Originalveröffentlichung in: Engelberg, Stefan/Lobin, Henning/Steyer, Kathrin/Wolfer, Sascha (Hrsg.): Wortschätze. Dynamik, Muster, Komplexität. - Berlin [u.a.]: de Gruyter, 2018. S. 269-292. (Jahrbuch des Instituts für Deutsche Sprache 2017)
DOI: *https://doi.org/10.1515/9783110579963-015*

Alexander Mehler/Rüdiger Gleim/Wahed Hemati/Tolga Uslu
(Frankfurt a.M.)

Skalenfreie online-soziale Lexika am Beispiel von Wiktionary

Abstract: Der Beitrag thematisiert Eigenschaften der strukturellen, thematischen und partizipativen Dynamik kollaborativ erzeugter lexikalischer Netzwerke am Beispiel von Wiktionary. Ausgehend von einem netzwerktheoretischen Modell in Form so genannter Mehrebenennetzwerke wird Wiktionary als ein skalenfreies Lexikon beschrieben. Systeme dieser Art zeichnen sich dadurch aus, dass ihre inhaltliche Dynamik durch die zugrundeliegende Kollaborationsdynamik bestimmt wird, und zwar so, dass sich die soziale Struktur der entsprechenden inhaltlichen Struktur aufprägt. Dieser Auffassung gemäß führt die Ungleichverteilung der Aktivitäten von Lexikonproduzenten zu einer analogen Ungleichverteilung der im Lexikon dokumentierten Informationseinheiten. Der Beitrag thematisiert Grundlagen zur Beschreibung solcher Systeme ausgehend von einem Parameterraum, welcher die netzwerkanalytische Betrachtung von Wiktionary als *Big-Data*-Problem darstellt.

1 Soziale Lexika als Mehrebenennetzwerke[1]

Dieses Kapitel analysiert Mehrebenennetzwerke als Modelle lexikalischen Wissens im Rahmen der verteilten Online-Kommunikation. In diesem Zusammenhang sprechen wir in Analogie zu Wikipedia als Medium des *collective problem solving* (Wang/Ye/Huberman 2012) bzw. zu *online-sozialen Netzwerken* (Thelwall 2009) von *online-sozialen Lexika*. Um den sozialen Charakter dieses Forschungsgegenstands zu betonen, beschreibt unser Ansatz mehrere Konstitutionsebenen (Abb. 1). Grundlegend hierfür ist 1) die Einbettung des untersuchten sprachlichen Systems in das soziale System jener Gemeinschaft von Sprachproduzenten, welche ersteres hervorbringen, und 2) die zirkuläre Konstitutionsbeziehung, welche für das Sprachsystem und seine Manifestation in unzähligen Situationen des Sprachgebrauchs konstatiert wird. In diesem Konstitutionsprozess sehen wir die Bedin-

1 Wir danken Henning Lobin für wertvolle wie auch kritische Hinweise zu diesem Beitrag. Ferner danken wir Daniel Baumartz für seine wertvolle Unterstützung dieses Beitrags.

DOI 10.1515/9783110579963-015

gung der Möglichkeit der Selbstorganisation, welcher die betrachteten Systeme ihre Gesetzmäßigkeit verdanken (Köhler/Altmann 1993).

Den Ausgangspunkt unseres Ansatzes bildet die Konzeption, wonach empirisch beobachtbare Einheiten etwa der Wortebene Manifestationen *relationaler Einheiten* eines auf die Produzenten der jeweiligen Sprachgemeinschaft verteilten Sprachsystems sind (de Saussure 1997). Für bestimmte Beobachtungseinheiten als Komponenten von Aggregaten der schriftlichen oder mündlichen Kommunikation postulieren wir folglich sprachsystematische Korrelate, von deren Vernetzungen wir annehmen, dass sie systemkonstitutiv wirken. Empirisch beobachtbare Repräsentationen dieser systemischen Netzwerke resultieren als *Beobachtungen 2. Ordnung* aus der Analyse ersterer Kommunikationseinheiten. Im Falle von Beobachtungen 2. Ordnung sprechen wir genauer von *Sprachgebrauchsnetzwerken* (SG-Netzwerken), welche eine Doppelrolle einnehmen: Erstens bilden sie gebrauchsorientierte Repräsentationen jener Korpora, auf deren Grundlage sie erzeugt wurden. Zweitens entsprechen sie den in der Theorie sprachlicher Netzwerke (Cong/Haitao 2014) bevorzugten Modellen der hypothetischen sprachsystematischen Netzwerke.

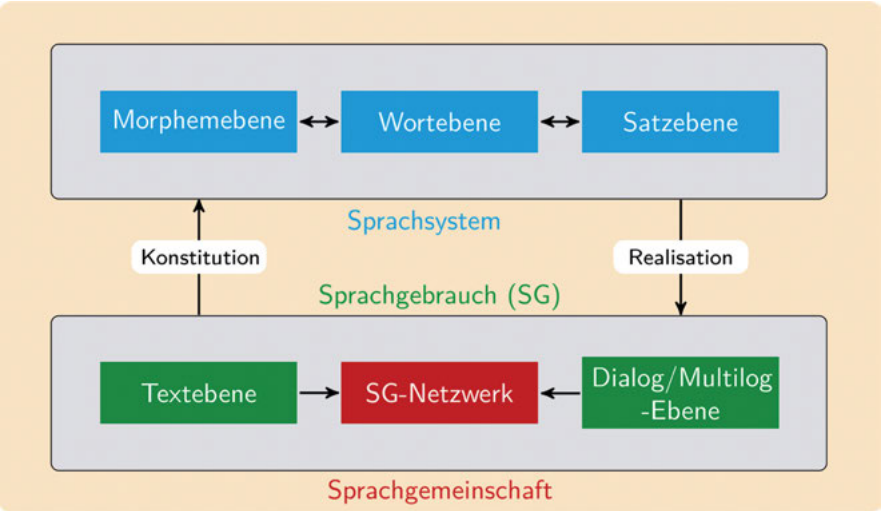


Abb. 1: Schema eines sozial-semiotischen Mehrebenennetzwerks

Der zirkuläre Konstitutionsprozess aus Abbildung 1 betrifft die Eigenart des über die Sprachgemeinschaft verteilten Lernens, wonach sprachliches Wissen zugleich Voraussetzung und Ergebnis sprachlichen Handelns ist (Rieger 1996;

Schnotz 1994) und also einer Änderungsdynamik unterliegt, welche mittels Fließgleichgewichten beschreibbar ist (Köhler 1986). Eine für soziale Systeme häufig beobachtbare Manifestation solcher Gleichgewichte besteht in *Zentrum-Peripherie-Strukturen*, die in Zusammenhang mit extremen Formen der Ungleichverteilung stehen (Stegbauer/Mehler 2011), wie sie das Präferenzgesetz für semiotische Systeme beschreibt (Tuldava 1998) und wofür das Zipfsche Gesetz (Zipf 1949) berühmt ist. Bezogen auf online-soziale Lexika führt diese Überlegung zu der Annahme, dass wir von einer Gemeinschaft von Lexikonproduzenten auszugehen haben, bei welcher der Output einer Minderheit hochaktiver Akteure den Output der mehrheitlich geringaktiven Akteure dominiert, und zwar so, dass sich die Topologie der sozialen Gemeinschaft der Topologie der resultierenden sprachsystematischen Netzwerke aufprägt, um schließlich mit Hilfe von SG-Netzwerken beobachtbar zu sein. Insofern wir für die soziale Topologie ebenso wie für ihr semiotisches Pendant eine Zentrum-Peripherie-Struktur annehmen, gelangen wir zu der Hypothese, dass sich extreme, skalenfreie Ungleichverteilungen in der Aktivitätsstruktur sozialer Agenten in ebenso skalenfreien Ungleichverteilungen der durch diese Aktivitäten hervorgebrachten sprachlichen Einheiten widerspiegeln. Anders formuliert: *Die soziale Struktur prägt sich der betroffenen Zeichenstruktur auf* (Mehler 2007). Lexika, welche diese mehrstufige Form der Skalenfreiheit aufweisen, bezeichnen wir als *skalenfrei*: Ein skalenfreies Lexikon ist also ein sprachliches Mehrebenennetzwerk im Sinne von Abbildung 1, dessen Fließgleichgewichte durch extreme Ungleichverteilungen auf sozialer und semiotischer Ebene gekennzeichnet sind. Solchen Lexika gilt unser Erkenntnisinteresse – hier am Beispiel des deutschsprachigen Wiktionarys.[2] Da wir das Wiktionary nur einer Sprache thematisieren, bildet das Kapitel eine messtheoretische Vorstudie zu einer Vergleichsstudie, welche Wiktionarys mehrerer Sprachen vergleicht. Das Kapitel ist wie folgt organisiert: Abschnitt 2 thematisiert eine Art von graphentheoretischer Polymorphie, derzufolge dasselbe Wiktionary eine Vielzahl von SG-Netzwerken induziert. Abschnitt 3 beschreibt eine Schar skalenfreier Eigenschaften von Wiktionary, während Abschnitt 4 Besonderheiten gegenüber vergleichbaren Ressourcen aufzeigt. Die Abschnitte 5 und 6 ergänzen Belege für skalenfreie thematische und partizipatorische Dynamiken und komplettieren somit das Bild von Mehrebenennetzwerken aus Abbildung 1.

