-Clique und dem „Ressentiment“-Cluster⁹, wird außerdem untersucht, ob die in diesen Gruppen besonders häufig nachgeschlagenen Knoten auch besonders häufig im Korpus sind. Damit gehen wir der Frage nach, ob diese zwei Einzelbefunde dem in Abschnitt 4.2 erwähnten allgemeinen Trend entsprechen, dass häufig nachgeschlagene Stichwörter auch häufig im Korpus sind.

In den Abbildungen 14 und 15 wird gezeigt, wie Informationen über die Nachschlagehäufigkeit visuell in die Clique- und Clusterdarstellung integriert werden können: Die Größe der Knoten verdeutlicht die Anzahl der Zugriffe im Jahr 2014. Es ist allerdings zu beachten, dass die Größen der Knoten in den Abbildungen 14 und 15 untereinander nicht vergleichbar sind, da die Kausalkonnektoren, insbesondere „ergo“, im Vergleich so häufig nachgeschlagen wurde, dass bei gleicher Skalierung wie in Abbildung 14 in Abbildung 15 nichts mehr zu erkennen gewesen wäre.

⁹ Andere Cliques und Cluster mit den Nachschlagehäufigkeiten sind in komprimierter Form unter <http://www.ids-mannheim.de/fileadmin/lexik/bilder/cliques-visits.zip> und <http://www.ids-mannheim.de/fileadmin/lexik/bilder/cluster-visits.zip> zu finden.

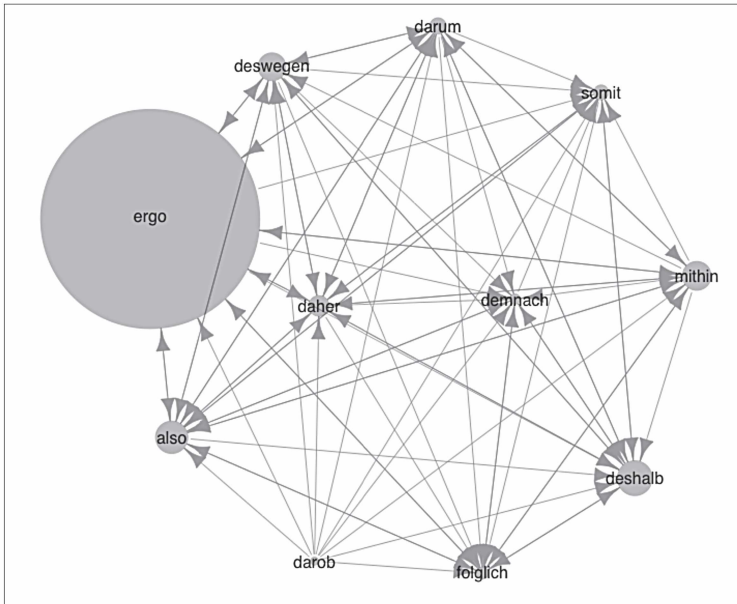


Abb. 14: „ergo“-Clique mit Kennzeichnung der Nachschlagehäufigkeit (Größe der Kreise verdeutlicht die Nachschlagehäufigkeit im Jahr 2014).

