Originalveröffentlichung in: Engelberg, Stefan/Kämper, Heidrun/Storjohann, Petra (Hrsg.): Wortschatz: Theorie, Empirie, Dokumentation. - Berlin [u.a.]: de Gruyter, 2018. S. 61-89. (Germanistische Sprachwissenschaft um 2020, Bd. 2)
DOI: *https://doi.org/10.1515/9783110538588-004*

Rainer Perkuhn

3 Kontexte und ihre Verteilung

Kookkurrenz und Distribution

Abstract: Die typischen sprachlichen Kontexte, in denen ein Wort verwendet wird, spannen den Rahmen auf, über den sowohl Sprecher als auch Forscher einer Sprache wesentliche Aspekte der Bedeutung des Wortes erschließen und vermitteln. Über große Korpora und entsprechende korpus-, aber auch computerlinguistische Methoden stehen nunmehr systematische Zugänge zu den typischen Verwendungsweisen zur Verfügung, am Institut für Deutsche Sprache etwa über die Kookkurrenzanalyse seit 1995. Auf den Ergebnissen des letztgenannten Verfahrens operieren weitere Methoden, die Bedeutungsbeziehungen zwischen Wörtern auf Ähnlichkeitsbeziehungen des Kontextverhaltens zurückführen. In jüngerer Zeit werden Ansätze vor allem aus der Computerlinguistik und dem *information retrieval* diskutiert, die mit einem ähnlichen Ziel antreten. Dieser Beitrag soll einen prinzipiellen Überblick bieten, wie die verschiedenen Forschungsstränge den Begriff *Kontext* interpretieren, wie sie ihn systematisch erfassen und zum Vergleich einsetzen. Neben Bedeutungsnähe wird vor allem Mehrdeutigkeit besondere Beachtung finden.

Keywords: Bedeutungsnähe, Distribution, Kontext, Kookkurrenz, Polysemie/Homonymie, vector space models, word embeddings

1 Einleitung

Am Institut für Deutsche Sprache werden seit seiner Gründung 1964 Daten für insbesondere schriftsprachliche Korpora gesammelt und aufbereitet (Teubert & Belica 2014). Auf welche Art und Weise Nutzer Zugang zu den Daten hatten, war zunächst den technischen Randbedingungen geschuldet – auf die Idee,

Anmerkung: Die überzeugenden Argumente in diesem Bericht gehen zurück auf die gemeinsame Arbeit mit Cyril Belica, Marc Kupietz, Holger Keibel, Harald Lüngen und Peter Fankhauser. Die übrigen Gedankengänge verantwortet der Autor alleine.

Rainer Perkuhn, IDS Mannheim, R5, 6–13, D-68161 Mannheim,
E-Mail: perkuhn@ids-mannheim.de

eine Kopie eines Stapels Lochkarten mit nach Hause nehmen zu wollen, wäre in der Anfangszeit sicher kein Nutzer gekommen. Lange Zeit, bis weit in die Achtzigerjahre des letzten Jahrhunderts, wurden interessierte Linguisten im Umgang mit einfachen Rechercheinstrumentarien geschult, so wie sie von den Rechnerumgebungen bereits mit bereitgestellt wurden. Um diesen Aufwand zu reduzieren, aber auch, um auf die Anfrage- und Analysebedürfnisse der linguistischen Nutzer besser eingehen zu können, wurden und werden eigene Recherchesysteme entwickelt und eingesetzt, beginnend mit REFER (Brückner 1983, 1988/1989) über Cosmas I (al-Wadi 1994) und Cosmas II (Bodmer 2005, 2014) – bis hin zu dem aktuell im Beta-Betrieb anlaufenden KorAP (Bański et al. 2013). Das Angebot der dedizierten Schnittstellen ermöglichte, die Daten vor dem unmittelbaren Zugriff abzuschirmen; im Zuge der technischen Fortentwicklung ergab sich dann quasi automatisch die Recherchemöglichkeit über das zunächst lokale Netzwerk des IDS, später dann über das Internet. Die in den Neunzigerjahren eingeführte Registrierungspflicht für Endnutzer erlaubte die Überwachung individueller Rechercheaktivitäten. Dies war die Voraussetzung dafür, dass bei den Lizenzverhandlungen mit den Textspendern eingeräumt werden konnte, dass die Nutzung der Daten auf rein wissenschaftliche und nicht-kommerzielle Zwecke für einen begrenzten Nutzerkreis (und im Sinne des Zitationsrechts) beschränkt wird. Dieser Auflage müssen Nutzer durch Akzeptieren der Endnutzervereinbarung zustimmen und anerkennen, dass die Zuwiderhandlung sanktioniert wird.