2 Stand: 1.1.2017.

2 Wiktionary aus graphentheoretischer Sicht

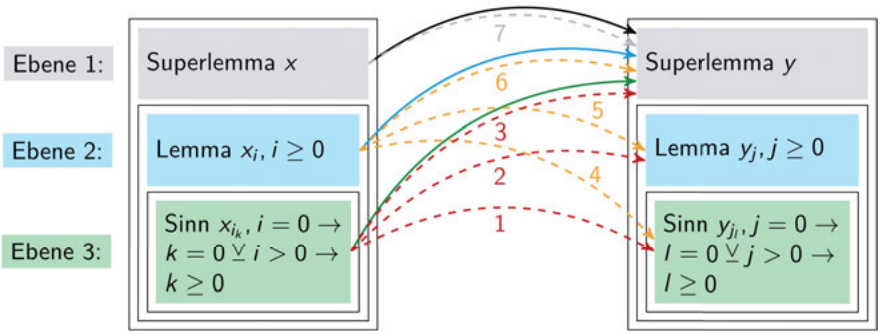


Abb. 2: Kanteninduzierende Einheiten in Wiktionary: Superlemma *x* als Einheit der *Ebene 1* umfasst Lemma x_i als Einheit der *Ebene 2* mit Lesart x_{ik} als Einheit der *Ebene 3* ($\veebar$ steht für die Kontravalenz). Durchgezogene Kanten von der Art parallel zu Kante 3 verbinden Lesarten mit Superlemmata (SL); durchgezogene Kanten von der Art parallel zu Kante 6 verbinden Lemmata mit SL und Kanten parallel zu Kante 7 verbinden SL untereinander. Ferner sind 7 Kanten inferierbar (gestrichelt) – teils basierend auf virtuellen Start- oder Endknoten. Alternativ werden Kanten durch Abstraktion von Lesarten zu (Super-)Lemmata oder durch Spezifikation unterspezifizierter Sinnrelationen inferiert.

Wiktionary ist mit den Mitteln simpler Graphen nicht unmittelbar abbildbar. Der Grund hierfür besteht in der hierarchischen Gliederung seiner Artikel, derzufolge das einen Artikel identifizierende Lemma (nachfolgend *Superlemma* genannt) mehrere untergeordnete Lemmata aufweisen kann, zu denen unter der Rubrik *Bedeutungen* je mehrere Bedeutungsbeschreibungen (*Sinne* oder *Lesarten* genannt) aufgeführt sein können (zur Zahl der Rubriken in Wiktionary und ihrer Behandlung im vorliegenden Beitrag siehe Tab. 1). Abbildung 2 bringt diesen Umstand schematisch zum Ausdruck. Sie zeigt eine 1:*m*:*n*-Beziehung, welche zwischen Superlemmata, Lemmata und Lesarten besteht.

Tab. 1: Zur Zahl der aus Wiktionary extrahierten Rubriken

	Zahl	Häufigkeit	%
extrahiert	54	3 189 073	93,77
nicht-extrahiert	1 979	211 984	6,23
Summe	2 033	3 401 057	100

Tab. 2: Manifestation von Sinnrelationen. Hyperlinks (unterstrichen) referieren auf Pages, die mehrere Superlemmata enthalten können. Überstriche denotieren *Broken Links*. In den entsprechenden Beispielen werden die betroffenen Wörter mittels][eingefasst. Wir setzen die Interpretation voraus, dass die Ziele von Hyperlinks Superlemmata eindeutig zugeordnet sind. Zahlen in Klammern geben an, wie oft die unverlinkte Aufzählungseinheit als Superlemma aufgefunden wurde. Dabei ist unabhängig von der Verlinkung von Konstituenten zu prüfen, ob die Einheit als Ganzes als Superlemma vorkommt. Gleichnamige Vorkommen von Superlemmata müssen nicht bedeuten, dass diese als Zieleinheiten intendiert sind.

Nr.	Relatum ⌈·⌉	Beispiel (Link)	#de	#alle
1.	⌈A⌉	Anschauung (https://de.wiktionary.org/wiki/Lehre)	549 185	778 612
2.	⌈A⌉	]Muskeldystrophie[(https://de.wiktionary.org/wiki/Muskel)	563 823	715 998
3.	⌈A⌉	Akupunkturpunkte (https://de.wiktionary.org/wiki/Chakra)	8 916 (1 827)	12 830 (2 683)
4.	⌈A B⌉	Gelbe Karte (https://de.wiktionary.org/wiki/Karte)	18 340	26 397
5.	⌈A B⌉	]angewandte Mathematik[(https://de.wiktionary.org/wiki/Numerik)	67 231	89 587
6.	⌈A B⌉	geometrisches Objekt (https://de.wiktionary.org/wiki/Punkt)	1 495	2 110
7.	⌈A B⌉	]ventrikuläre[ ]Fibrillation[(https://de.wiktionary.org/wiki/Kammerflimmern)	113	255
8.	⌈A B⌉	menschliche Gesellschaft (https://de.wiktionary.org/wiki/Welt)	1 899 (1 627)	2 626 (2 019)
9.	⌈A B⌉	sich]entäußern[(https://de.wiktionary.org/wiki/abkommen)	675 (508)	1 245 (847)
10.	⌈A B⌉	semantische Einheit (https://de.wiktionary.org/wiki/Semen)	731 (452)	971 (610)
11.	⌈A B⌉	]rotsterniges[Blaukehlchen (https://de.wiktionary.org/wiki/Blaukehlchen)	347 (158)	672 (280)
12.	⌈A B⌉	landwirtschaftliche]Nutzfläche[(https://de.wiktionary.org/wiki/Wiese)	144	507
13.	⌈A B⌉	]mehrzellige[Lebewesen (https://de.wiktionary.org/wiki/Eukaryota)	172	416
14.	⌈A B⌉	Homer Simpson (https://de.wiktionary.org/wiki/Zeichentrickfigur)	2 729 (51)	3 837 (91)

Die Vernetzungsstruktur von Einheiten dieser Ebenen resultiert aus dem Umstand, dass Artikel Rubriken enthalten, deren Inhalte auf Tokenebene mit den Superlemmata anderer Artikel relationiert sein können, und zwar explizit (mittels Hyperlinks) oder implizit. Das entsprechende Möglichkeitsspektrum exemp-

lifiziert Tabelle 2: Eine Besonderheit markieren gestrichelt unterstrichene *broken links*, welche die Anforderung zur Aufnahme entsprechender Superlemma/Lemma/Sinn-Tripel ausdrücken, wofür zum Browsing-Zeitpunkt kein Artikel existiert. Solche von Artikel-Autoren implizit referierten Superlemmata bezeichnen wir als *Typ-1-virtuell*: Ihnen ordnen wir jeweils genau ein virtuelles Lemma und einen virtuellen Sinn zu (siehe *angewandte Mathematik* in Tab. 2), um die Drei-Ebenen-Struktur der Grundeinheiten von Wiktionary beizubehalten. Bezogen auf referierte Einheiten unterscheidet Tabelle 2, Zeile 14 einen weiteren Fall von Virtualität, den wir als *Typ-2-virtuell* bezeichnen: Er betrifft gänzlich unverlinkte Relata von Sinnrelationen, zu denen ebenfalls keine Superlemmata existieren.

Wiktionary enthält drei Arten von Rubriken, deren Token auf (nicht-)virtuelle Einheiten referieren können: 1) Aufzählungen von Relata, welche in der durch die Rubrik benannten Relation (z.B. der Synonymie) zum Startargument stehen, 2) textuelle Beschreibungen oder satzwertige Aufzählungen, deren Gliederungseinheiten i.d.R. nicht Superlemmata entsprechen, und 3) Mischformen beider Varianten (z.B. *Charakteristische Wortkombinationen*). Im ersten Fall sprechen wir von *Aufzählungsrubriken* (i.d.R. zur Spezifikation von Sinnrelationen), im zweiten von *Beschreibungsrubriken* (bestehend aus erklärenden oder beispielgebenden Texten).

Ausgehend von Lesarten verweisen Relata von Aufzählungsrubriken zumeist auf Superlemmata mittels entsprechender Hyperlinks. *Verweisbeziehungen dieser Art bilden das Grundgerüst der expliziten Vernetzungsstruktur von Wiktionary*. Dabei tritt als Besonderheit die Möglichkeit *virtueller* Sinne aufseiten *referierender Artikel* auf, die in Zusammenhang mit impliziten Knoten und Kanten stehen: Referenzen auf Sinne als Startargumente von Sinnrelationen können explizit oder implizit durch Verwendung von Platzhaltern realisiert werden. Wir interpretieren entsprechende Vorkommen der Art von Zeile 2 und 3 aus Tabelle 3 als Hinweise auf virtuelle (referierende) Sinne: Hierzu führen wir modellseitig für jeden Platzhalter einen virtuellen Sinn (vom Typ 3 und 4) in der Rubrik *Bedeutungen* des betroffenen Lemmas ein. Zeile 4 aus Tabelle 3 demonstriert den Fall, dass bei der Spezifikation der Sinnrelationen eines Worts (im Beispiel *Auslegung*) das referierte Superlemma keinem der Sinne des referierenden Worts zugeordnet ist – auch nicht mittels Platzhalter. Hier hat man von einer Kante zwischen startbildendem Lemma und zielbildendem Superlemma auszugehen (siehe die Kanten von der Art parallel zu Kante 6 in Abb. 2). An dieser Stelle legen wir modellseitig einen virtuellen Sinn vom Typ 5 an. Während Typ-1- und Typ-2-virtuelle Sinne nur als Zielknoten von Sinnrelationen auftreten, können Typ-3-, Typ-4- und Typ-5-virtuelle Sinne als Start- und Zielknoten solcher Relationen auftreten.

