Rainer Perkuhn / Marc Kupietz

Visualisierung als aufmerksamkeitsleitendes Instrument bei der Analyse sehr großer Korpora

Abstract Sehr große Korpora – wie das Deutsche Referenzkorpus DeReKo – bieten eine breite Basis für die empirische Forschung. Sie bringen aber auch Herausforderungen mit sich, da sich weder Eigenschaften ihrer Zusammensetzung noch derer von Recherche- und Analyseergebnissen mit einfachen Mitteln erschließen lassen. Dafür bedarf es Verfahren geschickter Sortierung, Gruppierung oder des Clusterings, kurzum: strukturentdeckender Methoden. In Kombination mit Visualisierungstechniken kann so die Wahrnehmung bestimmter Eigenschaften und Zusammenhänge unterstützt und die Aufmerksamkeit auf bestimmte Phänomene, ggf. in Anlehnung an präferenzrelationale Befunde, gelenkt werden. Neben der illustrativen Funktion geht es in diesem Beitrag vor allem um das erkenntnisleitende Potenzial derartiger Verfahren in Kombination. Aus verschiedenen Bereichen werden Beispiele gezeigt, die am IDS oder in Kooperationen zum Einsatz kommen, sowohl zur dokumentarischen und reflexiven Kontrolle von Eigenschaften der Korpuszusammensetzung als auch hinsichtlich korpusanalytischer Methodik, um die qualitative Interpretation von Analysebefunden und die Abduktion von Hypothesen stimulierend zu unterstützen.

1. Einleitung

Visualisierung, insbesondere die Forschung zur Visualisierung, ist nicht das primäre Forschungsfeld der Projekte, aus deren Arbeit im Folgenden berichtet wird. Die Schwerpunkte der Projekte sind vielmehr angesiedelt im Umfeld des Deutschen Referenzkorpus (DeReKo) des Instituts für Deutsche Sprache (IDS, Institut für Deutsche Sprache 2016a). Der Bericht soll daraus überblicks- und querschnittsartig Themenbereiche kurz vorstellen, in denen Visualisierungstechniken zum Einsatz kommen. Neben vereinzelten Anwendungen zu illustrativen Zwecken wird – und perspektivisch in noch stärker zunehmendem Maße – ein Aspekt in den Mittelpunkt gerückt, den Schumann/Müller (2000) als Informationsvisualisierung

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

dem Bereich Data Mining bzw. Knowledge Discovery in Databases (KDD) zuordnen. Als Fortführung unseres Forschungsparadigmas, das das Aufspüren präferenz-relationaler Zusammenhänge im Sprachgebrauch zum Ziel hat, setzt diese Art der Visualisierung auf „die Idee, das menschliche visuelle System mit seinen unnachahmlichen Fähigkeiten zum Auffinden von Strukturen und Korrelationen zur Analyse der Informationen zu nutzen" (ebda, S. 342). Ihre Aufgabe in diesem Kontext ist z. T. weniger die einer adäquaten und objektiven Ergebnisdarstellung als vielmehr die eines Hilfsmittels zur Abduktion vielversprechender Hypothesen. Die Güte des gesamten Vorgehens hängt aber nicht nur von einer geeigneten Visualisierung ab, sondern insgesamt von einem harmonischen Zusammenspiel einer angemessenen Datengrundlage, geeigneter Analyseverfahren und der Fähigkeit, die durch die Visualisierung hervorgehobenen Aspekte zu interpretieren. Die Rolle des Interpretierenden übernehmen wir dabei teilweise vollständig selbst. Insbesondere bei der Erschließung der Zusammensetzung des Korpus fließen die Erkenntnisse in die Dokumentation und rückgekoppelt in die weitere Akquisitionsstrategie ein. Für die Entwicklung oder Verfeinerung von Analysemethoden beziehen wir darüber hinaus als Interpretierende aber auch die Methodenanwender für ihre diversen linguistischen Fragestellungen mit ein.

Der stärker rückgekoppelte Einsatz der Visualisierungstechniken ergibt sich in unserem Umfeld aus der extremen Größe des Archivs und einer zum Teil „opportunistischen" Akquisitionsstrategie. Im Vergleich zu anderen Korpora, die gezielt nach vorgegebenen Kriterien aufgebaut werden, wird das Archiv des IDS dynamisch weiterentwickelt und kontinuierlich ausgebaut. Der Erfolg der Bestrebungen hängt dabei von konzeptuell Gewünschtem ab, aber natürlich auch von dem, was rechtlich, finanziell und technisch – grundsätzlich von den gegebenen Kapazitäten her – machbar ist (vgl. Kupietz/Schmidt 2015). Da das Archiv als Ur-Stichprobe dienen soll, ist eine ungleichmäßige Zusammensetzung der Daten weniger relevant. Jeder Nutzende kann eine Arbeitsversion aus dieser Datensammlung als ein sogenanntes virtuelles Korpus zusammenstellen (vgl. Kupietz et al. 2010). Trotzdem ist es natürlich hilfreich, möglichst viele Informationen über die Zusammensetzung der Daten zu sammeln, einerseits, um den Nutzenden bei der Definition des Arbeitskorpus zu unterstützen, andererseits, um Bereiche aufzuspüren, die noch besser ausgebaut werden könnten.

Zu Anfang des nächsten Abschnitts wollen wir aber zunächst am Beispiel eines als ausgewogen geplanten Korpus zeigen, wie Visualisierung im primär illustrierenden Sinne eingesetzt werden kann. Anstelle von hierarchischen Strukturen, die für didaktische und/oder dokumentarische Zwecke syntaktische oder morpho-syntaktische Zusammenhänge veranschaulichen oder auch die Genese und Verwandtschaft von Sprachfamilien (vgl. Bubenhofer in diesem Band, S. 63), deuten wir nur einige Diagrammtechniken für Verteilungen an, wie sie in den meisten Tabellenkalkulationsprogrammen integriert angeboten werden.

2. Eigenschaften/Zusammensetzung des Archivs

Korpora werden vielfach nach bestimmten Vorgaben geplant, etwa dass (zumindest abschnittsweise) ein bestimmter Umfang angestrebt wird oder dass die Zusammensetzung nach bestimmten Kriterien (wie Textsorte o.Ä.) einer bestimmten Verteilung folgt. Für das DWDS-Kernkorpus ist laut Webseite (http://www.dwds.de/ressourcen/kernkorpus/, letzter Zugriff am 22. August 2016) ein Umfang von 10.000.000 Token je Dekade des 20. Jahrhunderts und eine Zusammensetzung von Belletristik : Gebrauchsliteratur : Wissenschaft : Zeitung im Verhältnis 28,42% : 21,05% : 23,15% : 27,36% vorgegeben (verschriftlichte gesprochene Sprache ist bei diesen Zusammenstellungen herausgenommen). Ohne die Plausibilität dieser Vorgaben diskutieren zu wollen, zeigen wir hier eine selbst erzeugte Grafik für die geplante Verteilung nach Textsorten, ein sogenanntes Tortendiagramm (s. Abb. 1).

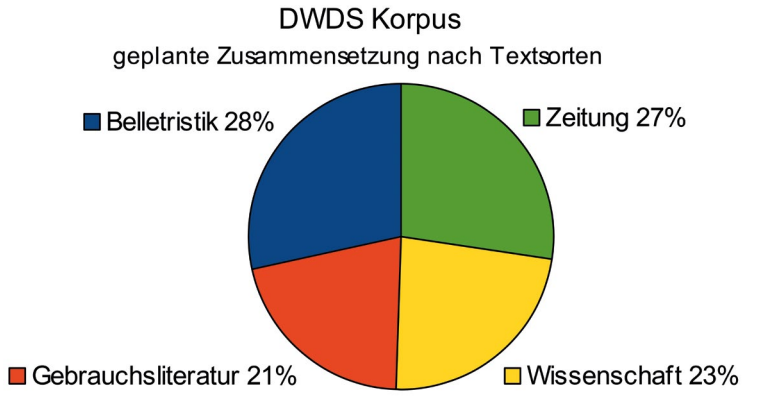

Abb. 1: Geplante Verteilung nach Textsorten (DWDS-Kernkorpus) (Copyright IDS).

Die reine Darstellung der SOLL-Werte liefert nur wenig zusätzliche Information – am wenigsten bei einer Darstellung der Umfänge pro Jahrzehnt. Interessant wird es aber auch in diesem Umfeld schon, wenn die geplanten und die tatsächlich erreichten Werte nebeneinandergestellt präsentiert werden. Eine Darstellung der beiden Dimensionen getrennt voneinander lässt sich noch gut umsetzen (vgl. Abb. 2).

Der Versuch, beide Dimensionen zusammenzuführen, zeigt ansatzweise die Grenzen des Machbaren auf (vgl. Abb. 3). Um die Bereiche lokalisieren zu können, die auf der Webseite als unterbesetzt erwähnt werden, braucht es ein glückliches Händchen für die Wahl der Perspektive bzw. der Anordnung in der Tiefe zwischen Vorder- und Hintergrund.

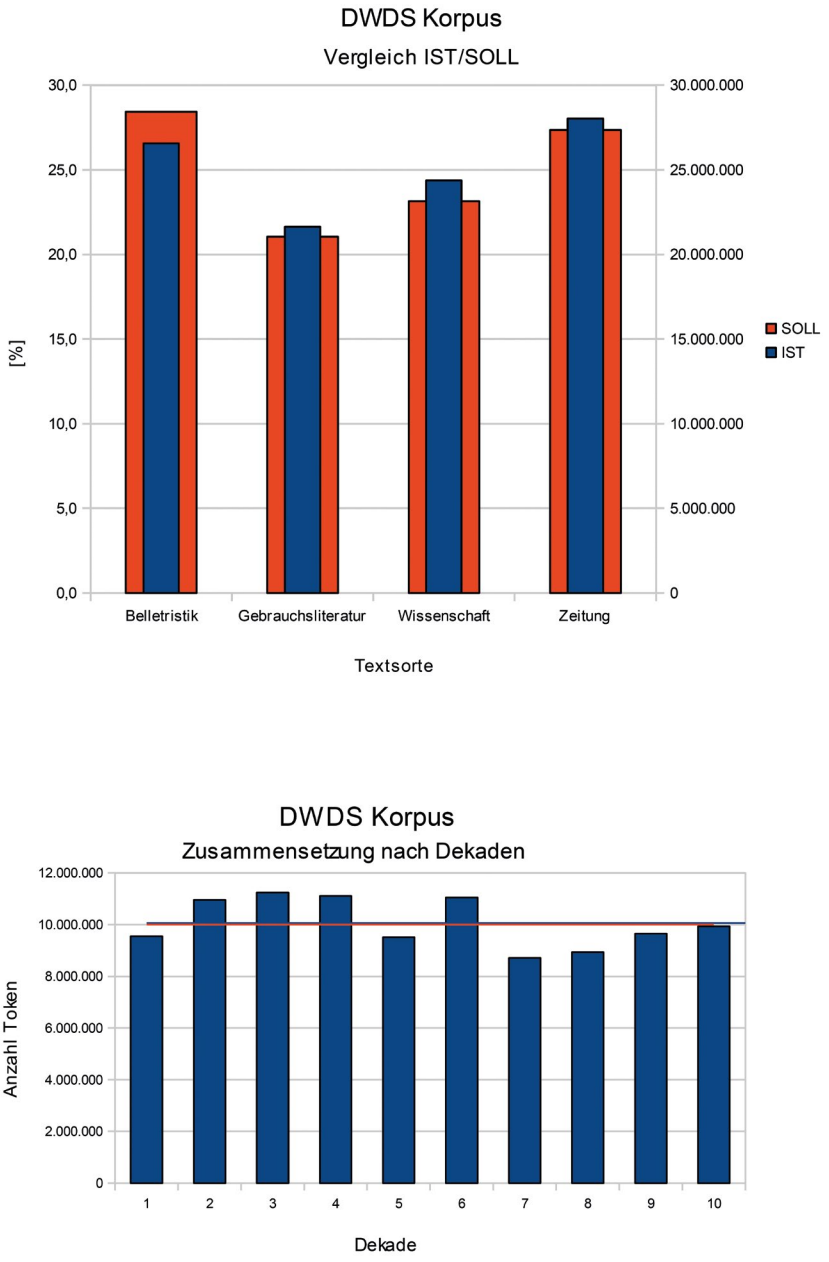

Abb. 2: Vergleich IST/SOLL getrennt nach Textsorte bzw. Umfang (DWDS-Kernkorpus) (Copyright IDS).