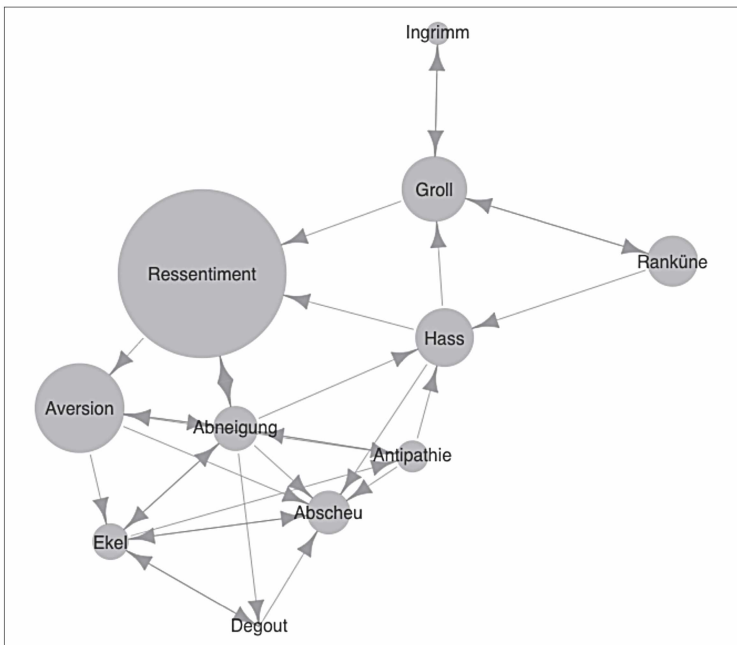


Abb. 15: „Ressentiment“-Cluster mit Kennzeichnung der Nachschlagehäufigkeit (Größe der Kreise verdeutlicht die Nachschlagehäufigkeit im Jahr 2014).

Es ist in diesen Darstellungen klar zu erkennen, dass jeweils ein Stichwort besonders häufig nachgeschlagen wird: In der ersten Gruppe das Stichwort „ergo“, in der zweiten „Ressentiment“. Ob diese Stichwörter auch diejenigen sind, die in den jeweiligen Gruppen besonders häufig im Korpus vorkommen, zeigen die Tabellen 5 und 6.

Tab. 5: Korpusfrequenzen von Mitgliedern der „ergo“-Clique.

<i>Stichwort</i>	<i>Rohfrequenz</i>	<i>Frequenzdezil</i>	<i>Frequenzrang</i>
also	1.148.545	10	245
deshalb	807.498	10	351
daher	491.402	10	606
darum	338.752	10	883
somit	234.359	10	1.331
deswegen	115.970	10	2.686
demnach	93.245	10	3.272
folglich	24.112	10	9.752
mithin	17.022	10	12.301
darob	3.285	9	31.469
ergo	2.523	9	35.653

Tab. 6: Korpusfrequenzen von Mitgliedern des „Ressentiment“-Clusters.

<i>Stichwort</i>	<i>Rohfrequenz</i>	<i>Frequenzdezil</i>	<i>Frequenzrang</i>
Hass	25.033	10	9.517
Abneigung	11.451	10	15.836
Groll	7.951	9	19.623
Ekel	7.880	9	19.719
Abscheu	6.981	9	21.184
Ressentiment	2.047	8	39.202
Antipathie	1.571	8	43.976
Aversion	1.519	8	44.557
Ingrimm	809	7	57.327
Ranküne	347	7	75.516
Degout	91	5	105.283

In den Tabellen ist zu sehen, dass „ergo“ die niedrigste Rohfrequenz (2523) und den niedrigsten Rang (35.653) innerhalb dieser Clique von Stichwörtern im deutschen Wiktionary hat; in einem ähnlichen Frequenzbereich liegt nur noch „darob“. Das

Stichwort „Ressentiment“ liegt bezüglich der Korpushäufigkeit in der Mitte (an sechster Stelle von elf Mitgliedern) und ist im Korpus deutlich seltener als „Hass“ oder „Abneigung“. Man kann vermuten, dass sowohl die Eigenschaft als Fremdwort wie auch die dementsprechend schwierigere Schreibung zu einem häufigen Nachschlagen beider Stichwörter geführt hat. Diese zwei Fälle sind damit auch schöne Beispiele für die Abweichung von Einzelfällen gegenüber dem Durchschnitt. Zwar werden im Durchschnitt häufigere Wörter im Korpus auch häufiger nachgeschlagen, in den beiden gezeigten Beispielen verhält es sich jedoch anders.

5 Ausblick

Wir haben in diesem Beitrag aufgezeigt, wie man Teile der Vernetzungsstruktur eines digitalen Wörterbuchs auf Basis einer Gesamtanalyse aller relevanten Daten beschreiben kann und welche Formen zusätzlicher Analysen unter Einbeziehung weiterer (Meta-) Datentypen möglich sind. Unser Vorgehen unterscheidet sich damit prinzipiell von der Analyse von Mediostrukturen in gedruckten Wörterbüchern, bei der Stück für Stück einzelne verweisrelevante Angaben wie ein Puzzle zusammengesetzt und in eine strukturelle Ordnung gebracht werden. Abbildung 16 illustriert dieses unterschiedliche Vorgehen: Bei der Analyse eines gedruckten Wörterbuchs ist der Ausgangspunkt eine Stelle im Buch, die zu einer weiteren führt, von der wieder auf eine andere verwiesen wird usw.; unser Ausgangspunkt ist dagegen eine Draufsicht, eine Art Weltkarte vernetzungsrelevanter Angaben, von der aus wir auf einzelne Teilstrukturen schauen können.