Tab. 3: Zur Unterspezifikation von Lesarten als Startknoten von Sinnrelationen. [*i*], *i*∈*N*, denotiert einen auf Sinnebene spezifizierten Startknoten, [?], [*] und ^ hingegen nicht. Das Beispiel *Zierquitte* ist in][eingefasst und denotiert auf diese Weise einen *Broken Link*.

Nr.	Sinnbezug	Startknoten	Kante	Zielknoten	#de	#alle
1.	[*i*]	Rinne	[Synonyme].[1] →	Furche	1 078 494	1 475 064
2.	[?]	Quitte	[Unterbegriffe].[?] →	]Zierquitte[	3 129	4 907
3.	[*]	Video	[Unterbegriffe].[*] →	Montevideo	1 099	1 256
4.	^	Auslegung	[Synonyme].— →	Deutung	196 506	258 773

Tab. 4: Inferenzregeln für Sinnrelationen: V_1 ist die Menge der Superlemmata, V_2 der Lemmata und V_3 der Lesarten, $V = V_1 \cup V_2 \cup V_3$. R^* ist die zu R gegenläufige Sinnrelation, $R = R^{**}$. V^{**} ist die Menge aller Triaden $x.x'.x''$, so dass $x \in V_1$, $x' \in V_2$ und $x'' \in V_3$: $x.x'.x'' \in V^{**} \leftrightarrow (x, x') \in SL \wedge (x', x'') \in LS$; SL verbindet Superlemmata mit Lemmata und LS Lemmata mit Sinnen. $\lambda(x)$ ist die Menge aller Lemmata und $\sigma(x)$ die Menge aller Sinne, die $x \in V_1$ dominiert: $|\lambda(y)| = |\sigma(y)| = 1 \leftrightarrow (\exists!\, y', y'' \in V\colon y.y'.y'' \in V^{**})$. Gegenläufig sind Hyperonymie/Hyponymie, Holonymie/Meronymie und Augmentativ/Diminutiv. Alternativ ist $R^* = R$, falls R symmetrisch ist. Hierzu zählen die Synonymie, die Antonymie, die Übersetzungsrelation und die Relation der Sinnverwandtschaft (*Sinnverwandte Wörter*). Die letzte Zeile zeigt die Zahl der Anwendungen (die Regeln werden von links nach rechts angewandt).

	Inferenzregel 1		Inferenzregel 2		Inferenzregel 3
	$\forall R \forall y \in V_1 \forall x.x'.x'' \in \ddot{V}$:		$\forall R \forall y \in V_1\, x.x'.x'' \in \ddot{V}$:		$\forall R \forall x.x'.x'', y.y'.y'' \in \ddot{V}$:
$\wedge$	$\lvert\lambda(y)\rvert = 1 \wedge \sigma(y) = \{y''\}$	$\wedge$	$\lambda(y) = \{y'\} \wedge \lvert\sigma(y)\rvert \neq 1$	$\wedge$	$(x'', y) \in R \wedge (x'', y'') \notin R$
$\wedge$	$(x'', y) \in R \wedge (x'', y'') \notin R$	$\wedge$	$(x'', y) \in R \wedge (x'', y') \notin R$	$\wedge$	$(y'', x) \in R^*$
$\rightarrow$	$R \leftarrow R \cup \{(x'', y'')\}$	$\rightarrow$	$R \leftarrow R \cup \{(x'', y')\}$	$\rightarrow$	$R \leftarrow R \cup \{(x'', y'')\}$
#	1 323 503	#	267 976	#	140 445

Die Unterscheidung von Aufzählungs- und Beschreibungsrubriken macht die Differenzierung von Verweisbeziehungen, die mittels Hyperlinks manifestiert werden, und Sinnrelationen, die parallel hierzu verlaufen, nötig. In Zeile 8 aus Tabelle 2 etwa ist ⌈A B⌉ Startknoten einer Sinnrelation, wobei nur B per Hyperlink auf das entsprechende Superlemma bezogen ist. Zeile 4 aus Tabelle 2 zeigt demgegenüber den Paradefall von paralleler Sinnrelation und Verweisbeziehung. Beschreibungsrubriken sind dahingehend charakterisiert, dass sie ausschließlich Verweisbeziehungen manifestieren, während Aufzählungsrubriken Verweis- und Sinnrelationen ausweisen. Wiktionary ist sozusagen ein Lexikon, in dem eine

Hyperlink-basierte Verweisstruktur die zugrundeliegende semantische Vernetzungsstruktur basierend auf Sinnrelationen überlagert, ohne dass ein Isomorphismus beide Netzwerkebenen verbindet.

Unser Modell erlaubt die Unterscheidung von drei Arten *inferierbarer*, in Wiktionary nicht explizierter Sinnrelationen. Man beachte, dass Wiktionary Sinnrelationen i.d.R. zwischen Lesarten als Startknoten und Superlemmata als Zielknoten definiert (siehe die durchgezogene Kante parallel zur gestrichelten Kante 3 in Abb. 2). *Immer dann also, wenn das referierte Superlemma mehrere Sinne vermittels eines oder mehrerer Lemmata subsumiert, bleibt unterspezifiziert, auf welchen/welche dieser Sinne die Sinnrelation bezogen ist.* Geht man zudem von der Tatsache aus, dass Lemmata und Sinne ontologisch divergieren, so steht man vor dem Problem, tripartite Graphen anzunehmen, in denen Sinne stets auf Superlemmata verweisen. Mit einem solchen Modell ließe sich schwerlich die Vernetzungsstruktur von Wiktionary mit Ressourcen vergleichen, welche nicht dieser Logik folgen. Unter der Annahme also, dass Sinnrelationen über der Menge der Sinne definiert sind, ist die Inferenz entsprechender Kanten unabdingbar. Tabelle 4 unterscheidet hierzu Inferenzregeln (IR), um das Grundmuster des sinnrelationalen Verweises von Lesarten auf Superlemmata auf Lesarten herunterzubrechen: Während IR 1 und IR 3 dazu dienen, Sinnrelationen vom Typ 1 aus Abbildung 2 zu inferieren, sind Sinnrelationen vom Typ 2 aus Abbildung 2 das Ergebnis von IR 2. Weitere sechs, hier aus Platzgründen nicht aufgeführte Inferenzregeln betreffen die Übertragung letzterer Regeln auf die Lemma-Ebene bzw. die Abstraktion von Kanten unterer Ebenen auf darüber liegende Ebenen.

Unsere Darstellung spannt schließlich einen Parameterraum auf, welcher mehr Möglichkeiten ausweist, als derzeit mittels *Big Data Analysis* verarbeitbar sind. Dies zeigt folgende Formel:

$$(1) \quad |\dot{P}(\mathbb{E})| \times |\dot{P}(\mathbb{P})| \times |\dot{P}(\mathbb{R})| \times |\dot{P}(\mathbb{H})| \times |\dot{P}(\mathbb{V})| \times |\dot{P}(\mathbb{I})| = 2^3 - 1 \cdot 2^{100} - 1 \cdot 2^{54} - 1 \cdot 2 \cdot 2^5 - 1 \cdot 2^7 - 1 = 2^{170} - 5(1)$$

$\dot{P}(X)$ ist die Potenzmenge von X ohne leere Menge, $\mathbb{E}$ die Menge der Grundeinheiten (Superlemma, Lemma, Sinn), $\mathbb{P}$ die Menge der Wiktionary-Wortarten, $\mathbb{R}$ die Menge der Rubriken (siehe Tab. 1), $\mathbb{H}$ die Alternative, wonach Hyperlinks zur Kanteninduktion auszuwerten sind oder nicht, $\mathbb{V}$ die fünf Alternativen zur Erzeugung virtueller Sinne und $\mathbb{I}$ die sieben Typen inferierbarer Kanten aus Abbildung 2. Selbst unter Fokussierung auf Nomen, Adjektive, Verben und Sinnrelationen der Menge

$$(2) \quad \mathbb{R}' = \{\textit{Antonymie}, \textit{Heteronymie}, \textit{Holonymie}, \textit{Hyperonymie}, \textit{Hyponymie}, \textit{Meronymie}, \textit{Synonymie}\}$$

werden nahezu 70 Mio. Graphmodelle unterscheidbar, so dass selbst Stichproben mit 100 SG-Netzwerken allzu klein erscheinen. Daraus ergeben sich folgende

Schlussfolgerungen: 1) Wiktionary ist in Bezug auf den Ausweis von Lesarten und Sinnrelationen unterspezifiziert; 2) es induziert einen Raum möglicher Graphmodelle, deren Verarbeitung *Big-Data*-Methoden erforderlich macht; 3) die aus diesem Möglichkeitsraum ausgewählten Alternativen schränken den Aussagegehalt entsprechender Analysen ein. Im Folgenden konzentrieren wir uns auf SG-Netzwerke, deren Knoten Lesarten und deren Kanten Sinnrelationen entsprechen.

3 Skalenfreie lexikalische Strukturen

Tab. 5: Anpassung von Potenzgesetzen. SL steht für Superlemma, S für Sinn, [∀] für alle Sprachen, [de] für Deutsch. opt meint *optimierte Anpassung*. α ist der Exponent des angepassten Potenzgesetzes, *acd*∈[0,1] der *adjusted coefficient of determination*. Die letzte Spalte zeigt die Einheit mit der größten Wertausprägung.