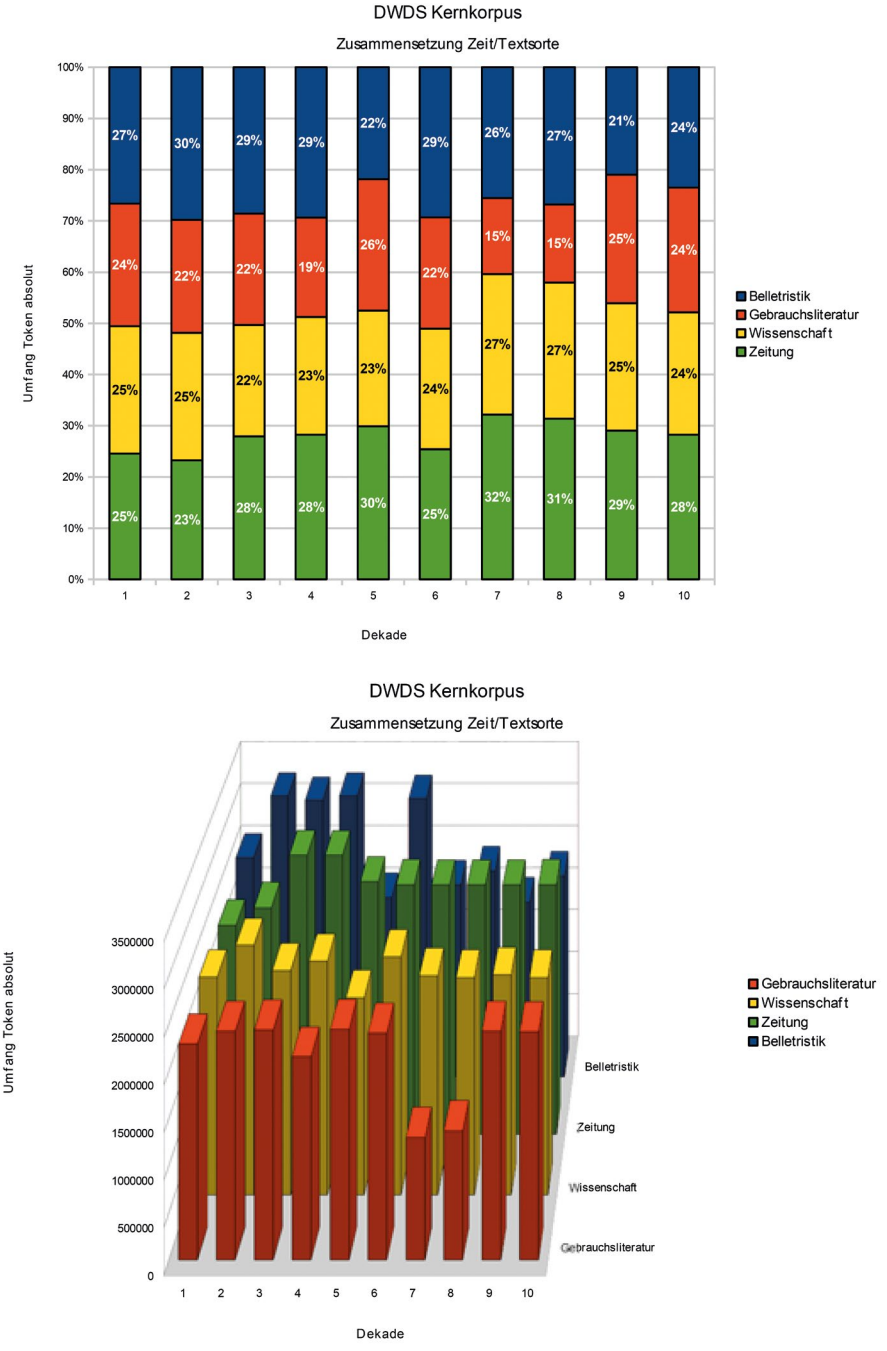

Abb. 3: Differenzierung der Dekaden-IST-Werte nach Textsorte und Umfang, gestapelt vs. dreidimensional (DWDS-Kernkorpus) (Copyright IDS).

Den Umfang der Korpora kumuliert nach Dekaden vor- und anzugeben, eröffnet einen schwer einzuschätzenden Spielraum für die Verteilung auf die einzelnen Jahre von absoluter Gleichverteilung bis hin zu extremen Schieflagen. Um die Zusammensetzung des IDS-Archivs zu dokumentieren, beziehen wir uns auf die Einheit „pro Publikationsjahr“ (vgl. Abb. 4).

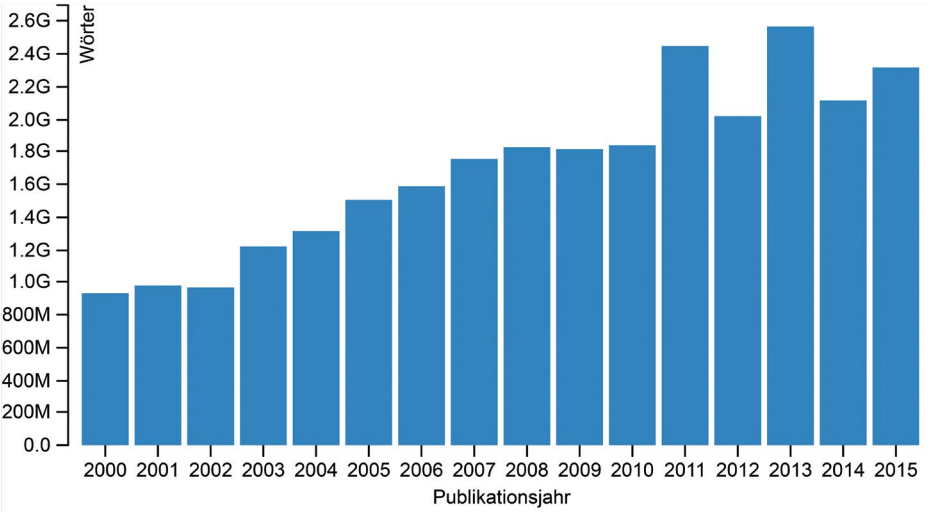

Abb. 4: DeReKo-Umfang pro Publikationsjahr ab 2000 (Copyright IDS).

Die Botschaft dieses Diagramms ist nicht die Dokumentation des Vergleichs mit einem geplanten Ziel. Auch wenn Recherchen im gesamten Archiv durchaus sinnvoll sein können, soll dieser ansteigende Verlauf vor allem dafür sensibel machen, dass Trefferzahlen pro Jahr nur in engen Bereichen absolut miteinander verglichen werden können. Auch die Relativierung auf den jeweiligen Jahresumfang kann bei extremen Abweichungen zu Überraschungen führen. Für Untersuchungen, die auf die Dimension Zeit schauen, ist es in vielen Fällen ratsam, ein virtuelles Korpus zu definieren, dessen Umfang je nach zugrunde gelegter Zeiteinheit nicht allzu sehr schwankt (vgl. Abschnitt zu Zeitverläufen). Interessant hierfür sind vor allem auch Zeitungsdaten, da sie sich – im Vergleich zu internetbasierter Kommunikation – gut datieren lassen. Zudem erscheinen Zeitungstexte – im Kontrast zu Belletristik – in einem kurzen Takt, und Textproduktion und -publikation liegen zeitlich nah beieinander. Sie bilden verhältnismäßig gut und vor allem eben zeitnah den allgemeinen Sprachgebrauch ab, zugegebenermaßen stark vom Zeitgeschehen, teilweise mit einem gewissen Lokalkolorit, beeinflusst.

2.1 Geografische Verteilung der Archiv-Quellen

Zeitungsdaten, auf die wir uns in diesem Abschnitt konzentrieren wollen, weisen regionale Unterschiede auf. In geringem Maße ergibt sich dies sicher auch aus dem umgebungsbedingten Substandard der allgemeinen Sprache (wobei vermutlich die wenigsten Redakteure „unverfälschte" gebürtige Sprecher der jeweiligen Region sind). Vielfach sind aber gerade die Themen, über die geschrieben wird, und somit Wortwahl und das Vokabular stark durch die Region geprägt. Für Untersuchungen, die gerade gezielt darauf eingehen wollen oder die Effekte in einem größeren Zusammenhang gedämpft sehen möchten, wird häufig der Wunsch an uns herangetragen, mehr Daten aus möglichst vielen verschiedenen Regionen zur Verfügung zu stellen. In einem benachbarten IDS-Projekt „Deutsch heute", das sich zum Ziel gesetzt hat, flächendeckend Audio-Aufnahmen für ein Korpus gesprochener Sprache zu erheben, wurde dieses Bestreben Teil der Projektplanung (vgl. Abb. 5).

Demhingegen wollen wir für die schriftsprachlichen Korpora im Allgemeinen nicht zur Textproduktion auffordern, sondern uns bei Quellen bedienen, die in einem „natürlichen" Prozess Texte gestalten. Eine entsprechende Karte wäre deutlich dünner besetzt als die in Abb. 5 gezeigte, liegt uns aber nicht vor. Anstelle der Produktionsorte haben wir die Verlagsorte unserer Textspender auf eine Karte geplottet (vgl. Abb. 6). Sie stellt einen Ausschnitt aus Mitteleuropa dar, im Kern bestehend aus Deutschland und den Nachbarländern, in denen Deutsch zumindest als Minderheitensprache gesprochen wird. Dazu haben wir auf die Orte Kreissignaturen platziert, die in ihrer Größe dem Umfang der Texte entsprechen, die aus den Quellen des gleichen Ortes in unser Archiv eingespeist sind.