Wir denken, dass diese Art des Vorgehens vielfältiges Potential für die Arbeit mit Vernetzungsstrukturen digitaler Wörterbücher hat. In der mediostrukturellen Analyse verbessert der neue Ansatz die Quantifizierbarkeit von Aussagen wie:

„Wesentlich häufiger erfolgen Wortfeldverweise im Rahmen von bauteilexternen Verweisen auf den Nachspannabschnitt *Wortfelder* [...]“ (Müller 2002: 492)

Auf Basis einer Gesamtanalyse vernetzungsrelevanter Daten lässt sich genau messen, welche Art von Angaben wie oft und wo vorkommt. Entsprechend haben wir für die paradigmatischen Angaben gezeigt, bei welchen Wortarten welche Art von Angabeklasse wie häufig in Wortartikeln enthalten ist und mit wie vielen Angaben diese Klassen im Durchschnitt gefüllt sind. Man kann noch weiter denken: Auch für die Entwicklung von Zugriffsstrukturen in digitalen Wörterbüchern könnte das Konzept der Cliques und Cluster genutzt werden. Auf diese Weise könnte man paradigmatisch eng vernetzte Stichwortgruppen detektieren, die dann Nutzerinnen und Nutzer in einem gruppenbasierten Zugriff angeboten werden könnten. Bei einem Wörterbuch mit kleinem Stichwortumfang ließe sich darüber hinaus mit der Darstellung des Gesamtgraphen eine neue Sicht auf das Wörterbuch entwickeln. Erste Tests mit dem

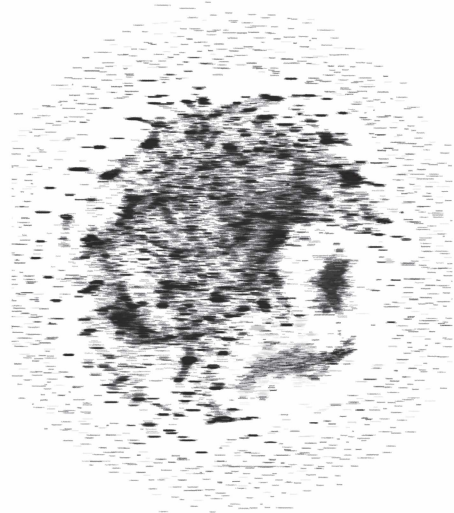


Abb. 16: Jonathan Wolstenholme: Cross References, 2003 (links)¹¹, Gesamtgraph der paradigmatischen Vernetzungen des Wiktionary (rechts)

Diskurswörterbuch zum „Schulddiskurs 1945–55“¹⁰, einem Wörterbuch mit unter 100 Hauptstichwörtern, erscheinen in dieser Hinsicht vielversprechend.

So ergibt sich beispielsweise auch über eine graphentheoretische Herangehensweise, dass der Wortartikel „Schuld“ der mit Abstand am besten vernetzte Artikel im Wörterbuch ist und vom Visualisierungsalgorithmus an zentraler Stelle verortet wird. Auf diese Weise wird die zentrale Stellung des Stichworts „Schuld“ auch auf Basis der ein- und ausgehenden Vernetzungen deutlich.¹² Auch im lexikografischen Prozess eines im Aufbau befindlichen Wörterbuchs oder bei einer Überarbeitung könnten solche Analysen gewinnbringend eingesetzt werden. Das Konzept der Cliques und Cluster könnte man beispielsweise nicht nur nutzen, um eng vernetzte Gruppen zu finden, sondern auch, um diese Gruppen kritisch dahingehend zu untersuchen, ob

10 Schulddiskurs 1945–55 (2015), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim, <<http://www.owid.de/wb/disk45/einleitung.html>> (zuletzt eingesehen am 14.2015).