Nr.	Verteilung	α	acd	α_{opt}	acd_{opt}	Top-Ausreißer
1	Rubriken (alle)	0,21	0,909 40	0,32	0,995 24	*Worttrennung*
2	Aufzählungsrubriken	0,13	0,826 50	0,15	0,829 98	*Ü-Tabelle*
3	Beschreibungsrubriken	0,23	0,879 85	0,40	0,994 46	*Worttrennung*
4	Wortarten [∀]	0,24	0,968 80	0,24	0,964 11	*Deklinierte Form*
5	Wortarten [∀+$SL_{virtuell}$]	0,20	0,940 88	0,20	0,931 60	*Substantiv*
6	Wortarten [de]	0,26	0,988 70	0,28	0,994 88	*Deklinierte Form*
7	Wortarten [de+$SL_{virtuell}$]	0,23	0,986 89	0,26	0,994 42	*Deklinierte Form*
8	Superlemmata/Sprache	0,30	0,980 23	0,35	0,992 45	*Deutsch*
9	Superlemmata/Seite	4,24	0,999 98	3,65	0,999 997	*a*
10	Lesarten/Lemma [∀]	2,04	0,991 80	3,67	0,999 27	*take*
11	Lesarten/Lemma [de]	2,32	0,996 80	3,50	0,999 65	*BA*
12	Lesarten/Lemma [∀+$S_{virtuell}$]	1,66	0,983 58	3,24	0,998 57	*take*
13	Lesarten/Lemma [de+$S_{virtuell}$]	1,73	0,986 35	4,01	0,998 56	*BA*
14	Synonyme/Lemma [de]	3,19	0,998 10	2,37	0,999 91	*Schickse*
15	Synonyme/Sinn [∀]	1,96	0,999 43	2,65	0,999 97	*Prostituierte*
16	Synonyme/Sinn [de]	1,66	0,996 59	2,59	0,999 80	*Prostituierte*
17	Sinnverwandte/Sinn [de]	2,80	0,997 72	2,46	0,999 85	*Prostituierte*
18	Hyperonyme/Sinn [de]	1,75	0,979 06	4,40	0,996 65	*Hund*
19	Hyponyme/Sinn [de]	2,27	0,987 56	2,33	0,999 62	*Taube*
20	Übersetzungen/Sinn [de]	0,94	0,956 50	2,25	0,998 83	*Wasser*

Wiktionary ist durch Skalenfreiheit charakterisiert, wie sie bereits für zahlreiche soziosemiotische Systeme konstatiert wurde. Im Fall von Wiktionary ist sie jedoch polymorph. Skalenfreiheit betrifft Systeme, in denen eine kleine Klasse von Wertausprägungen die jeweilige Verteilung dominiert und dabei äußerst schnell in die Gruppe der sehr zahlreichen, aber seltenen Ausprägungen übergeht. Diese Ordnung nach dem *semiologischen Präferenzgesetz* (Tuldava 1998) ist oft dadurch gekennzeichnet, dass es *keine typischen Systemausprägungen* gibt, und zwar in dem Sinne, dass der beobachteten empirischen Verteilung eine divergente theoretische Wahrscheinlichkeitsverteilung in Form eines Potenzgesetzes der Gestalt $y = c \cdot x^{-\alpha}$, $0<\alpha<2$, $\alpha \in R$ entspricht (Newman 2005). Potenzgesetze sind skalenfrei, da sie ihre Gestalt unter jedweder Skalenänderung beibehalten (Newman 2005). Sie sind charakteristisch für Ungleichverteilungen in sozialen (Stegbauer/Mehler 2011), semiotischen und lexikalischen Systemen (Tuldava 1998). *Polymorphie* bedeutet, Skalenfreiheit anhand einer Vielzahl systemrelevanter Bezugsgrößen zu beobachten. In Wiktionary entspricht sie dem dominanten Verteilungsmodell. Um dies zu belegen, dokumentiert Tabelle 5 die Skalenfreiheit von 20 Bezugsgrößen.

Zeile 1 aus Tabelle 5 thematisiert die Häufigkeitsverteilung von Rubriken. Abbildung 3 zeigt hierzu die empirische Ranghäufigkeitsverteilung der von uns identifizierten Rubriken. Dabei können wir von einer gelungenen Anpassung ausgehen (siehe Tab. 5, Zeile 1), ohne eine mittlere Rubrikenhäufigkeit erwarten zu dürfen (siehe oben): Der Inhalt von Wiktionary wird von wenigen Rubriken dominiert. Die häufigsten drei Rubriken halten einen Anteil von 0,47 an allen Rubrikangaben. Selbst wenn man die Verteilung auf die Sinnrelationen der Menge aus Formel 2 beschränkt, erhält man ein Potenzgesetz (Tab. 5, Zeile 2), was auf Selbstähnlichkeit hindeutet.

Im Fall der Häufigkeitsverteilung von „Wiktionary-Wortarten“ (WW) dokumentiert Tabelle 5 das gleiche Bild, wobei wir vier Modalitäten unterscheiden: bezogen auf Superlemmata aller Sprachen (Zeile 4 und 5), des Deutschen (6 und 7), unter Aus- (4 und 6) oder Einschluss (5 und 7) virtueller Superlemmata. Auch hier dominieren wenige Häufigkeitsklassen, während die überwiegende Mehrheit der WW selten oder nur einmal belegt wird. Ferner beobachten wir, dass es keine typische Vorkommenshäufigkeit gibt – sprach- und inferenzunabhängig.

Tabelle 5 zeigt, dass diese Diagnose für eine Reihe weiterer Verteilungen gilt, etwa die Verteilung der Superlemmata je Seite (Zeile 9), der Lesarten je Lemma (Zeilen 10–13) oder der Synonyme je Lesart. Um diese Polymorphie allgemein nachzuweisen, ist es erforderlich, für die Mehrheit aller systemrelevanten Attribute deren Skalenfreiheit aufzuzeigen. Aus Platzgründen entfällt dies hier. Tabelle 5 zeigt allemal, dass Skalenfreiheit eine prägnante Eigenschaft der in Wiktionary dokumentierten Strukturen ist.

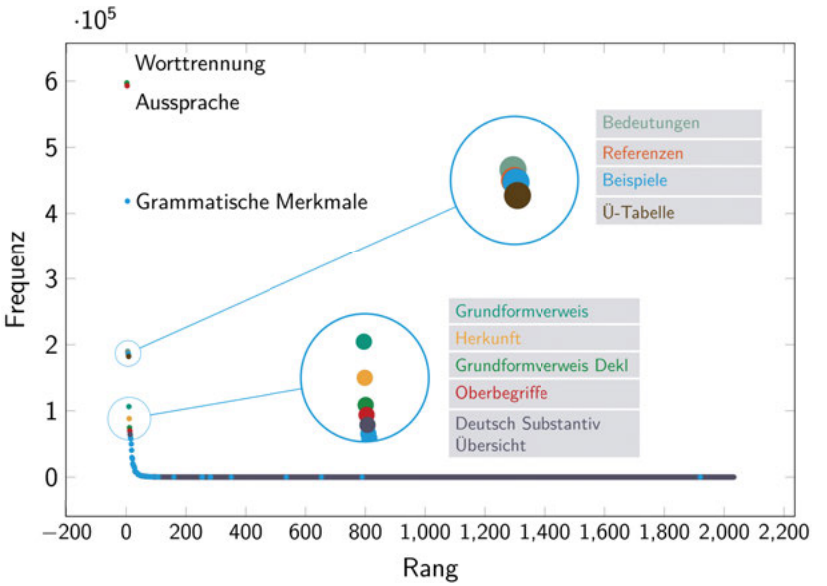


Abb. 3: Rang-Häufigkeitsverteilung der Rubriken in Wiktionary. *x*-Achse: Rang der Rubrik; *y*-Achse: ihre Häufigkeit in Wiktionary

4 Lexikalische Vernetzung

Dieser Abschnitt thematisiert Vernetzungsindikatoren, welche mit Eigenschaften lexikalischer Systeme in Verbindung stehen (Sigman/Cecchi 2002; Motter et al. 2002; Steyvers/Tenenbaum 2005; Gravino et al. 2012). Dabei konzentrieren wir uns auf konnektierte Komponenten, durchschnittliche geodätische Distanzen von Lesarten und deren Transitivität (Newman 2010). Maßgeblich hierfür sind Beobachtungen zu vergleichbaren Systemen, welche hohe Transitivitätsraten, skalenfreie Gradverteilungen (siehe Zeilen 15–20 in Tab. 5) und durchschnittlich kurze Pfade aufweisen (Steyvers/Tenenbaum 2005).

Naheliegende Interpretationen dieser Eigenschaften betreffen die Effizienz und Ausfallsicherheit von Gedächtnisoperationen (Motter et al. 2002) ebenso wie die flexible semantische Interpretierbarkeit situationsgebundener Sprache: Ausgehend von immer neuen Situationen können Sprachteilnehmer dieselben Einheiten auf immer neue Weisen anwenden, ohne Interpretationsmöglichkeiten einzubüßen, was etwa infolge des Abbruchs der Aktivierungsausbreitung im Gedächtnis aufgrund vermeintlich fehlender Relationen oder diskonnektierter Komponenten geschehen kann.