Um den Erfolg einer der letzten Akquisitionsbestrebungen zu dokumentieren, sind die Kreise in unterschiedlichen Farben jeweils für die Zeit vor und nach der Aktion markiert. Wie die Karte zeigt, sind sehr viele Lücken geschlossen worden. Sie zeigt aber auch, dass auch vor der Akquisition der Schwerpunkt gar nicht so sehr durch die Großregion Mannheim geprägt war (was dem Archiv gelegentlich unterstellt wird) – und dass noch einige Lücken geblieben sind. Neben vielen anderen Aspekten bleibt bei dieser Darstellung auch unberücksichtigt, wie sich die Bevölkerung auf die Regionen verteilt und wie hoch etwa die Auflagenstärke der Printmedien ist, um womöglich einen rezeptiven Wirkungsgrad abschätzen zu können.

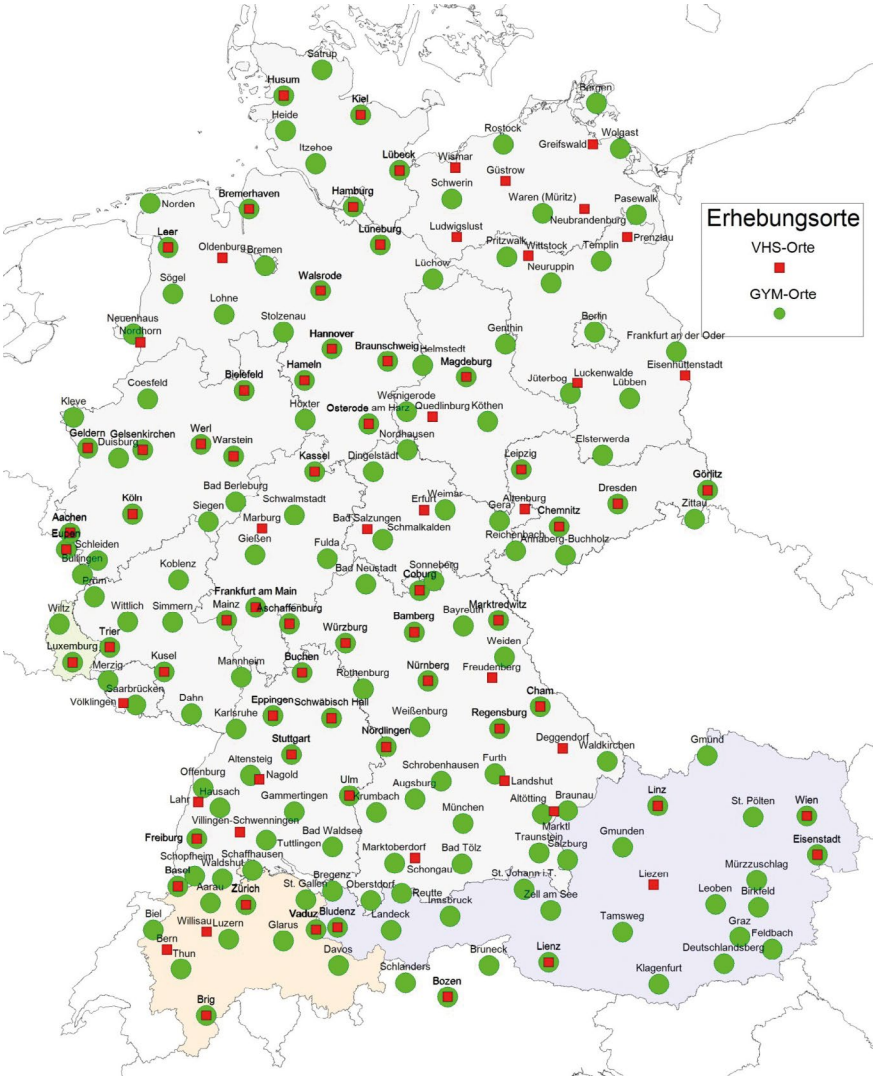

Abb. 5: Geplante Orte der Datenerhebung (Deutsch heute) (Copyright IDS, http://www1.ids-mannheim.de/fileadmin/prag/AusVar/Deutsch_heute/Erhebungsorte_DH_70.jpg).

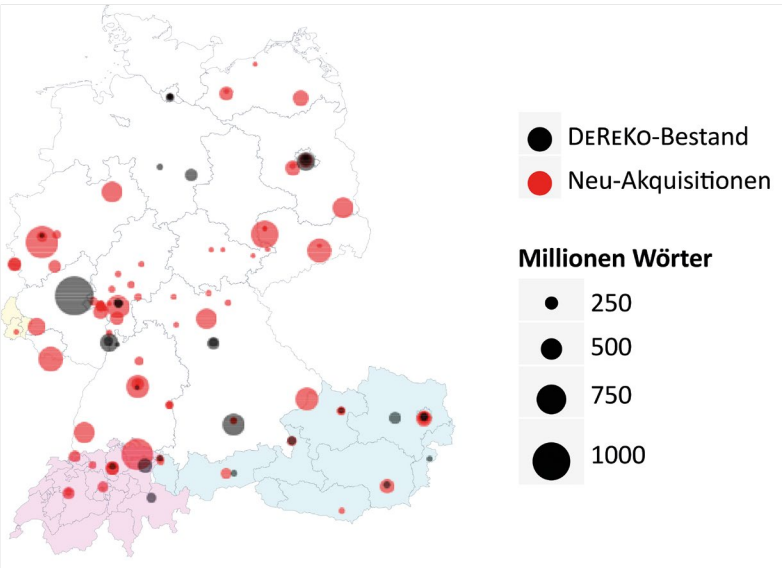

Abb. 6: Regionale Verteilung der bestehenden DeReKo-Zeitungsquellen und der in 2013 neu akquirierten Quellen. (Copyright IDS)

2.2 Textdublettenerkennung

Aus verschiedenen Gründen kann es vorkommen, dass nahezu identische Texte zur Aufnahme in das Archiv bereitgestellt werden. Dies kann versehentlich zustande kommen aufgrund von Überschneidungen des redaktionellen Prozesses mit der Übernahme der Daten in unsere Arbeitsabläufe. In anderen Fällen kann es sich um schematisch angelegte Texte handeln, wie etwa Wetterberichte, Kinoankündigungen oder -rezensionen, oder auch um Variationen von Agenturmeldungen. Während im ersten Fall die Einschätzung relativ eindeutig ausfallen sollte, dass es sich dabei um (echte) unerwünschte Textdubletten handelt, ist dies bei den anderen Fällen weniger klar. Je nach Fragestellung (z. B. Untersuchung der Produktion vs. Untersuchung der Rezeption) können gerade die Variationen von Serientexten zum Gegenstand der Betrachtung werden. Die Ähnlichkeiten zwischen den Texten unseres Archivs werden mithilfe eines Dublettenerkennungsverfahrens (Kupietz 2005) ermittelt, dem zum Zweck einer einfacheren Qualitätsprüfung eine Visualisierung nachgelagert ist. Diese basiert auf einer einfachen Alignierung und farblichen Hervorhebung der Gemeinsamkeiten und Unterschiede, setzt damit aber genau die erforderliche Expressivität um (vgl. Abb. 7).

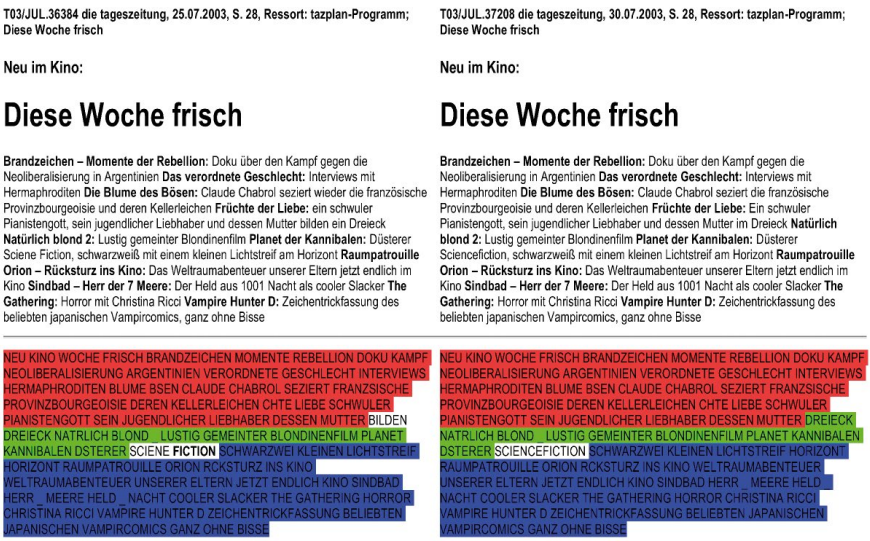

Abb. 7: Visualisierung von Textdubletten (DeReKo) (Copyright IDS).

Die Dublettenerkennung wird in erste Linie dazu eingesetzt, um die Beziehungen zwischen den Texten zu dokumentieren und als Metadaten festzuhalten. Nur in eindeutigen Fällen werden die Texte tatsächlich aussortiert. Perspektivisch soll es den Nutzenden virtuell ermöglicht werden, diese Entscheidung für die Gesamtmenge der markierten Texte über einen einstellbaren Schwellwert selbst zu treffen.

2.3 Inhaltliche Eigenschaften der Archiv-Quellen

Neben der stärksten Ausprägung der Ähnlichkeit von Texten als nahezu vollständige Übereinstimmung gibt es weitere Beweggründe, um Texte oder ganze Korpora bezüglich weicherer Aspekte zu vergleichen. Zum Beispiel: Entstammen Texte etwa ähnlichen Registern oder handeln sie von ähnlichen Themen? Bereits beim Vergleich unseres Archivs mit dem zum Zeitpunkt der Untersuchung größenmäßig vergleichbaren webbasierten Korpus deWaC (Baroni et al. 2009) hat sich gezeigt, wie aussagekräftig Vergleiche schon allein auf der Ebene des lexikalischen Inventars, also des Vokabulars, sind. In Fortführung dieser Gedanken haben wir alle Teilkorpora unseres Archivs mithilfe eines Maßes von Kilgarriff (2001) verglichen, das auf den vorderen Ausschnitten der frequenzsortierten Vokabulare basiert (Kupietz et al. 2012). Hintergrund der

Studie war der Versuch, wiederum aus einer webbasierten Ressource durch optimierende wiederholte Stichprobenziehungen eine Sammlung von Texten („litDeWaC“) zusammenzustellen, die einem echten Literaturkorpus („lit“: zusammengesetzt aus allen Belletristik-Teilkorpora aus DeReKo) möglichst nahekommt. In Abb. 8 sind die Abstände zwischen den Korpora mithilfe nichtmetrischer multidimensionaler Skalierung (NMDS) zweidimensional projiziert (vgl. Cox & Cox 2001). Sie zeigt durchaus plausible räumliche Anordnungen, wobei sich in diesem Fall die Semantik des Hintergrunds erst aus der Interpretation der Gruppen ergibt.