11 Den Hinweis auf dieses Gemälde verdanken wir dem Vortrag von Helmut Feilke auf der IDS-Jahrestagung 2015 (<http://www.ids-mannheim.de/org/tagungen/program2015.html>) S. auch <<https://www.pinterest.com/pin/502151427172528612/>> und <<http://coffeenuts.tumblr.com/post/53848686855/painting-by-jonathan-wolstenholme>> (zuletzt eingesehen am 1. April 2015).

12 Ein Artikel für die IDS-Zeitschrift *Sprachreport* zur Anwendung des hier vorgestellten Ansatzes auf das Diskurswörterbuch 1945–55 ist in Vorbereitung (Sascha Wolfer/Heidrun Kämper). Zum Schulddiskurs s. auch Kämper 2005.

bestimmte Relationspartner fehlen und Vernetzungen nachgetragen werden sollten. Auch nicht gut in den Vernetzungsgraphen integrierte Gruppen könnten auf diese Weise gefunden werden. In einem kollaborativen Wörterbuch wie dem Wiktionary könnte man auf diese Weise auch Stichwortgruppen finden, die im Wörterbuch noch fehlen.

Es wäre darüber hinaus interessant, wie bereits anfangs gesagt wurde, die hier dargestellte Form der mediostrukturellen Analyse von lexikographischen Daten auf ein von lexikographischen Expertinnen und Experten erarbeitetes Wörterbuch anzuwenden. Einige Daten im Wiktionary, wie die Lemmatisierung flektierter Formen oder Eigennamen, widersprechen der gängigen lexikografischen Praxis. Außerdem scheinen oft nur die Einträge systematisch miteinander vernetzt, die von einem und demselben Autor bearbeitet wurden. In einem über Jahre erarbeiteten und regelmäßig überarbeiteten Wörterbuch könnte man mit solchen Analysen vielleicht sogar einen anderen Blick auf Beziehungen in der Sprache gewinnen, natürlich nur soweit sie im Wörterbuch abgebildet sind. Wir denken jedenfalls, dass die Möglichkeiten, die dieser neue Ansatz der Analyse von Vernetzungsstrukturen in digitalen Wörterbüchern bietet, mit den hier gezeigten Analysen noch lange nicht ausgeschöpft ist.

6 Literatur

- Blumenthal, Andreas/Lemnitzer, Lothar/Storrer, Angelika (1988): Was ist eigentlich ein Verweis? Konzeptionelle Datenmodellierung als Voraussetzung computerunterstützter Verweisbehandlung. In: Harras, Gisela (Hg.): *Das Wörterbuch. Artikel und Verweisstrukturen. Jahrbuch 1987 des Instituts für deutsche Sprache (=Sprache der Gegenwart 74)*. Düsseldorf: Schwann, 351–373.
- Fuertes-Olivera, Pedro A. (2009): The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary. In: Bergenholtz, Henning/Nielsen, Sandro/Tarp, Sven (Hgg.): *Lexicography at a crossroads: dictionaries and encyclopedias today, lexicographical tools tomorrow. (= Linguistic Insights - Studies in Language and Communications)*. Bern et al.: Peter Lang, 99–134.
- Hanks, Patrick (2012): Corpus Evidence and Electronic Lexicography. In: Granger, Sylviane/Paquot, Magali (Hgg.): *Electronic lexicography*. Oxford: Oxford University Press, 57–82.
- Kammerer, Matthias (1998): Die Mediostruktur in Langenscheidts Großwörterbuch Deutsch als Fremdsprache. In: Wiegand, Herbert Ernst (Hg.): *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von „Langenscheidts Großwörterbuch Deutsch als Fremdsprache“ (= Lexicographica. Series Maior 86)*. Tübingen: Niemeyer, 315–330.
- Kämper, Heidrun (2005): Der Schulddiskurs in der frühen Nachkriegszeit. Ein Beitrag zur Geschichte des sprachlichen Umbruchs nach 1945. Berlin/New York: de Gruyter.
- Köhler, Reinhard (2012): *Quantitative Syntax Analysis (= Quantitative linguistics 65)*. Berlin/Boston: de Gruyter.
- Koplenig, Alexander/Meyer, Peter/Müller-Spitzer, Carolin (2014): Dictionary users do look up frequent words. A log file analysis. In: Müller-Spitzer, Carolin (Hg.): *Using Online Dictionaries (Lexicographica: Series Maior 145)*. Berlin/Boston: de Gruyter, 229–250.

- Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta/Tapias, Daniel/Rosner, Mike/Piperidis, Stelios/Odijk, Jan/Mariani, Joseph/Choukri, Khalid (Hgg.): Proceedings of the 7th conference on International Language Resources and Evaluation. (LREC-10). Valetta, Malta: European Language Resources Association (ELRA), 1848–1854.
- Lindemann, Margarete (1999): Mediostrukturen in modernen italienischen Wörterbüchern. In: *Lexicographica* 15, 38–65.
- Prinsloo, Danie J./Heid, Ulrich/Bothma, Theo/Faaß, Gertrud (2012): Devices for Information Presentation in Electronic Dictionaries. In: *Lexikos* 22, 290–320.
- Mann, Michael (2010): Internet-Wörterbücher am Ende der „Nullerjahre“: Der Stand der Dinge. Eine vergleichende Untersuchung beliebter Angebote hinsichtlich formaler Kriterien unter besonderer Berücksichtigung der Fachlexikographie. In: *Lexicographica* 26, 19–45.
- Meyer, Peter (2014): Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von *lexiko*. In: Abel, Andrea/Lemnitzer, Lothar (Hgg.): Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern (= OPAL - Online publizierte Arbeiten zur Linguistik 2/2014). Mannheim: Institut für Deutsche Sprache, 9–21.
- Müller, Peter O. (2002): Die Mediostruktur im De Gruyter Wörterbuch Deutsch als Fremdsprache. In: Wiegand, Herbert Ernst (Hg.): Perspektiven der pädagogischen Lexikographie des Deutschen II (= *Lexicographica*. Series Maior 110), Tübingen: Niemeyer, 485–496.
- Müller-Spitzer, Carolin (2007): Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis (= Studien zur Deutschen Sprache 42). Tübingen: Narr.
- Müller-Spitzer, Carolin (2013): Textual structures in electronic dictionaries compared with printed dictionaries. A short general survey. In: Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herbert Ernst (Hgg.): Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography (= *Handbücher zur Sprach- und Kommunikationswissenschaft*; 5.4). Berlin/Boston: de Gruyter, 367–381.
- Müller-Spitzer, Carolin/Wolfer, Sascha/Koplenig, Alexander (2015): Observing Online Dictionary Users: Studies Using Wiktionary Log Files. In: *International Journal of Lexicography* 28 (1), 1–26.
- Tarp, Sven. (1999): Theoretical foundations of the so-called cross-reference structures. In: *Lexicographica* 15, 114–137.
- Tarp, Sven (2008): Lexicography in the borderland between knowledge and non-knowledge: general lexicographical theory with particular focus on learner's lexicography. Tübingen: Niemeyer.
- Tarp, Sven, 2012: Theoretical challenges in the transition from lexicographical p-works to e-tools. In: Granger, Sylviane/Paquot, Magali (Hgg.): *Electronic lexicography*. Oxford: Oxford University Press, 107–118.
- Wiegand, Herbert Ernst (1996): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Zettersten, Arne/Hjørnager-Pedersen, Viggo (Hgg.): Symposium on Lexicography VII. Proceedings of the Seventh Symposium on Lexicography May 5–6, 1994 at the University of Copenhagen (= *Lexicographica*. Series Maior 76). Tübingen: Niemeyer, 11–49.
- Wiegand, Herbert Ernst (2002): Altes und Neues zur Mediostruktur in Printwörterbüchern. In: *Lexicographica* 18, 168–252.
- Wiegand, Herbert Ernst/Smit, Maria (2013): Mediostructures in printed dictionaries. In: Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herbert Ernst (Hgg.): Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography (= *Handbücher zur Sprach- und Kommunikationswissenschaft*; 5.4). Berlin/Boston: de Gruyter, 214–253.