Tab. 6: Potenzgesetze angepasst an die komplementär-kumulativen Größenverteilungen konnektierter Komponenten von SG-Netzwerken basierend auf Nomen, Verben und Adjektiven. *GKK* ist die *Größte Konnektierte Komponente*.

Nr.	St.	PoS	Modell	α	acd	α_{opt}	acd_{opt}	$\|GKK\|$	Anteil
1	i	N	{Hypo, Hyper}	4,14	0,998 7	1,91	0,999 9	73 748	0,32
2	ii	N	{Hypo, Hyper, Ant, Syn}	3,67	0,999 1	2,25	0,999 9	114 573	0,49
3	iii	N	$\mathbb{R}'$	3,67	0,999 1	2,25	0,999 9	114 608	0,49
4	iv	N	$\mathbb{R}''$	3,15	0,999 9	2,75	1,000 0	142 905	0,61
5	i	V	{Hypo, Hyper}	4,93	0,999 6	2,70	0,999 9	619	0,02
6	ii	V	{Hypo, Hyper, Ant, Syn}	4,23	0,999 7	2,55	1,000 0	7 524	0,30
7	iii	V	$\mathbb{R}'$	4,23	0,999 7	2,55	1,000 0	7 524	0,30
8	iv	V	$\mathbb{R}''$	3,79	0,999 9	2,75	1,000 0	9 955	0,40
9	i	A	{Hypo, Hyper}	4,99	0,999 6	2,73	0,999 9	711	0,02
10	ii	A	{Hypo, Hyper, Ant, Syn}	3,91	0,999 6	2,03	0,999 9	9 125	0,32
11	iii	A	$\mathbb{R}'$	3,91	0,999 6	2,03	0,999 9	9 125	0,32
12	iv	A	$\mathbb{R}''$	3,26	0,999 9	2,29	1,000 0	11 832	0,41

Tab. 7: Netzwerkstatistiken der Graphvarianten aus Tabelle 6: $|V|$ ist die Zahl der Knoten des Graphen, $|E|$ die Zahl seiner Kanten. D ist der Diameter, L die durchschnittliche geodätische Distanz (beide bzgl. der GKK), C_{ws} der Clusterwert nach Watts/Strogatz (1998). Kursive Werte basieren auf 10.000 Knotenpaaren.

Nr.	St.	PoS	Modell	$\|V\|$	$\|E\|$	D	L	C_{ws}
1	i	N	{Hypo, Hyper}	232 613	236 086	*19*	*8,59*	0,15
2	ii	N	{Hypo, Hyper, Ant, Syn}	232 613	360 706	*24*	*10,97*	0,28
3	iii	N	$\mathbb{R}'$	232 613	360 808	*24*	*10,90*	0,28
4	iv	N	$\mathbb{R}''$	232 613	477 786	*40*	*12,49*	0,20
5	i	V	{Hypo, Hyper}	25 162	6 996	30	9,67	0,05
6	ii	V	{Hypo, Hyper, Ant, Syn}	25 162	24 970	27	9,34	0,18
7	iii	V	$\mathbb{R}'$	25 162	24 970	27	9,34	0,18
8	iv	V	$\mathbb{R}''$	25 162	34 693	23	8,64	0,18
9	i	A	{Hypo, Hyper}	28 517	7 786	20	6,80	0,08
10	ii	A	{Hypo, Hyper, Ant, Syn}	28 517	35 969	29	10,02	0,24
11	iii	A	$\mathbb{R}'$	28 517	35 969	29	10,02	0,24
12	iv	A	$\mathbb{R}''$	28 517	52 680	27	8,16	0,24

Während nun lexikalische Systeme wie WordNet (Fellbaum 1998) solche Netzwerkeigenschaften aufweisen, stellt sich die Frage, ob das auch für Wiktionary gilt. Im negativen Fall deutet dies auf seine „Unnatürlichkeit" hin. Hierzu betrachten wir SG-Netzwerke separat für einzelne Wortarten. Dabei gehen wir vierstufig vor, indem wir nach der Vernetzung von Nomen, Verben und Adjektiven unter Anwendung der Regeln aus Tabelle 2 fragen, wobei Stufe 1 nur Hyponymie- und Hyperonymie-, Stufe 2 zusätzlich Synonymie- und Antonymie-, Stufe 3 zusätzlich die Menge R' und Stufe 4 alle 34 extrahierten Relationen (siehe Tab. 2) berücksichtigt, welche aus Aufzählungsrubriken resultieren (siehe Tab. 6). Offenbar gilt: je höher die Stufe, desto höher die Vernetzungswahrscheinlichkeit, desto eher sollte Wiktionary Eigenschaften vergleichbarer Netzwerke (Mehler 2008) aufweisen, wobei die Berücksichtigung von Inferenzregeln einen monotonen Anstieg letzterer Vernetzungswahrscheinlichkeit impliziert. Finden wir also heraus, dass unsere SG-Netzwerke sehr viel weniger konnektiert sind, belegt dies die oben thematisierte Unnatürlichkeit. Dies entspricht offenbar unserer Beobachtung. Tabelle 6 zeigt zunächst, dass in keinem der betrachteten Netzwerke die jeweilige GKK alle Nomen, Verben oder Adjektive umfasst. Zwar wächst die GKK mit der betrachteten Vernetzungsstufe (i–iv), jedoch verbleibt diese stets weit unter der Ordnung des Graphen. Ferner fällt auf, dass im Übergang von Stufe ii zu iii kaum (bei Nomen) oder keine zusätzliche Konnektivität erzielt wird: Über die Relationen der Stufe ii hinaus wirken die übrigen Sinnrelationen kaum vernetzend. Dass im Fall von Stufe i die GKK zumeist winzig ist, hat nichts damit zu tun, dass Hyponymie-/Hyperonymie-Relationen hierarchische Grundgerüste aufspannen (Sigman/Cecchi 2002).[3] Denn bereits auf dieser Ebene ist erwartbar, dass der entsprechende Graph mittelbare Konnektiertheit nahezu aller Lesarten herstellt. Dass diese Konnektiertheit auf höheren Stufen zwar zunimmt, jedoch weit unterhalb des erwarteten Maximums verbleibt, deutet darauf hin, dass Wiktionary Komponenten enthält, welche unabhängig voneinander entwickelt wurden.

Die Hypothese, wonach diese Komponenten thematisch bedingt sind, wird im folgenden Abschnitt untersucht. Tabelle 7 zeigt ergänzend, dass die Netzwerke zu langen Wegen neigen, was der Hypothese widerspricht, es handele sich bei ihnen um kleine Welten (Milgram 1967; Watts/Strogatz 1998) – ein Attribut, das vergleichbaren Strukturen zugesprochen wird (Mehler 2008). Die Transitivität ist im untersuchten Ausschnitt von Wiktionary zwar hoch (C_{ws}), die kürzesten Pfade sind dabei jedoch lang. Im Sinne der Netzwerktheorie ist Wiktionary daher ungewöhnlich.

3 De facto sind sie in Wiktionary an manchen Stellen zirkulär, wie im Falle anderer sozialer Ontologien (Mehler/Pustylnikov/Diewald 2011).

5 Thematische Dynamik

Die Frage, welche Themenfelder Wiktionary mit seinen Lesarten abdeckt, ist nur mit Methoden der Automatisierung zu beantworten. Anhand der Antwort erwarten wir weiteren Aufschluss über die Skalenfreiheit von Wiktionary – nun in Bezug auf seinen Inhalt. Unser Ansatz besteht darin, Lesarten unter Bezug auf die zweite Ebene der *Dewey Decimal Classification* (DDC) zu klassifizieren. Die 100 Themenfelder dieser Ebene bilden unser Inhaltsmodell. Da die DDC eine im Bibliotheksbereich weit verbreitete Inhaltsklassifikation ist, eignet sie sich offenbar für diese Zwecke. Unser Algorithmus basiert auf *fastText* (Joulin et al. 2016), einem Klassifikator in Form eines neuronalen Netzes mit nur einem *hidden layer*. Zwecks Erstellung von Trainingsdaten extrahieren wir alle Wikipedia-Referenzen aller extrahierten Lesarten auf die entsprechenden Wikipedia-Seiten unter Absehung von Disambiguierungsseiten. In einem zweiten Schritt explorieren wir die Hyperlinks der referierten Wikipedia-Artikel auf die *Gemeinsame Normdatei* (GND), um schließlich Lesarten auf DDC-Klassen abzubilden. Auf diese Weise werden 7.270 Lesarten 95 DDC-Klassen zugeordnet. Den Input für das Trainieren von *fastText* bildet der Textinhalt, welcher der betrachteten Lesart in Wiktionary zugeschrieben wird (Synonyme, Beispieltexte etc.). Um hiervon unabhängig die Güte unseres Ansatzes zu bemessen, teilen wir das Trainingskorpus von 7.270 Texten nach der Regel 80/20 in eine Trainings- und eine Testmenge auf. Für die Testmenge erzielen wir einen F-Wert von über 75% – angesichts von 95 Zielklassen ist dies ein gutes Ergebnis.