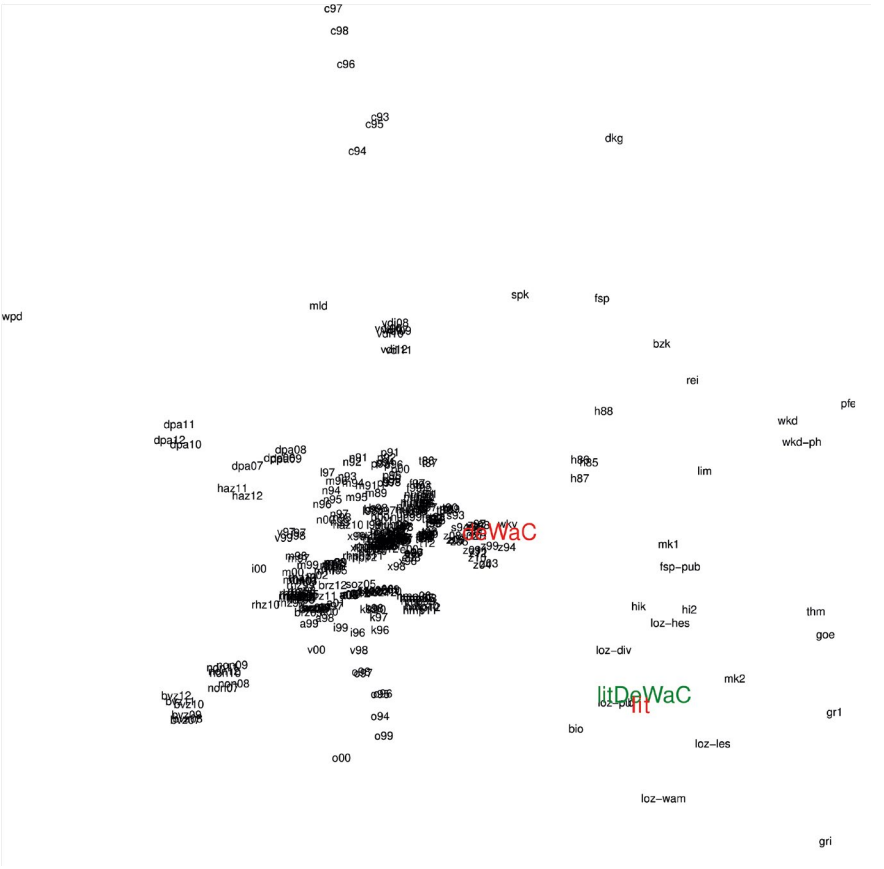

Abb. 8: Ergebnisdarstellung der Extraktion eines Pseudo-Literaturkorpus litDeWaC nach dem Vorbild des Korpus lit aus deWaC vor dem Hintergrund aller DeReKo-Teilkorpora (weitere Abkürzungen siehe http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html (interaktiv)). Die Ähnlichkeitsmatrix der Teilkorpora wurde wie in Kilgarriff (2001) mit Spearmans Rangkorrelationskoeffizienten berechnet. Zur Projektion der Ähnlichkeitsmatrix auf eine Karte wurde NMDS verwendet. (Copyright IDS)

Dabei ist weniger die Positionierung des neuen Korpus litDeWaC überraschend, da dessen Zusammenstellung ja gerade auf die Nähe zu bekannten Literaturdaten (lit) ausgerichtet war. Beeindruckend ist aber schon das Gesamtbild, das sich aus der Anordnung der verschiedenen Teilkorpora abzeichnet, beispielsweise im unteren rechten Bereich die Nähe der LOZ-Korpora (**L**iteratur, **O**riginalsprache: Deutsch, des **Z**wanzigsten Jahrhunderts) untereinander und zu den Korpora einzelner Schriftsteller (thm – Thomas Mann, goe – Goethe), aber auch zu den Märchen der Gebrüder Grimm (gri). Insgesamt deutet sich eine Topographie anhand verschiedener Texteigenschaften wie Register, Medium, Textsorte, Thematik u. Ä. an: Im mittleren rechten Bereich schimmern etwa Aspekte mündlicher Interaktion durch, durch das gemeinsame Arrangement der Korpora „Reden und Interviews“ (rei), dem Wendekorpus (wkd) und dem Pfeffer-Korpus (pfe).

3. Ergebnisübersichten der Treffermengen

Neben den bisher diskutierten, stärker kumulativen Betrachtungen von Texten oder Korpora finden Visualisierungstechniken auch bei eher wortbezogenen Auswertungen Anwendung. Bereits seit den Anfängen der Korpuslinguistik werden die auf die Suchobjekte passenden Texteinheiten bei der Beleganzeige durch Textattribute (fett, Farbe) hervorgehoben. Bei der auf einen unmittelbaren Vergleich ausgerichteten kompakten Darstellung einer Konkordanz wird dies zusätzlich durch die positionelle Anordnung unterstützt: Mehrere Treffer werden zeilenweise untereinander so angeordnet, dass die gefundenen Objekte mittig aligniert untereinander platziert werden. Nach links und rechts werden dann – je nach Platz aufgrund des genutzten Mediums – gleichermaßen so viele Zeichen aufgefüllt, wie eine Zeile aufnehmen kann. Während man in den frühen Phasen aufgrund der gegebenen Rahmenbedingungen dafür auf nichtproportionale Schriften zurückgegriffen hat, lässt sich dies mit heutigen Mitteln auch mit proportionalen Schriften quasi tabellenartig darstellen.

Vor allem für große Treffermengen ist es oft hilfreich, sich zunächst anhand eines Ergebnisüberblicks einen ersten Eindruck zu verschaffen. Wie viele Treffer gibt es etwa pro Jahr, wie viele pro Quelle oder je Region? So wie oben aber bereits angedeutet, sind absolute Häufigkeiten nur bei annähernd gleich großen Schnitten vergleichbar. Aber auch relative Frequenzen verlieren dann ihre Aussagekraft, wenn die Schnitte um mehrere Größenordnungen auseinanderliegen. Das Recherchesystem des IDS Cosmas II (Bodmer Mory 2014, Institut für Deutsche Sprache 2016b) bietet Ergebnisübersichten nur in tabellarischer Form an. Diagramme wie die hier gezeigten können mit anderen Softwaretools auf der Grundlage dieser Angaben in einem weiteren

nachgelagerten Bearbeitungsschritt erzeugt werden, wie das Balkendiagramm für die Häufigkeiten pro Jahr oder das Tortendiagramm für die Verteilung nach Quellen, Themen oder Region (vgl. Abb. 9 und 10).

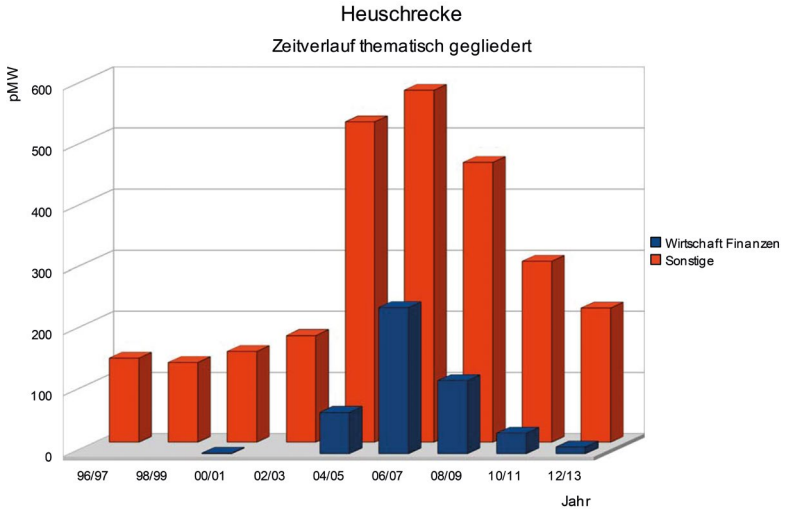

Abb. 9: Thematische Verteilung der Ergebnismenge zu dem Lemma „Heuschrecke“ (DeReKo/Cosmas II) (Copyright IDS).

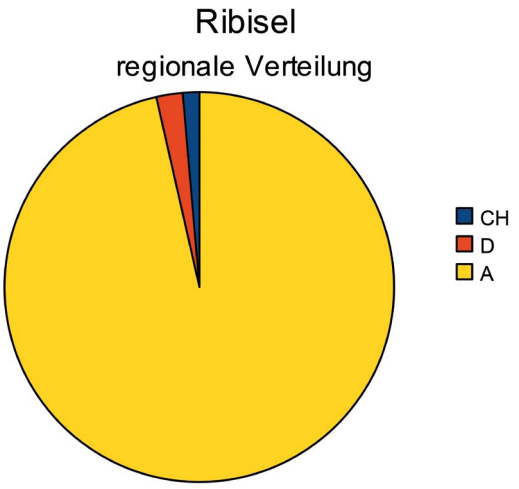

Abb. 10: Regionale Verteilung der Ergebnismenge zu dem Lemma „Ribisel“ (DeReKo/Cosmas II) (Copyright IDS).

Die Darstellung der regionalen Verteilung für das Wort *Ribisel* zeigt dabei einen eindeutigen Befund, der unabhängig von absoluten oder relativen Häufigkeiten ist: Dieses Wort wird fast ausschließlich im österreichischen Deutsch verwendet.

3.1 Ergebnisübersicht der Treffermengen nach Zeit (Zeitverlaufsgrafiken)

Die relativen Vorkommen einer sprachlichen Einheit pro Zeitabschnitt genießen für verschiedene Fragestellungen eine besondere Aufmerksamkeit. Eine geeignete Datengrundlage vorausgesetzt, lassen sich Häufigkeitsverschiebungen als Veränderungen im Sprachgebrauch deuten: als Indizien für die Lebendigkeit bestimmter Diskurse oder, im einfachsten Fall, für das Aufkommen neuer (oder das Aussterben alter) Wörter. Für Wörter, die in ihrer Form neu in einem bestimmten Zeitabschnitt zu beobachten sind, sogenannte Neulexeme (z. B. der Jahre 2000 bis 2010), generieren wir auf der Grundlage eines speziell dafür zusammengestellten virtuellen Korpus Zeitverlaufsgrafiken (Lüngen/Keibel 2013, vgl. Abb. 11), die über das Online-Informationssystem OWID, Rubrik Neologismenwörterbuch, öffentlich angeboten werden.