Die Zugangsbündelung unter ein System unter eigener Kontrolle bot und bietet weiterhin die Möglichkeit, die Funktionalität des Systems um fremd- oder eigenentwickelte Methoden nach und nach für alle Nutzer zu erweitern. Das zukünftige System KorAP setzt dies *par excellence* um, indem es Andockmöglichkeiten für externe Module anbietet. Bei den bisherigen Systemen wurde diese Integration noch IDS-intern bewerkstelligt. Außer eigenen Verfahren der Tokenisierung und einem auch heute noch durchaus wettbewerbfähigen Lemmatisierer (Belica 1994) hat das IDS im traditionellen Bereich vor allem extern entwickelte Werkzeuge eingesetzt – zum einen, weil es vermessen gewesen wäre, zu glauben, mit den sehr begrenzten Kapazitäten am IDS etwas noch Besseres entwickeln zu können, zum anderen aber auch, weil jedes Werkzeug nur eine angenäherte Umsetzung einer linguistischen Sichtweise darstellen kann, insbesondere abgesteckt durch den Rahmen des Formalisierbaren und Operationalisierbaren. Insofern ist es durchaus sinnvoll, für einen vorsichtigen Umgang mit derartigen Angaben zu plädieren. Mit dem neuen System KorAP sind nun erstmals die Voraussetzungen geschaffen, mit mehreren konkurrierenden Angaben mehrerer Werkzeuge zu arbeiten, um die Übereinstimmung oder Widersprüchlichkeit dahinterliegender Annahmen oder Meinungen mit untersuchen zu können.

Die meisten Anfrageszenarien zielten darauf, eine übersichtliche Treffermenge zu erzielen, mit der eine Hypothese überprüft werden konnte oder aus der Nutzer Belege zur Dokumentation und zur Illustration einer bestimmten Verwendungsweise auswählen konnten. Für die Formulierung entsprechender Suchanfragen haben die Anfragesprachen immer ausgefeiltere Ausdrucksmöglichkeiten bereitgestellt. Die Ergebnisse einer Recherche konnten – je nach verfügbaren Metadaten – in verschiedenen Übersichtsformen präsentiert werden oder in der Gesamtdarstellung – etwa gemäß der Dimension „Zeit" – nach diesen sortiert werden. Kontextuelle Beziehungen ließen sich bei der Sortierung der Ergebnisgesamtansicht über benachbarte Wörter ansatzweise erahnen – oder durch die Formulierung von Mehrwortsuchausdrücken vorgeben, in denen wort- und/oder satzbezogene Abstände festgelegt werden.

Mit den immer größer werdenden Datenmengen wuchs das Bedürfnis nach Filter- und Sortierangeboten; gleichzeitig ermöglichten diese den Einsatz statistischer Verfahren, z. B. zur thematischen Klassifikation. Während hierbei wieder im Wesentlichen auf externe Vorarbeiten zurückgegriffen werden konnte, musste in einem anderen Bereich viel Eigenentwicklung betrieben werden, da es noch kein ausreichend elaboriertes Verfahren gab, das den Besonderheiten der deutschen Sprache gerecht geworden wäre. Die Methode der Kookkurrenzanalyse (Belica 1995) ermöglicht es, zu einer gegebenen sprachlichen Einheit herauszufinden, in welchen typischen Kontexten sie verwendet wird. Die wichtigsten Begleiter eines Wortes treten dabei aus der Analyse hervor; sie müssen nicht vorher erahnt und als Teil einer Mehrwortsuchanfrage benannt werden. Bevor die Methode in Kapitel 3 kurz beschrieben wird, soll im folgenden Kapitel zunächst die Tragweite der Dimension „Kontext" motiviert werden.

2 Kontext

Wird das Korpus in einer herkömmlichen Art und Weise genutzt, um Belege zu einer sprachlichen Einheit, einem Wort oder einer Mehrworteinheit, oder zu einem grammatischen Konstrukt zu finden, hoffen Nutzer auf eine kleine Treffermenge oder eine schnelle Reduktion, um dann hermeneutisch die einzelnen (Gesamt-)Trefferkontexte auf die Eignung der Belege zu sichten. Für lexikographische Zwecke können dann daraus Belegbeispiele gewählt oder andere illustrative Angaben abgeleitet werden. Bei größeren Treffermengen kommt diese Vorgehensweise schnell an ihre Grenzen; dafür bräuchte es ein Ordnungssystem der Menge anhand kontextueller Eigenschaften. Wenn wir den Blick ein wenig weiter fassen und nicht nur auf Lexikographie begrenzen, so kann man durchaus die These aufstellen, dass jeder Muttersprachler in der Lage ist, dieses Ordnungssystem aus dem eigenen Spracherleben abzuleiten.