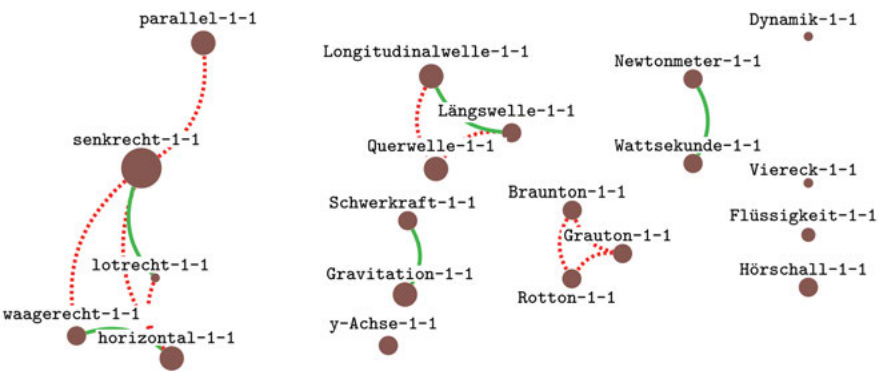


Abb. 4: Die 20 Lesarten, welche Formel 3 der DDC-Klasse für Physik zuoberst zuordnet: Verlinkung mittels Synonymie- (grüne bzw. durchgezogene Linien) und Antonymie-Relationen (rote bzw. gestrichelte Linien). Je höher der GSS-Wert einer Lesart, desto größer ihr Knoten. Suffixe kodieren Lemma- und Sinn-IDs.

Indem wir den anhand aller 7.270 Texte trainierten Klassifikator auf alle extrahierten Sinne aus Wiktionary anwenden, erhalten wir für jeden Sinn einen Vorschlag für die wahrscheinlichste DDC-Klasse, der er angehört. Im nächsten Schritt erstellen wir ein Netzwerk der in Wiktionary mittels Lesarten-Instanzen nachweisbaren DDC-Klassen, um Informationen über deren Zusammenhangsstruktur zu gewinnen. Hierzu gewichten wir jede Sinninstanz jeder Zielklasse mittels des Gewichtungsschemas (Galavotti/Sebastiani/Simi 2000):

(3) $$GSS(s_k, c_i) = P(s_k, c_i) \cdot P(\neg s_k, \neg c_i) - P(s_k, \neg c_i) \cdot P(\neg s_k, c_i)$$

Hierzu werten wir die Sinnrelationen der Ziellesart aus, um deren Zugehörigkeit zur Kategorie und also die Parameter aus Formel (3) zu schätzen. Grob gesprochen berechnen wir, wie konsistent die Endpunkte der Sinnrelationen von s_k in dem durch c_i vorgegebenen Bereich verbleiben. Abbildung 4 exemplifiziert dies am Beispiel der DDC-Klasse für Physik.

Auf dieser Basis weist unser Algorithmus das in Abbildung 5 gezeigte Netzwerk von Themenfeldern in Wiktionary nach. Hierzu wird für jede betrachtete Relationsart (hier *Synonymie* und *Antonymie*) immer dann eine Kante zwischen zwei DDC-Klassen bzw. derselben Klasse gezogen bzw. deren Gewicht erhöht, wenn zwei Instanzen dieser Klassen identifiziert werden, welche in der betrachteten Sinnrelation zueinander stehen. Somit werden Kanten zwischen Themenfeld-Knoten mittels sinnrelational verwandter Lesarten als Instanzen der Felder inferiert, was ein makroskopisches Bild der Themenverteilung und -vernetzung in Wiktionary liefert. Abbildung 5 zeigt, dass diese Verteilung ungleichmäßig ist. Dies betrifft nicht nur die Verteilung der Klassen der ersten, sondern auch der zweiten DDC-Ebene. Um hierüber weiteren Aufschluss zu erlangen, zeigt Abbildung 6 die Ranghäufigkeitsverteilung für die nachgewiesenen 95 Klassen der zweiten Ebene. Abbildung 7 zeigt die entsprechend separierten Verteilungen für die zehn Klassen der ersten Ebene.

Die Verteilungen ergeben folgendes Bild: In Wiktionary sind Lesarten auf Themenbereiche schief verteilt; diese Schiefe entspricht nur mit Abstrichen einem Potenzgesetz (siehe Abb. 6). In jedem Fall aber existieren wenige, dominante Themenfelder, denen die Mehrheit der extrahierten Sinne angehört. Sie entstammen den DDC-Klassen 3 *Social Sciences*, 5 *Science* und 6 *Technology*. Die 10 größten DDC-Klassen der zweiten Stufe decken 50,7% aller klassifizierten Sinne ab, während sich die Restmenge auf die übrigen 85 Klassen verteilt (Abb. 6). Die mit Abstand dominantesten Themenbereiche sind Klasse 3 (*Social Sciences*) und 6 (*Technology*): Sie decken 26,1% bzw. 26,4% aller Sinne ab. Diesem Modell nach ist Wiktionary semantisch verzerrt: wenige Bereiche dominieren die thematische

Provenienz der Sinne, während die Mehrzahl aller übrigen Bereiche unterrepräsentiert ist. Dieser Interpretation ist Folgendes entgegenzuhalten: Zum einen verwenden wir einen fehlerbehafteten Klassifikator mit einem F-Wert von ca. 75%. Zum anderen setzen wir voraus, dass die DDC eine repräsentative Klassifikation zur Inhaltsbestimmung ist. Drittens kennen wir nicht die Größe der Vokabulare solcher Themenfelder im Deutschen, so dass wir deren Abdeckungsrate nicht genau bestimmen können. Solange jedoch kein Alternativmodell existiert, das unseren Ansatz falsifiziert, können wir von einer thematischen Präferenz bzw. Ungleichheit in Wiktionary sprechen.

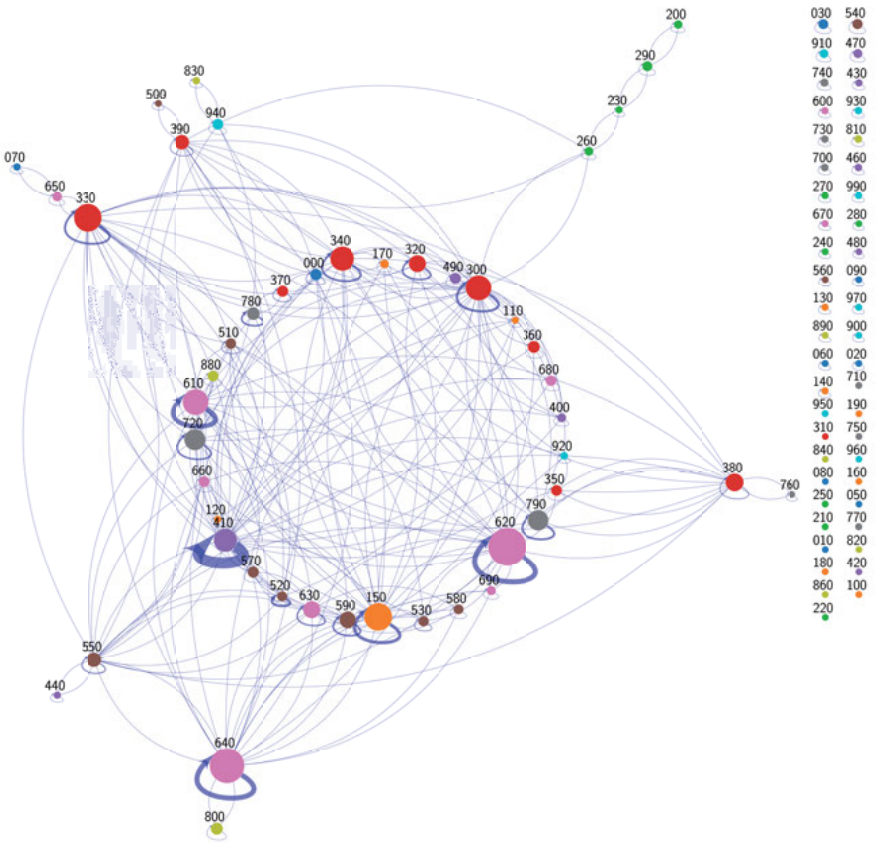


Abb. 5: Verteilung und Vernetzung von Themenfeldern der zweiten DDC-Ebene in Wiktionary: je höher die Zahl der für eine Klasse nachweisbaren Lesarten, desto größer der sie abbildende Knoten; je größer die Zahl der Sinnrelationen von Instanzen zweier oder derselben Klasse, desto breiter die sie verbindende Kante.

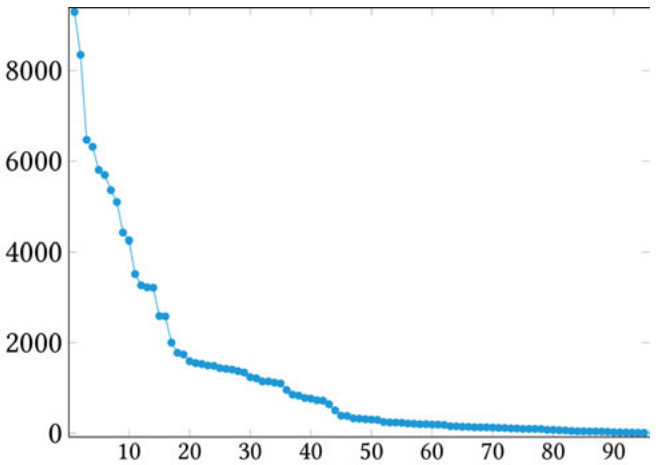


Abb. 6: Ranghäufigkeitsverteilung der Sinne in Wiktionary je Klasse der zweiten DDC-Ebene. *x*-Achse: Rang beginnend mit der umfangreichsten Klasse; *y*-Achse: Zahl der Sinne der rangbildenden Klasse. Die Anpassung eines Potenzgesetzes führt zu $\alpha = 0.349$, $acd = 0.927$.