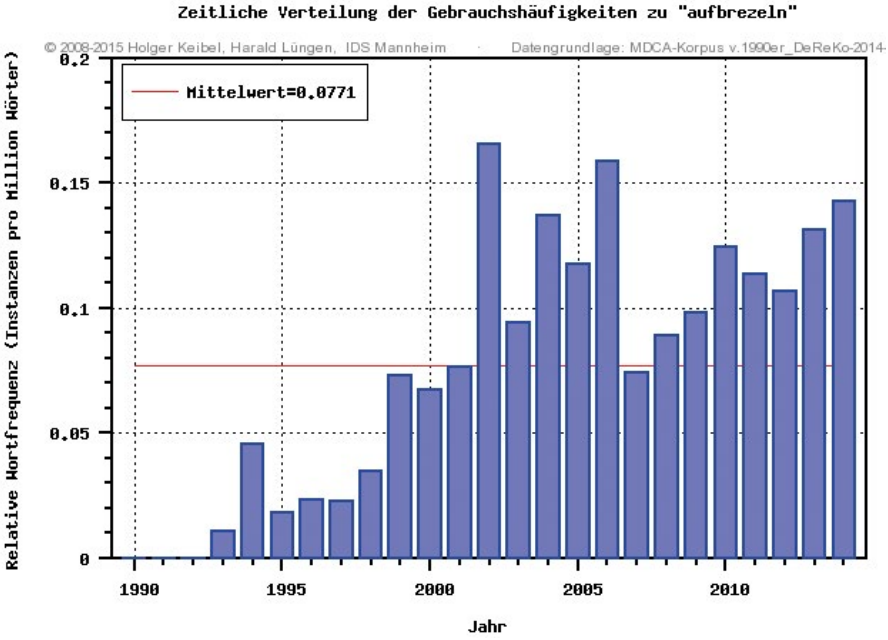

Abb. 11: Zeitverlaufsgrafik eines Neulexems (DeReKo/OWID) (Copyright IDS, http://www.owid.de/artikel/298252, http://www.ids-mannheim.de/kl/neoplots/owid/298252.html).

Neben dieser Visualisierung als Bestätigung eines Befundes setzen wir auf den Gesamtbestand des Korpus auch Verfahren auf, um Neulexem-Kandidaten aufzuspüren (Keibel et al. 2010). Visualisierung spielt in diesem Zusammenhang aber nur eine untergeordnete Rolle, um didaktisch zu illustrieren, dass Neulexem-Kandidaten als notwendige Bedingung einen Zeitverlauf aufweisen müssten, der auf eine schematisch angedeutete Zeitverlaufsschablone passt. Wörter, die eine zusätzliche, neue Bedeutung angenommen haben (Neubedeutungen), lassen sich mit diesen einfachen quantitativen Verfahren nicht aufspüren und somit darauf aufbauend auch nicht visualisieren.

3.2 Ergebnisübersicht der Treffermenge nach Kontext (Kookkurrenz)

Eine andere Form, die Treffermenge eines Suchausdrucks zu sortieren, basiert auf der Idee, dafür die Wörter in der unmittelbaren textuellen Umgebung heranzuziehen, die besonders systematisch in der Nähe des Suchausdrucks vorkommen. Für die Berechnung dieser typischen Wortverbindungen bietet das IDS das parametrisierbare Verfahren der Kookkurrenzanalyse (Belica 1995) an, das auch in dem System Cosmas II integriert ist. Die Präsentation des Analyseergebnisses ähnelt in diesem System, wenn auch mit höherer Komplexität, den bereits bekannten tabellarischen Darstellungen. Gerade für diesen Ergebnisüberblick sind aber speziellere Zugangsformen wünschenswert, da die Skala der Sortierung nicht vorgegeben ist, sondern sich erst aus der Analyse heraus ergibt. Ein Vorgehensmodell für die Erschließung dieser Ergebnisstrukturen wurde für ausgewählte Aspekte als Prototyp operationalisiert (Perkuhn 2007a/b). Für dieses Werkzeug sind die zwei zentralen Aspekte die Visualisierung der Gesamtstruktur und die Möglichkeit, in Form von Annotationen gewonnene Erkenntnisse festhalten zu können, wobei wir auf den zuletzt genannten Aspekt hier nicht weiter eingehen werden. Die Darstellung der Gesamtstruktur steht vor der Herausforderung, dass bei den üblicherweise verwendeten Medien nur begrenzt Raum zur Verfügung steht. Der Ausweg, nur einen kleinen ausgewählten Ausschnitt anzubieten – z. B. über scrollbare Fensterausschnitte wie bei der Cosmas II-Präsentation –, steht im Widerspruch zur Forderung einer Gesamtsicht. Um alle Elemente der Gesamtstruktur anzeigen zu können, müssten diese jedoch so weit verkleinert werden, dass sie quasi nicht mehr zu erkennen sind. Ein Ausweg aus diesem Dilemma bietet ein einfacher Ansatz, der diese Übersicht von „Miniaturen“ um interaktiv angebotene Möglichkeiten ergänzt. So können beispielsweise einzelne Elemente oder Ausschnitte ausgewählt werden, die vergrößert und dadurch erkennbar dargestellt werden (analog Abb. 15). Die Interaktionsmöglichkeiten sind zurzeit über Maus-Bewegungen und -Aktionen umgesetzt. Die sogenannten Fokus&Kontext-Techniken (vgl. Lamping et al. 1995) könnte

man als Variante dieses Ansatzes verstehen, bei denen ein ausgewählter Ausschnitt zu Anfang bereits gesetzt ist. Das, was im Fokus steht, wird gut erkennbar dargestellt eingebettet in dessen Kontext, d. h. vor dem Hintergrund der ggf. zur Unkenntlichkeit verkleinerten Gesamtstruktur. Unsere Präsentation bedient sich hierzu eines hyperbolischen Modells (vgl. Abb. 12). Dabei wird eine hierarchische Struktur sozusagen auf eine Halbkugel, die von oben betrachtet wird, projiziert. Hätten wir die hierarchische Struktur plan in der Ebene radial um die Wurzel herum aufgezeichnet, wären die Verbindungslinien zwischen den verschiedenen Hierarchiestufen gleich lang. Dadurch, dass dieses netzartige Gebilde quasi über die Halbkugel gelegt wird, wirken die Linien in der Nähe der Wurzel fast so lang wie in der Ebene, während die weiter außen liegenden quasi perspektivisch in der dritten Dimension „nach hinten“ nahezu verschwinden. Die Größen der verbundenen Objekte werden entsprechend angepasst, so dass die oben liegenden (der oberste und sein enger Kontext) gut zu erkennen sind, die weiter außen liegenden im Normalfall aber unkenntlich klein sind.

Die Projektion auf die Halbkugel ist der Trick, um auf dem begrenzt zur Verfügung stehenden Raum eine beliebig komplexe hierarchische Struktur abbilden zu können. Der äußere Rand sammelt die Objekte der untersten Ebene der Hierarchie mit der maximalen Verzweigung der Verästelung. Jedes Element der Struktur kann aber zum obersten Punkt der Halbkugel verschoben, somit

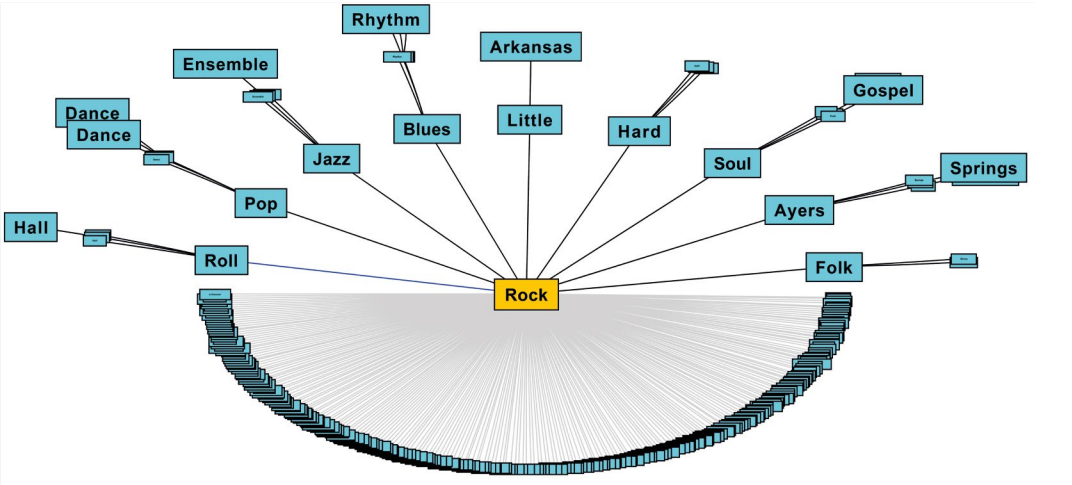

Abb. 12: Hyperbolische Repräsentation des Kookkurrenzprofils des Lemmas „Rock“ (Vicomte) (Copyright IDS).

in den Fokus gerückt und gut erkennbar dargestellt werden. Auf diesem Weg lässt sich nach und nach die gesamte Struktur erschließen.

Bei unseren Analyseergebnissen handelt es sich meist um sehr flache, aber sehr schnell stark verzweigende Strukturen. Die herkömmliche Art der Fokusverschiebung ermöglicht eine nur sehr kleinschrittige Navigation durch die Gesamtstruktur. Neben der interaktiven Vergrößerung von Elementen über die Maus-Bewegung bietet unser Ansatz deshalb eine überlagernde Fokussierung auf einen Ausschnitt der ersten Ebene der Hierarchie an. In der oberen Hälfte der Halbkugel wird ein sehr kleiner Ausschnitt auf den zur Verfügung stehenden Platz gestreckt, während sich der Rest der Gesamtstruktur mit der unteren Hälfte begnügen muss. Zusätzlich zu den anderen Fokussierungsmöglichkeiten lässt sich die Gesamtstruktur drehen, sodass nach und nach jeder Abschnitt der ersten Hierarchieebene gut erkennbar dargestellt wird. Einzelne Elemente können hervorgehoben werden, was auch fixiert für die Darstellung im unteren Bereich beibehalten wird.

Untersuchungen (z. B. in Storjohann 2007a/b, Schnörch 2015) haben gezeigt, dass diese Visualisierungsform wie auch die hier nicht näher ausgeführte Möglichkeit der Annotation die Erschließung der Gesamtstruktur für verschiedene Fragestellungen gut unterstützt und sie dadurch nachvollziehbar dokumentiert wird.

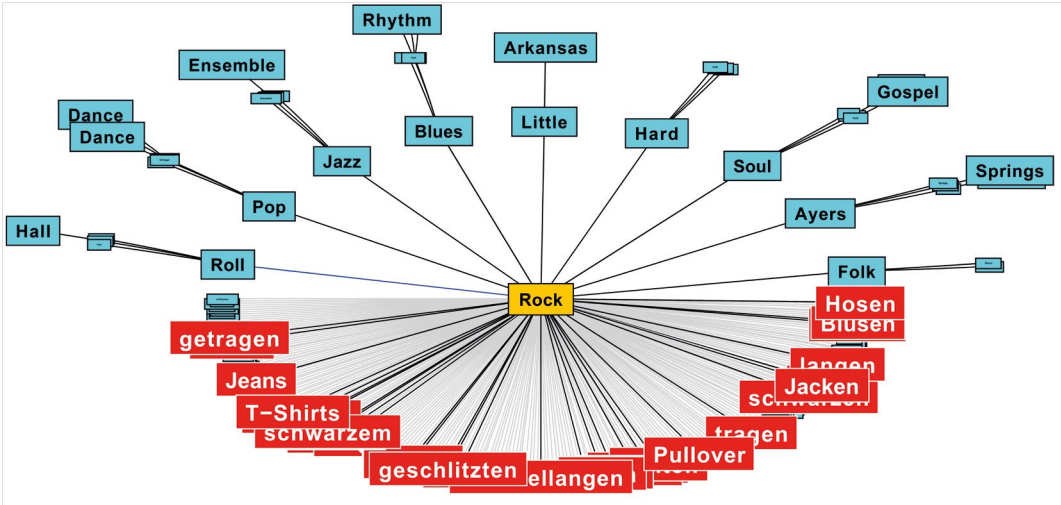

Abb. 13: Hervorgehobene Übereinstimmungen des Profils des Lemmas „Rock“ im Vergleich zum Lemma „Kleid“ (Vicomte) (Copyright IDS).

3.3 Kookkurrenz und Zeit

Während sich Neulexemkandidaten über einen typischen Zeitverlauf charakterisieren lassen, ist dieser für viele Neubedeutungen eher unscheinbar. Die Ausschläge der Amplituden sind nicht auffälliger als bei bedeutungsstabilen Wörtern, bei denen diese durch Veränderungen in der Zusammensetzung der Daten, durch das Weltgeschehen oder einfach durch „Modeerscheinungen" bedingt sind. Vor allem fehlt aber das Gegenstück für eine einfache quantitative Messung: Anders als bei einem Neulexem, dessen Häufigkeit vor dessen Initiation vernachlässigbar klein gewesen sein muss, kann das Aufkommen einer neuen Bedeutung nicht ohne Weiteres zu einem bestimmten Zeitpunkt zahlenmäßig erfasst werden. Einen Anhaltspunkt gibt es aber dennoch: Wenn sich Bedeutungsaspekte im Kookkurrenzverhalten eines Wortes niederschlagen, so sollten sich auch Bedeutungsveränderungen in Änderungen des Kookkurrenzverhaltens abzeichnen. Um die oben angedeuteten alternativen Gründe für Veränderungen ein wenig kontrollieren zu können, achten wir für entsprechende Untersuchungen verstärkt auf eine durchgängig homogene Zusammensetzung des zugrunde gelegten virtuellen Korpus. Der Preis, den wir dafür zahlen, besteht in kleineren Treffermengen. Wenn diese so gering ausfallen, dass Kookkurrenzanalysen für einzelne Jahrgänge zu unergiebig sind, sind die Definitionen der Zeitscheiben auf mehrere Jahrgänge zu erweitern. Ein weiterer Nachteil des kritischen Mindestdatenumfangs besteht darin, dass lediglich Untersuchungen für ca. die letzten 25 Jahre umsetzbar sind. Aufgrund noch vieler anderer Unwägbarkeiten können wir auch noch keine Methode präsentieren, die auffälligen Bedeutungswandel aufdeckt. Wir benutzen in diesem Zusammenhang allerdings Visualisierungen, die die Plausibilität des Ansatzes unterstützen. Die Ergebnisse explorativer Untersuchungen zeigen, dass die Rangverläufe im Kookkurrenzverhalten bestimmter Partnerwörter, die die Neubedeutungen indizieren, Parallelen zu den Zeitverläufen von Neulexemen aufweisen (vgl. Abb. 14).

Für andere Fragestellungen lässt sich derselbe Ansatz verwenden, zum Beispiel, um die Entwicklung auffälliger Diskurse nachzuzeichnen (vgl. das Beispiel Konflikt in Perkuhn/Belica 2016). Das automatische Erkennen derartiger Indikatoren gestaltet sich zurzeit noch schwierig, da einerseits das gesamte Kookkurrenzprofil „ständig in Bewegung" ist und andererseits für eine genauere Isolierung tatsächlich markanter Veränderungen mehr Messpunkte, somit breitere homogene Datengrundlagen erforderlich wären.

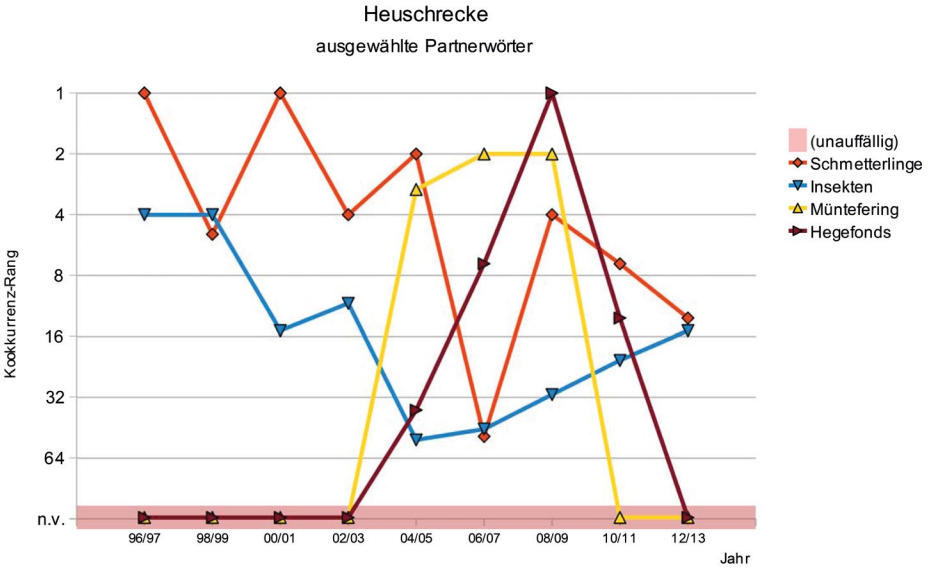

Abb. 14: Rangverlauf auffälliger Partnerwörter des Lemmas „Heuschrecke“ (DeReKo) (Copyright IDS).

3.4 Kookkurrenzverhalten im Vergleich

Wir haben im vorherigen Abschnitt beschrieben, wie das Koookkurrenzverhalten eines Wortes im Laufe der Zeit anhand zeitscheibenbezogener Kookkurrenzprofile analysiert werden kann, um den Bedeutungswandel eines Wortes zu dokumentieren. Der elementare Schritt war hierbei der Vergleich zweier Kookkurrenzprofile. Ausgangspunkt des Vergleichs war in diesem Fall dasselbe Wort, die Profile wurden auf der Grundlage unterschiedlicher virtueller Korpora ermittelt. Übertragen wir jetzt die Vorgehensweise auf die Profile zweier verschiedener Wörter, die auf der Grundlage desselben Korpus erstellt wurden, so können wir hoffen, etwas über die Beziehung zwischen den Wörtern zu lernen: Wörter, die eine enge semantische Beziehung vermuten lassen, sollten viele Gemeinsamkeiten im Kookkurrenzverhalten aufweisen. Eine Erweiterung zu dem oben beschriebenen Vorgehensmodell für die Erschließung von Kookkurrenzanalysen visualisiert die Verteilung der Partnerwörter bei einem Vergleich von bis zu drei Bezugswörtern, wobei topographisch zwischen den dedizierten und den paarweise (ggf. auch allen drei) gemeinsamen Partnern unterschieden wird. Erstere werden vertikal nach Rang, letztere nach gemitteltem Rang und horizontal näher bei dem Bezugswort mit dem höheren Rang angeordnet (vgl. Abb. 15).

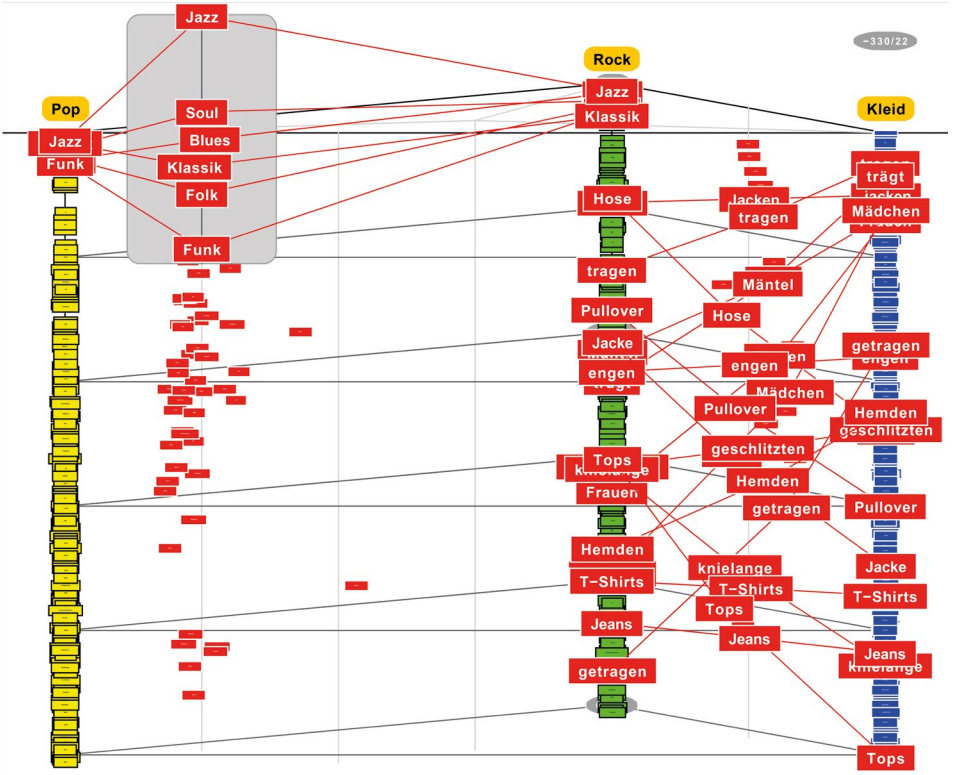

Abb. 15: Hervorgehobene Übereinstimmungen des Profils des Lemmas „Rock“ im Vergleich zu den Lemmata „Pop“ (Ausschnittsvergrößerung) und „Kleid“ (Markierungen) (Vicomte) (Copyright IDS).

Die Darstellung dieser Struktur leidet unter demselben Dilemma, das komprimierte Ansichten wie die Fokus-und-Kontext-Technik erforderlich gemacht hat. Auch hier wird nur überblicksartig die gesamte Information verkleinert dargestellt; eine Vergrößerung von Bestandteilen ist interaktiv über einen Ausschnitt oder durch explizite Markierung möglich. Die Gesamtansicht liefert einen ersten Eindruck von der Verteilung der Partnerwörter etwa bei Synonymen oder Paronymen. Im Detail zeigen sich dann die charakteristischen Partnerwörter. Bei den gemeinsamen Partnern wird durch die gewichtete Anordnung bei der Visualisierung der Eindruck unterstützt, wie relevant das jeweilige Partnerwort bei den jeweiligen Bezugswörtern ist.