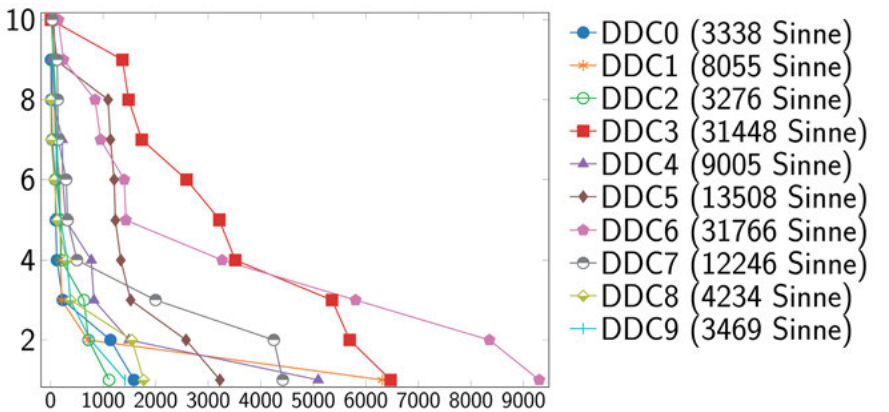


Abb. 7: Die zehn komplementär-kumulativen Verteilungen der Zahl der Sinne je Klasse der zweiten DDC-Ebene. *x*-Achse: Zahl der Sinne; *y*-Achse: Zahl der DDC-Klassen der zweiten Ebene derselben ersten Ebene, denen mindestens *x* Sinne zugeordnet sind.

6 Partizipative Dynamik

Dieser Abschnitt thematisiert abschließend die Partizipationsdynamik von Wiktionary. Ziel ist es, das Beitrags- und Kollaborationsverhalten von Wiktionary-Autoren zu beleuchten. Dieser Ansatz basiert auf dem Kollaborationsmodell und Aktivitätskonzept von Brandes et al. (2009). Wir vereinfachen dieses Modell dahingehend, dass wir lediglich Differenzen von Längen von Revisionen messen, um Aktivitäten zu schätzen. Mittels dieser Messgröße soll das Beitrags- und Kollaborationsverhalten von Autoren als Netzwerk analysiert werden. Hierbei denotieren Knoten Autoren, während Kanten deren Beziehungen als Funktion gemeinsam bearbeiteter Seiten repräsentieren. Auch im Hinblick auf die Gestaltung und Färbung von Knoten orientieren wir uns an Brandes et al. (2009). Dazu dient folgender Induktionsalgorithmus:

1. *Knotengestalt*: Die Gestalt eines Knotens ist eine Funktion der Aktivität des entsprechenden Autors und der Zahl der von ihm bearbeiteten Seiten:
   a) *Höhe:* Je höher die Aktivität, desto höher der Knoten.
   b) *Breite:* Je mehr Seiten ein Autor bearbeitet, desto breiter der Knoten.
2. *Knotenfarbe*: Die Knotenfarbe ist eine Funktion der Summe der Anteile des jeweiligen Autors an den von ihm bearbeiteten Seiten:
   a) Je höher die Summe, desto stärker die Grünfärbung.[4]
   b) Je geringer die Summe, desto stärker die Rotfärbung.[5]
3. *Knotenvernetzung*: Die Verlinkung der Knoten geschieht mittels folgender Gewichtungsfunktion der Koautorenschaft zweier Autoren x, y bezogen auf alle Revisionen aller Artikel T:

$$coauthorship(x, y) = \sum_{T} \frac{2 \cdot \min(activity(x, T), activity(y, T))}{\sum_{x \in authors(T)} (activity(x, T)} \in [0, 1] \quad (4)$$

   Es werden nur solche Kanten gezogen, für die $coauthorship(x, y)$ mindestens der Summe aus entsprechendem Mittelwert und Standardabweichung entspricht.

Abbildung 8 zeigt das Kollaborationsnetzwerk der 100 aktivsten Autoren (Bots sind umrandet). *BetterkBot* und *YS-Bot* entsprechen der Erwartung, wonach Bots zugleich flache (geringer Aktivitätsgrad), breite (Bearbeitung vieler Seiten) und rotgefärbte Knoten (geringe Autorenschaftsanteile) induzieren. Typischen Bots

4 In der Schwarz-Weiß-Darstellung erscheinen grün gefärbte Agentenknoten im Schachbrettmuster.

5 In der Schwarz-Weiß-Darstellung werden rot gefärbte Agentenknoten ausgefüllt.

gegenüber stehen hochaktive Autoren, welche auf vielen Seiten viel beitragen und dabei wenige Koautoren oder nur solche mit geringen Autorenschaftsanteilen haben. Bemerkenswert an Abbildung 8 ist der hohe Vernetzungsgrad der Akteure: hochaktive Agenten bilden offenbar ein dichtes Kollaborationszentrum.

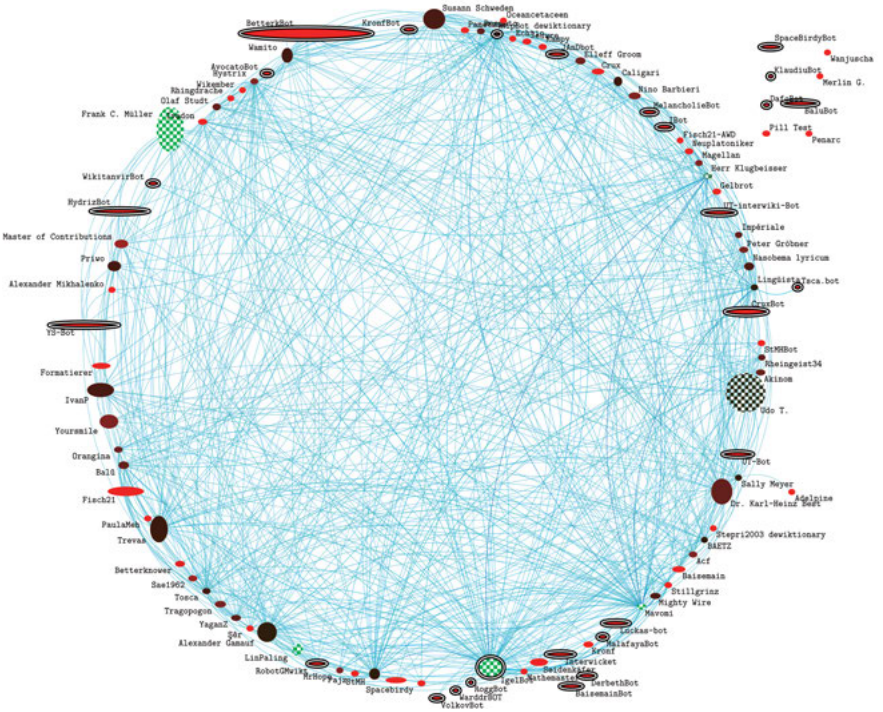
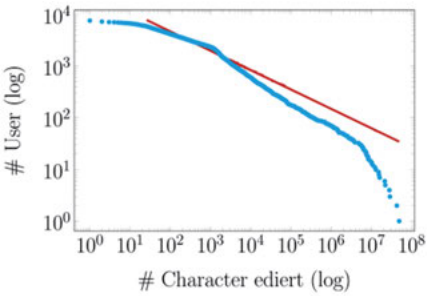
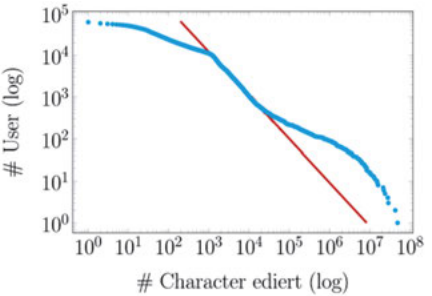


Abb. 8: Das Kollaborationsnetzwerk der 100 aktivsten Wiktionary-Autoren

An dieser Stelle wenden wir uns wieder der Frage nach der Skalenfreiheit zu. Anhand von Abbildung 8 fällt auf, dass flache rote Knoten überwiegen. Dies deutet darauf hin, dass wenige Autoren viel Inhalt produzieren, und zwar im Vergleich zur Mehrheit kaum aktiver Autoren, welche zudem dazu tendieren, an Seiten mit vielen Koautoren beteiligt zu sein – unabhängig davon, ob ihre Beteiligung auf viele Seiten bezogen ist. Diese Tendenz bestätigen die Verteilungen der Aktivitäten bzw. Beteiligungen aller (registrierten) User (siehe Abb. 9 und 10). Unabhängig davon, ob wir alle oder nur registrierte User betrachten, und unabhängig davon, ob wir Aktivitäten oder Seitenbeteiligungen betrachten, resultieren potenzgesetzliche Verteilungen, denen zufolge eine Minderheit hochaktiver, breit beteiligter Autoren einer großen Mehrheit gering-aktiver, auf wenige Seiten fokussierter Autoren gegenübersteht.