Diese eher anschauliche Erklärung und Visualisierung einer gewichteten Ähnlichkeitsbeziehung zwischen Kookkurrenzprofilen ist bereits seit längerer Zeit durch ein formales Vergleichsmaß modelliert, das in der Kookkurrenzdatenbank CCDB Anwendung findet.

3.5 Kartierung von Gebrauchsaspekten

Die Kookkurrenzdatenbank CCDB (Belica 2007) wurde als Sammlung von Kookkurrenzprofilen zu über 220.000 Einträgen aufgebaut und dient als Denk- und Experimentierplattform für die weitere methodische Auswertung der Kookkurrenz, u. a. für die systematische Anwendung des Ähnlichkeitsvergleichs zwischen allen vorhandenen Einträgen. Das Ergebnis dieses Abgleichs wird für jedes Wort als Liste der verwandten Profile absteigend nach dem ermittelten Maß (Related Collocation Profiles) in der CCDB angeboten. So überzeugend und plausibel fast alle Einträge in diesen Listen für sich alleine stehend wirken, so sehr deutet sich eine Vielfalt unterschiedlicher Begründungen der Ähnlichkeit an – im Extremfall bis hin zu disjunkten Aufteilungen des Kookkurrenzverhaltens aufgrund mehrerer Lesarten eines Homonyms. Ein weiteres Verfahren versucht, das Verwendungsspektrum eines Wortes, das diese Vielfalt begründet, in ein grobes Raster einzuordnen. Dazu wird angenommen, dass sich die Wörter, die einem vorgegebenen Wort ähnlich sind, je in Gruppen einordnen lassen, innerhalb derer alle Elemente dem Bezugswort auf eine vergleichbare, aber auch von anderen Gruppen abgrenzende Art ähneln. Als Ausgangsmaß für diese Einordnung wird die Ähnlichkeit nach dem beschriebenen Maß für alle Paare von Einträgen in der Liste der ähnlichen Profile herangezogen. Nach Vorgabe eines Rasters arrangiert dann ein selbst-organisierendes Verfahren eine Anordnung aller ähnlichen Profile auf einer zweidimensionalen Karte (SOM, Kohonen 1990). Die hochdimensionale Vielfalt der Ähnlichkeitsbeziehungen wird so auf eine planare Topologie reduziert, bei der die geometrische Distanz den bestmöglichen Kompromiss aus Ähnlichkeit bzw. Unähnlichkeit aller Einträge widerzuspiegeln versucht. Oberhalb eines Schwellwerts wird allerdings nicht weiter differenziert. Die Gruppen von Wörtern, die die höchste Ähnlichkeit untereinander aufweisen, werden gemeinsam in einem Feld des Rasters abgebildet, da zu vermuten ist, dass sie tendenziell denselben Gebrauchsaspekt des Bezugswortes zum Ausdruck bringen.

Als Raster hat sich in vielen Anwendungen eine 5×5-Matrix bewährt, deren Felder in fließenden Farbtönen hinterlegt sind (vgl. Abb. 16). Damit soll visuell unterstützt werden, dass die Aspekte, die in den einzelnen Feldern ausgedrückt werden, weich ineinander übergehen. In den Fällen, in denen aufgrund (Un-)Ähnlichkeitsbeziehungen kein weicher Übergang möglich ist (etwa wenn Aspekte unterschiedlicher, scharf trennbarer Lesarten aneinanderstoßen sollen), wird dieser Abstand durch unbesetzte und ungefärbte Felder umgesetzt, die sich häufig (z. T. in Kombination mit weiteren, dünn besetzten Feldern) wie ein Trenngraben durch das gesamte Diagramm ziehen.

Bei der (meist lexikographisch motivierten) Interpretation der Karten hat sich eine semiotische Ausrichtung herauskristallisiert: Ausgehend von einzelnen Feldern werden auch die umgebenden Felder miteinbezogen, um zu sichten,

Rock

Musikstil Klassik Stilrichtung Mixtur Salsa Calypso Folklore Melange	Reggae Techno Funk Musikrichtung Rap Punkrock Mix Weltmusik	Pop Folk Hop Jazz Ska Punk Grunge Rockabilly	Soul Blues Hip Hardcore Dancefloor Groove Disco Rave	Acid Jam Sampler Independent
Rockmusik Popmusik Schlager Dixieland Chanson Tanzmusik Volksmusik Sprechgesang	tanzbar fetzig funkig eingängig melodiös jazzig Rocksong Siebziger	Gospel Swing grooven rockig Ragtime Mambo Eigenkomposition soulig	Country Boogie Sound Revival rappen	Jump Attack Chart Rocker Underground Floor Pistol Voodoo
knallig	poppig	Dixie Ballade Evergreen Gassenhauer Popsong Spiritual Schnulze Band	Song Twist rocken Ohrwurm Musical Medley Abba Feeling	covern Rocks Coverversion Soundcheck Cover Dancer Hot Nirvana
Jeans Hemd Blouson Pullover Sweatshirt Latzhose beigen hellblau	pinkfarben gestickt		Hit Welthit Album Count Creole Alben Maid King	Diskografie Trouble Dust Sweet Live Boy Skin Straight
Hose Jacke Shirt Mantel Shorts Pulli Blazer Strickjacke	ärmellos knielang tailliert Bluse geschlitzt hauteng Hosenanzug Oberteil	Slip Petticoat Smoking	Let Titelsong Look Springfield	My Girl Go Rain Out Want On Gon

Abb. 16: Self-organized Map des Lemmas „Rock" (CCDB) (Copyright IDS).

inwieweit diese gemeinsame Aspekte zum Ausdruck bringen. Durch die Kartierung der Gebrauchsaspekte und ihre Interpretation konnten aufschlussreiche Aspekte über die Verwendungsspektren der betrachteten Wörter gewonnen werden (Vachkova/Belica 2009).

Eine Erweiterung des Verfahrens ist für die Anwendung auf Wortpaare konzipiert und operiert auf der Vereinigungsmenge aller Profile, die zu einem Wort (oder beiden) als ähnlich eingestuft wurden. Abweichend von der SOM-Farbgestaltung wird hierbei dann durch die Einfärbung eine weitere Information kodiert: Je nach Verhältnis der Ähnlichkeiten der Felder zu einem der beiden vorgegebenen Wörter wird die Feldfarbe aus Anteilen von Rot und Gelb zusammengemischt. Ein klares Votum für das eine oder andere Wort spiegelt

Abb. 17: Self-organized Map der Kontrastierung der Lemmata „Rock“ und „Kleid“ (CCDB) (Copyright IDS).

sich in den ihnen zugeordneten Primärfarben wider; unklare Zuordnungen zeigen sich durch entsprechende Abstufungen von Orangetönen.

Angewandt auf Paare, die als Synonyme oder Paronyme gelten (vgl. Abb. 17), zeigt das Verfahren und diese Einfärbung auf, in welchen Domänen oder Diskursbereichen die Wörter sich nahestehen oder sich scharf trennen lassen. Studien in diesen Bereichen (Marková 2012, Storjohann/Schnörch 2014) zeigen, dass der reflektierte Einsatz dieser Methoden gerade auch durch die Interpretation der Visualisierungen einen erheblichen Mehrwert empirischer Analysen darstellt.

Eine Besonderheit der Kartierung mithilfe von SOMs ist, dass sie kleinere Unterschiede vollständig ausblendet, indem sie verschiedene Wörter einer Zelle

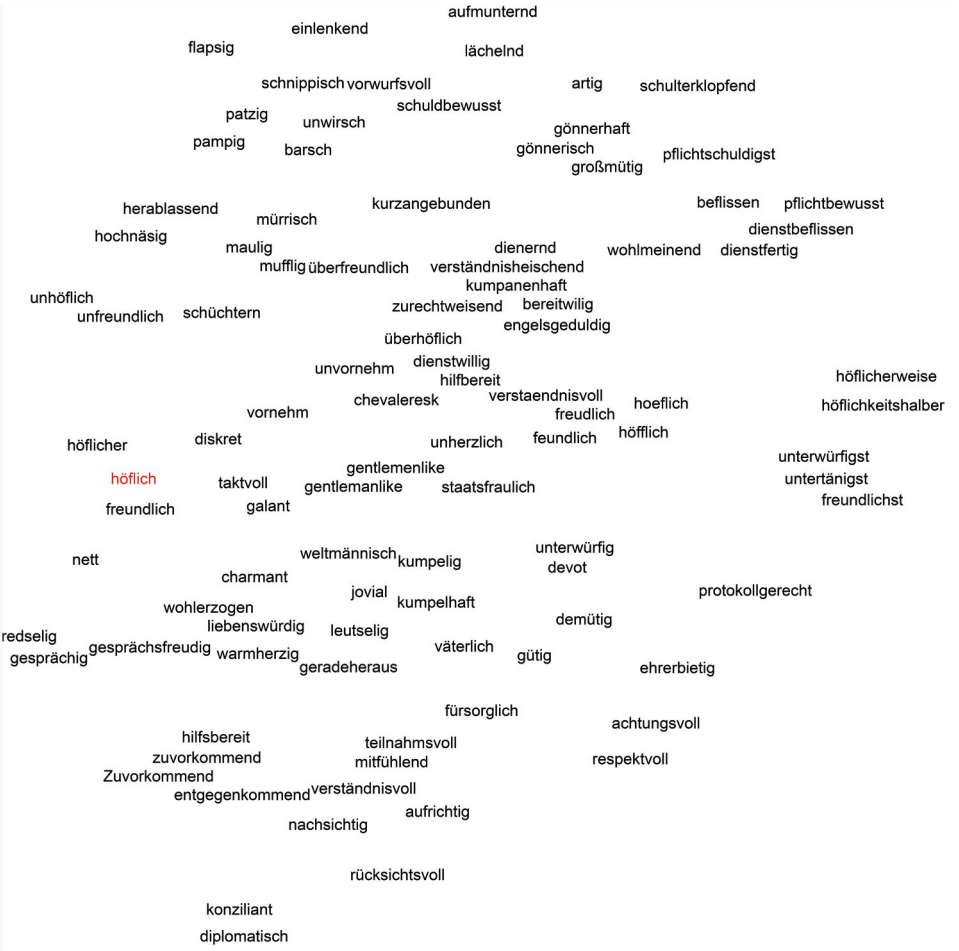

Abb. 18: „höflich“ und seine 99 distributionell ähnlichsten Nachbarn in DeReKo dargestellt mithilfe von t-SNE (Copyright IDS).

auf der Karte zuordnet, ohne deren Ähnlichkeitsbeziehungen und Abstände innerhalb dieser Gruppe im Diagramm darzustellen. Oft ist das Weglassen solcher potenziell ablenkender Detailinformation aber ein gewünschter Effekt, um den Blick auf das Wesentliche nicht zu verstellen. So kann eine solche Quantisierung bzw. ein solches Clustering erfahrungsgemäß die Abduktion neuer Hypothesen erleichtern. Dabei muss natürlich beachtet werden, dass man es nicht mit Ergebnissen zu tun hat, sondern nur mit Hypothesen, die erst noch überprüft

werden müssen (z. B. mithilfe neuer Analysen oder einer manuellen Untersuchung von Belegstellen).

Wenn gerade nicht das Ausblenden von Detailinformationen zur Abduktion allgemeinerer Hypothesen, sondern eine detaillierte, topographieerhaltende Darstellung engerer Ähnlichkeitsbeziehungen das Ziel ist, hat sich in den letzten Jahren t-SNE (van der Maaten & Hinton 2008) als eine in der Regel gut geeignete Methode zur Dimensionsreduktion etabliert. Abb. 18 zeigt „höflich" und seine 99 bezüglich ihrer distributionellen Eigenschaften ähnlichsten Nachbarn auf einer mithilfe von t-SNE erzeugten Karte. Zur Vektorrepräsentation der Wörter wurden in diesem Fall nicht Kookkurrenzprofile verwendet, sondern sogenannte „word embeddings" (Mikolov et al. 2013), die mittels einer Erweiterung der Programme word2vec bzw. wang2vec (Ling et al. 2015) auf der Basis von DeReKo-2016-I (Institut für Deutsche Sprache 2016a) berechnet wurden. Auch solche Streudiagramme können natürlich so angereichert werden, dass sie bestimmte Zusammengehörigkeitshypothesen nahelegen, indem etwa die Ergebnisse von Clusteranalysen über gemeinsame Farben oder Rahmen um zusammengehörige Knoten kodiert werden.

4. Fazit und Ausblick

In vielen Bereichen unserer Arbeitsfelder hat es sich als sinnvoll und hilfreich erwiesen, eine Mischung von quantitativ-qualitativen Vorgehensweisen mit Visualisierungstechniken zu kombinieren. Dabei können es auch durchaus schlichte Ansätze sein, die ausreichen, um die Interpretation auf Interessantes zu lenken. Aufwändigere Techniken und insbesondere Interaktionsmöglichkeiten bergen neben dem Einarbeitungsaufwand bisweilen die Gefahr, verstärkt Aufmerksamkeit zu binden und den Status der angebotenen Signale überzubewerten.

Man sollte stets im Hinterkopf behalten, dass nicht unbedingt neue Fakten, sondern nur zu interpretierende Hinweise angeboten werden. Diese können in verzerrender Weise zu Unrecht überspitzt sein, sie können auch weniger relevant sein als andere, die nicht in den Vordergrund gerückt wurden.

Auch Visualisierungstechniken müssen reflektiert und mit der nötigen Distanz eingesetzt werden. Sie sollten, wenn möglich, in verschiedenen Variationen verglichen werden können, um die ersten, schnellen Hypothesen auf den Prüfstand zu stellen. Ein in jeglicher Hinsicht kritischer Punkt der Verfahren ist dabei, die Menge an Information auf das Wesentliche zu reduzieren und auf das Relevante zu fokussieren. Denn gerade dafür erhofft sich der/die NutzerIn Unterstützung bei der Bewältigung des Reichtums an Hinweisen, die in einem sehr großen Korpus stecken.

Es ist absehbar, dass der Bereich der Visualisierung zukünftig an Bedeutung gewinnt und etwa durch Andockmöglichkeiten diverser Visualisierungsverfahren an unser Recherchesystem für ein breites Fachpublikum zu einer selbstverständlichen Ergänzung des üblichen Arbeitens werden wird (s. a. Kupietz et al. 2015.).

5. Bibliographie

Baroni, Marco, Silvia Bernardini, Adriano Ferrares und Eros Zanchetta. 2009. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora." *Language Resources and Evaluation* 43 (3): 209–226. wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf.

Belica, Cyril. 1995. „Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode." http://corpora.ids-mannheim.de/ (letzter Zugriff am 12. Oktober 2016).

Belica, Cyril. 2007. „Kookkurrenzdatenbank CCDB – V3: Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs." http://corpora.ids-mannheim.de/ccdb/ (letzter Zugriff am 12. Oktober 2016).

Bodmer Mory, Franck. 2014. „Mit COSMAS II ‚in den Weiten der IDS-Korpora unterwegs'." In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache,* herausgegeben vom Institut für Deutsche Sprache. Mannheim: Institut für Deutsche Sprache, 376–385.

Cox, Trevor F. und Michael A. A. Cox. 2001. *Multidimensional Scaling.* Boca Raton, Fla.: Chapman and Hall.

Institut für Deutsche Sprache. 2016a. „Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2016-I (Release vom 31.03.2016)." Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (letzter Zugriff am 12. Oktober 2016).

Institut für Deutsche Sprache. 2016b. „*Cosmas II*-Recherchesystem." https://cosmas2.ids-mannheim.de/cosmas2-web/ (letzter Zugriff am 12.Oktober 2016).

Keibel, Holger, Sophie Hennig und Rainer Perkuhn. 2010. *Effiziente halbautomatische Detektion von Neologismuskandidaten.* Mannheim: Institut für Deutsche Sprache (Technical Report IDS – KL-2010-01). www.ids-mannheim.de/kl/dokumente/ids-kl-2010-01.pdf

Kilgarriff, Adam. 2001. "Comparing Corpora." *International Journal of Corpus Linguistics,* 6 (1): 97–133.

Kohonen, Teuvo. 1990. "The Self-Organizing Map. New Concepts in Computer Science." In Informatique: nouveaux concepts scientifiques. Colloque en l'honneur de Jean-Claude Simon, Paris. AFCET, 181–190.

Kupietz, Marc. 2005. *Near-Duplicate Detection in the IDS Corpora of Written German.* Mannheim: Institut für Deutsche Sprache (Tech. Rep. KT-2006-01*). www1.ids-mannheim.de/fileadmin/kl/misc/ids-kt-2006-01.pdf.*

Kupietz, Marc, Cyril Belica, Holger Keibel und Andreas Witt. 2010." The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research." In *Proceedings of the seventh conference on International Language Resources and Evaluation* (LREC 2010), herausgegeben von Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner und Daniel Tapias. Malta : ELRA, 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Marc Kupietz, Harald Lüngen, Cyril Belica, Cyril und Rainer Perkuhn. 2012. "Webkorpora als qualitätsgesicherte Forschungsdaten." Unveröffentlichter Vortrag im Rahmen des GSCL-Workshops Webkorpora in Computerlinguistik und Sprachforschung am 27. September 2012.

Kupietz, Marc, Nils Diewald, Michael Hanl und Eliza Margaretha. 2016. „Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP." In *Grammatische Variation – empirische Zugänge und theoretische Modellierung*, herausgegeben von Marek Konopka und Angelika Wöllstein. Berlin: de Gruyter, 319–330.

Marc Kupietz und Thomas Schmidt. 2015. „Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung." In *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven,* herausgegeben von Ludwig M. Eichinger. Berlin: de Gruyter, 297–322(Jahrbuch des Instituts für Deutsche Sprache 2014).

Lamping, John, Ramana Rao und Peter Pirolli. 1995. "A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies." In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM, 401–408.

Ling, Wang, Chris Dyer, Alan Black und Isabel Trancoso. 2015. "Two/Too Simple Adaptations of word2vec for Syntax Problems." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics.* Denver, CO: Human Language Technologies. www.cs.cmu.edu/~lingwang/papers/naacl2015.pdf.

Lüngen, Harald und Holger Keibel. 2013. „Zur Erstellung und Interpretation der Zeitverlaufsgrafiken." In *Neuer Wortschatz: Neologismen im Deutschen 2001–2010,* herausgegeben von Doris Steffens und Doris al-Wadi. Mannheim: Institut für Deutsche Sprache, 561–567.

Marková, Věra. 2012. Synonyme unter dem Mikroskop: Eine korpuslinguistische Studie. Tübingen: Narr (CLIP 2).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado und Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality."

In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality (letzter Zugriff am 27. November 2017).

Perkuhn, Rainer. 2007a. "Systematic Exploration of Collocation Profiles." In: *Proceedings of the 4th Corpus Linguistics Conference (CL 2007)*. Birmingham: University of Birmingham. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4715 (letzter Zugriff am 27. November 2017).

Perkuhn, Rainer. 2007b. „'Corpus-driven': Systematische Auswertung automatisch ermittelter sprachlicher Muster." In *Sprach-Perspektiven: Germanistische Linguistik und das Institut für Deutsche Sprache,* herausgegeben von Heidrun Kämper und Ludwig M. Eichinger. Tübingen: Narr, 465–491 (Studien zur Deutschen Sprache 40).

Perkuhn, Rainer. 2012. *Diachrone Kookkurrenzanalyse.* Mannheim: Institut für Deutsche Sprache (Technical Report IDS-KL-2012-02).

Perkuhn, Rainer und Cyril Belica. 2016. „Konflikt, Sprache, korpuslinguistische Methodik." In *Linguistische Zugänge zu Konflikten in europäischen Sprachräumen. Korpus – Pragmatik – kontrovers,* herausgegeben von Friedemann Vogel, Stefaniya Ptashnyk und Janine Luth. Heidelberg: Winter 4), 339–364 (Schriften des Europäischen Zentrums für Sprachwissenschaften (EZS).

Schnörch, Ulrich. 2015. „Wortschatz." In *Handbuch „Wort und Wortschatz"*, herausgegeben von Ulrike Haß und Petra Storjohann. Berlin/Boston: de Gruyter, 3–26 (Handbücher Sprachwissen 3).

Schumann, Heidrun und Wolfgang Müller. 2000. *Visualisierung–- Grundlagen und allgemeine Methoden.* Berlin: Springer.

Storjohann, Petra. 2007a. „Wie viel Diskurs braucht ein Wörterbuch?" *German Life and Letters* 60, Issue 4: 569–592.

Storjohann, Petra. 2007b. „Der Diskurs ‚Globalisierung' in der öffentlichen Sprache. Eine korpusgestützte Analyse kontextueller Thematisierungen." *Aptum: Zeitschrift für Sprachkritik und Sprachkultur* 2007, Heft 2: 139–155.

Storjohann, Petra und Ulrich Schnörch. 2014. "Empirical Approaches to Paronyms." In: *Proceedings of the XVI EURALEX International Congress*, herausgegeben von Andrea Abel, Chiara Vettori und Natascia S. Ralli, 463–476. Bozen: Institute for Specialised Communication and Multilingualism.

Vachková, Marie und Cyril Belica. 2009. "Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography." *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis* 13 (2): 223–260.

van der Maaten, Laurens und Geoffrey Hinton. 2008. "Visualizing High-Dimensional Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.