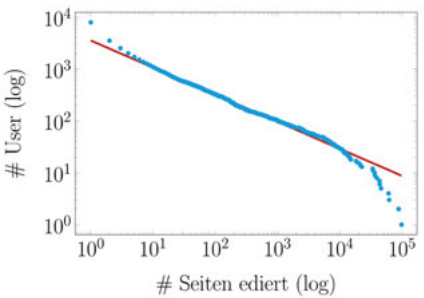


(a) Registrierte User, acd = 0,963740, α = 0,38

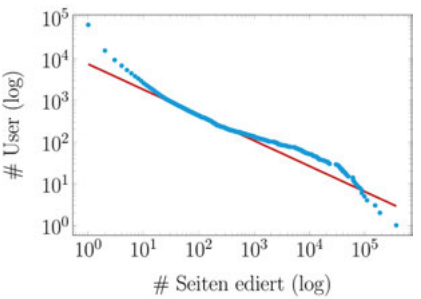


(b) Alle User, acd = 0,999673, α = 1,04

Abb. 9: Komplementär-kumulative Verteilung der Aktivitäten registrierter (links) bzw. aller User (rechts)



(a) Registrierte User, acd = 0,999294, α = 0,52



(b) Alle User, acd = 0,999 795, α = 0,61

Abb. 10: Komplementär-kumulative Verteilung der Seitenbeteiligungen registrierter (links) bzw. aller User (rechts)

Durch diese Beobachtung schließt sich der Kreis zu den Überlegungen in der Einleitung, wonach die oben beschriebenen Formen struktureller und thematischer Skalenfreiheit auf partizipatorischer Ebene gespiegelt werden. Das wirft die Frage auf, inwiefern erstere Skalenfreiheit das Produkt ihres sozialen Pendants ist. Wenige hochaktive Agenten haben das Potenzial, Themen zu setzen und somit die thematische Ausrichtung von Wiktionary zu verzerren. Das spricht dafür, dass soziale Skalenfreiheit eine der Ursachen ihres thematischen Pendants ist. Inwiefern diese vermutete Kausalität auch die Skalenfreiheiten aus Abschnitt 3 betrifft, ist auf dieser Grundlage nicht zu beantworten. Wie in der Einleitung erwähnt wurde, zielt dieser Beitrag nicht auf die Frage nach der Kausalität, sondern auf methodische Grundlagen, mit deren Hilfe entsprechende Ant-

worten eingrenzbar werden, und zwar mit Bezug auf das Konzept des skalenfreien Lexikons. Der Beitrag belegt am Beispiel des deutschsprachigen Wiktionarys jedoch bereits die vermutete Parallelität von semiotischer und sozialer Skalenfreiheit und damit eine notwendige Bedingung für skalenfreie Lexika.

7 Zusammenfassung

Der vorliegende Beitrag hat skalenfreie Lexika am Beispiel des Deutschen Wiktionarys thematisiert. Ausgehend von dem Konzept des Mehrebenennetzwerks wurde die strukturelle, thematische und partizipative Dynamik von Wiktionary untersucht. Unsere Ergebnisse werfen die Frage auf, ob sich die soziale Struktur von Wiktionary auf seine inhaltliche Struktur abbildet: *Ist also die skalenfreie Struktur dieses Lexikons Ausdruck der Skalenfreiheit seiner Partizipationsdynamik oder spiegelt Wiktionary lediglich die lexikalische Struktur der dokumentierten Sprache(n) wider?* Ist nämlich die lexikalische Struktur einer Sprache, wie es viele Untersuchungen nahelegen (vgl. Abschn. 4), skalenfrei, so ist Wiktionary möglicherweise bloß das Abbild dieser Struktur. Wir finden Hinweise darauf, dass dem nicht so ist. Sie betreffen die Untersuchung der thematischen Dynamik, die zeigt, dass die dokumentierten Lesarten wenigen Inhaltsbereichen entstammen und also thematische Vorlieben hochaktiver Artikelschreiber widerzuspiegeln scheinen. Auf dieser Grundlage ließe sich argumentieren, dass dominante Artikelschreiber die von Wiktionary abgedeckten Themenfelder zugunsten ihrer thematischen Präferenzen verzerren, während die dokumentierten Sinnrelationen Skalenfreiheit aufweisen (Abschn. 3) und dabei Netzwerke unterhalb der Konnektiertheit vergleichbarer Systeme aufspannen (Abschn. 4). In diesem Sinne wäre Wiktionary zwar eine hochinformative Ressource, seinem Inhalt nach jedoch weit verzerrter, als es die gestellte Aufgabe der Repräsentation lexikalischen Wissens aus der Sicht der jeweiligen Sprache erwarten ließe. Gegen diese Interpretation lässt sich einwenden, dass die hier dokumentierte thematische Schiefe nur mit Abstrichen als skalenfrei bezeichnet werden kann, dass unsere Methode der Messung thematischer Provenienz fehleranfällig ist (Abschn. 5), dass nur ein Wiktionary untersucht wurde, dass sich Wiktionary weiterentwickelt und die beobachtete Schiefe möglicherweise ausgleichen wird und dass die kausalen Beziehungen von struktureller, thematischer und partizipativer Dynamik einer genaueren Erforschung bedürfen. Diese Kritik impliziert, dass das Konzept des skalenfreien Lexikons durch die vorliegende Studie zwar belegt wird, jedoch einer Erweiterung um Wiktionarys vieler Sprachen bedarf. Dies soll die Aufgabe zukünftiger Arbeiten sein.

Literatur

Brandes, Ulrik et al. (2009): Network analysis of collaboration structure in Wikipedia. In: Proceedings of the 18th International Conference on World Wide Web (WWW '09). New York, S. 731–740.

Cong, Jin/Haitao, Liu (2014): Approaching human language with complex networks. In: Physics of Life Reviews 11, 4, S. 598–618.

de Saussure, Ferdinand (1997): Linguistik und Semiologie. Notizen aus dem Nachlaß. Frankfurt a.M.

Fellbaum, Christiane (1998): WordNet. An electronic lexical database. Cambridge.

Galavotti, Luigi/Sebastiani, Fabrizio/Simi, Maria (2000): Experiments on the use of feature selection and negative evidence in automated text categorization. In: Borbinha, José/Baker, Thomas (Hg.): Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL). Heidelberg, S. 59–68.

Gravino, Pietro et al. (2012): Complex structures and semantics in free word association. In: Advances in Complex Systems 15, 3–4, S. 1–22.

Joulin, Armand et al. (2016): Bag of tricks for efficient text classification. In: arXiv preprint arXiv:1607.01759.

Köhler, Reinhard (1986): Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. (= Quantitative Linguistics 31). Bochum.

Köhler, Reinhard/Altmann, Gabriel (1993): Begriffsdynamik und Lexikonstruktur. In: Beckmann, Frank/Heyer, Gerhard (Hg.): Theorie und Praxis des Lexikons. Berlin/New York.

Mehler, Alexander (2007): Evolving lexical networks. A simulation model of terminological alignment. In: Benz, Anton/Ebert, Christian/van Rooij, Robert (Hg.): Proceedings of the Workshop on Language, Games, and Evolution at the 9th European Summer School in Logic, Language and Information (ESSLLI 2007). Dublin, S. 57–67.

Mehler, Alexander (2008): Large text networks as an object of corpus linguistic studies. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus linguistics. An international handbook of the science of language and society. (= Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science 29). Berlin, S. 328–382.

Mehler, Alexander/Pustylnikov, Olga/Diewald, Nils (2011): Geography of social ontologies. Testing a variant of the Sapir-Whorf Hypothesis in the context of Wikipedia. In: Computer Speech and Language 25, 3, S. 716–740.

Milgram, Stanley (1967): The small-world problem. In: Psychology Today 2, S. 60–67.

Motter, Adilson E. et al. (2002): Topology of the conceptual network of language. In: Physical Review E 65, 6, 065102(R).

Newman, Mark E. J. (2005): Power laws, Pareto distributions and Zipf's law. In: Contemporary Physics 46, S. 323–351.

Newman, Mark E. J. (2010): Networks. An introduction. Oxford.

Rieger, Burghard (1996): Situation semantics and computational linguistics: Towards informational ecology. In: Kornwachs, Klaus/Jacoby, Konstantin (Hg.): Information. New questions to a multidisciplinary concept. Berlin, S. 285–315.

Schnotz, Wolfgang (1994): Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten. (= Fortschritte der psychologischen Forschung 20). Weinheim.

Sigman, Mariano/Cecchi, Guillermo A. (2002): Global organization of the WordNet lexicon. In: Proceedings of the National Academy of Sciences of the United States of America (PNAS) 99, 3, S. 1742–1747.

Stegbauer, Christian/Mehler, Alexander (2011): Positionssensitive Dekomposition von Potenzgesetzen am Beispiel von Wikipedia-basierten Kollaborationsnetzwerken. In: INFORMATIK 2011. Proceedings of 4th Workshop on Digital Social Networks. (= Lecture Notes in Informatics (LNI) P-19). Bonn.

Steyvers, Mark/Tenenbaum, Josh (2005): The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. In: Cognitive Science 29, 1, S. 41–78.

Thelwall, Mike (2009): Social network sites: Users and uses. In: Zelkowitz, Marvin (Hg.): Advances in computers 76: Social networking and the web. Amsterdam, S. 19–73.

Tuldava, Juhan (1998): Probleme und Methoden der quantitativ-systemischen Lexikologie. (= Quantitative Linguistics 59). Trier.

Wang, Chunyan/Ye, Mao/Huberman, Bernardo A. (2012): From user comments to on-line conversations. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12). Peking, S. 244–252.

Watts, Duncan J./Strogatz, Steven H. (1998): Collective dynamics of ‚small-world' networks. In: Nature 393. S. 440–442.

Zipf, George K. (1949): Human behavior and the principle of least effort. An introduction to human ecology. Cambridge.