> What people know when they know a word is not how to recite its dictionary definition – they know how to use it (when to produce it and how to understand it) in everyday discourse [...]. [...] people learn how to use words by observing how words are used. And because words are used together in phrases and sentences, this starting assumption directs attention immediately to the importance of context. (Miller & Charles 1991: 4)

Auch wenn wir bisher mit keinem Korpus die Datenmenge simulieren können, der ein einzelner Sprecher in seinem Spracherwerb ausgesetzt ist, schon gar nicht in einer chronologisch-inkrementellen Abfolge, so ist es doch einen Versuch wert, das Deutsche Referenzkorpus (Institut für Deutsche Sprache 2017b) als eine halbwegs vertretbare Annäherung an den rezeptiven Sprachstand eines Erwachsenen zu betrachten. Ganz im Sinne von Firth' berühmtem Zitat wäre ein systematischer Zugang zum Kontextverhalten eines Wortes der Schlüssel zu seinem Verständnis:

> You shall know a word by the company it keeps! (Firth 1957: 11)

Wie der Begriff *Wort* aus empirischer Sicht, insbesondere auf die deutsche Sprache angewandt, zu deuten ist, stellen wir noch kurz zurück. Für das folgende Beispiel schauen wir auf Wortformen. Auch wie konkret der Begriff *Kontext* auszulegen ist, wie viele Wörter oder Phrasen um das betrachtete Wort herum er umfassen soll und ob weitere hierarchisch-syntaktische oder kategorielle Bedingungen betrachtet werden, vereinfachen wir – wie durchaus üblich – zu einem einfachen Textfenster von fünf Wörtern vor und fünf Wörtern nach dem Bezugswort. Etwas ausführlicher klären müssen wir aber, was sich hinter „company" verbirgt. Es geht Firth offensichtlich nicht um zufälliges, gelegentliches Miteinandervorkommen, sondern um die übliche Gesellschaft, in der sich die Wörter miteinander befinden. Er nennt dies „habitual collocation" und führt dazu aus:

> The habitual collocations in which words [...] appear are quite simple the mere accompaniment, the other word-material in which they are most commonly or most characteristically embedded. (Firth 1957: 11 f.)

Firth spielt mit den beiden Adverbien im Superlativ natürlich auf häufiges oder systematisches Miteinandervorkommen an, ohne dass er an irgendeiner Stelle andeuten würde, wie sich die beiden Eigenschaften quantitativ genau fassen ließen. Um einen Eindruck zu bekommen, betrachten wir ein Beispiel. Wir haben mit dem Suchausdruck

> Weichsel /w1:5,s0 <Umgebungswort>

mit Hilfe von Cosmas II (Institut für Deutsche Sprache 2017a) die absoluten Häufigkeiten für die entsprechenden Wortkombinationen im Deutschen Refe-

renzkorpus ermittelt. Für den Ausdruck <Umgebungswort> ist eines der Wörter der ersten Zeile aus Tabelle 3.1 einzusetzen. Der Suchausdruck setzt die beiden Teilbegriffe zueinander in eine positionelle Beziehung, die mindestens ein, höchstens fünf Wörter betragen soll und auf denselben Satz begrenzt wird.

Tab. 3.1: Anzahl Vorkommen verschiedener Umgebungswörter im Kontext von *Weichsel*.

	Rhein	Elbe	Wechsel	fließt	Mündung	Ufer	gelegen	Kirsche	total
Weichsel	58	91	28	16	41	102	11	10	6109

Die Auswahl der betrachteten Umgebungswörter ist handverlesen und durch die später weiter darauf aufbauenden Ausführungen primär didaktisch motiviert. Abgesehen von dem letzten Wort *Kirsche*, auf das wir später eingehen, sind alle anderen Umgebungswörter auch naheliegend, sofern man weiß, dass mit dem Wort *Weichsel* ein Fluss bezeichnet wird. Wie können jetzt aber die Zahlen gedeutet werden, gerade auch mit Blick darauf, ob sie charakteristisch oder besonders geläufig sind – zumal andere, hier nicht gezeigte Wörter mit ähnlichen Häufigkeiten in der Umgebung auftreten (z. B. *Wochen* 13 Mal). Firth und auch später Harris sind sich dieser Vagheit auch bewusst; letzterer deutet aber auch nur die Richtung an, in der bis heute die beste Lösung gesucht wird:

> It is rather a question of the relative frequency of such environments ... (Harris 1970: 786)

Was heißt aber hier „relativ"? Der relative Anteil am Vorkommen des Bezugswortes oder der relative Anteil am Vorkommen des Umgebungswortes? Oder eine Mischung aus beidem? Genau genommen geht es darum, zu bewerten, ob die beobachtete Häufigkeit einer Wortkombination größer ist, als sie sein dürfte, wenn die beteiligten Wörter nur zufällig in den Umgebungen voneinander so verteilt wären. Folgendes Zitat bringt es auf den Punkt:

> [...] the aim is to compile a list of those syntagmatic items ('collocates') significantly co-occurring with a given lexical item ('node') within a specified linear distance ('span'). 'Significant collocation' can be defined in statistical terms as the probability of the item x co-occurring with the items a, b, c, ... being greater than might be expected from pure chance. (Berry-Rogghe 1973: 103)

Wie groß aber der erwartbare, zufällig noch vertretbare Wert bemessen wird und mit welcher Metrik die Abweichung davon erfasst werden soll, wird immer noch durchaus kontrovers diskutiert. Übereinstimmung herrscht nur bei dem Eingeständnis, dass die Verteilung sprachlicher Phänomene eigenwilliger Cha-

$$surprise(w_1, w_2) = \log_2 \left| \frac{\frac{f(w_1, w_2) + 1}{K}}{\frac{f(w_1)}{K} * \frac{f(w_2)}{K} * C} \right|$$

mit:

$f(w_1, w_2)$, $f(w_1)$, $f(w_2)$	*Frequenzen der Folge und der Wörter*
K	*Korpusumfang*
C	*Kontextbreite*

Formel 3.1: Häufigkeit der Kombination: Erwartungswertüberschreitung (ohne grau), Pendant zu *pointwise mutual information* (mit grau).

Tab. 3.2: Werte gem. Formel 3.1, ohne grau.

	Rhein	Elbe	Wechsel	fließt	Mündung	Ufer	gelegen	Kirsche
Weichsel	25,80	160,84	667,09	18,44	239,49	103,50	7,73	233,73

rakteristika unterliegt, sodass die Anwendung üblicher statistischer Modelle teilweise fragwürdig ist. Trotzdem wird dies in Kauf genommen, um einfache Ansätze entwickeln zu können, so auch an dieser Stelle unserer Argumentation für die Herleitung eines einfachen Maßes.

Um abzuschätzen, wie oft ein Wort (hier: *w*1) zufällig vorkommen dürfte, berechnen wir seine relative Häufigkeit, indem wir die absolute Häufigkeit durch den Korpusumfang teilen.

Wir ziehen aber nicht nur einmal ein Wort aus dem Korpus, das mit eben dieser Wahrscheinlichkeit das Wort *w*1 sein könnte (wenn wir ein ganz simples Modell zugrunde legen), sondern wir betrachten alle Wörter in der Umgebung aller Vorkommen des Bezugswortes w_2 auf so viel Positionen, wie es die Kontextdefinition vorgegeben hat. Diese Berechnung entspricht dem Nenner in Formel 3.1 ohne die grauen Bestandteile. Setzt man diesen nun ins Verhältnis zu der tatsächlich beobachteten Häufigkeit der Kombination (der Zähler ohne grau), erhält man den Faktor, um wie viel häufiger diese auftritt, als es erwartbar wäre; die entsprechenden Werte sind in Tabelle 3.2 wiedergegeben.

Offen gesagt, sind auch diese Angaben nur schwierig genauer zu deuten. Alle Umgebungswörter kommen häufiger vor, als bei einem zufälligen Vorkommen zu erwarten wäre. Der Faktor „um wie viel häufiger" liegt allerdings um Größenordnungen auseinander. Man ahnt aber schon, dass das Umgebungswort *Kirsche* trotz eigentlich geringer absoluter Häufigkeit durch diese Bewertung an Bedeutung gewonnen hat.

Anstelle der beiden Terme für absolute Häufigkeiten können wir auch äquivalent relative Angaben einsetzen, indem wir beide durch den Korpusumfang teilen. Nun sind wir noch zwei kleine Schritte von einer Formel entfernt, die tatsächlich in verschiedenen Varianten häufig Verwendung findet. Um die

Tab. 3.3: Werte gem. Formel 3.1, mit grau.

	Rhein	Elbe	Wechsel	fließt	Mündung	Ufer	gelegen	Kirsche
Weichsel	4,71	7,35	9,43	4,29	7,94	6,71	3,08	8,01

zunehmenden Größensprünge zu dämpfen, wenden wir den Zweierlogarithmus auf den gesamten Term an. Auch wenn es dazu selten eine Begründung gibt, könnte dessen Plausibilität auf eine Parallele zum Weber-Fechner-Gesetz zurückzuführen sein, nach dem unser Einschätzungsvermögen Unterschiede auf höheren Skalenordnungen nicht mehr linear, sondern nur noch nahezu exponentiell abzugrenzen vermag (Fechner 1860/1907). Da das Logarithmisieren des Wertes 0 nicht definiert ist, bedient man sich an der Stelle eines weiteren Tricks und erhöht die Häufigkeiten der Kombinationen um 1, sodass die einzige Gefahr, dass der Zähler Null werden könnte, vermieden wird.

Die Werte dieser Berechnung finden sich in Tabelle 3.3 für die Formel 3.1 mit allen Bestandteilen, mit der wir uns quasi an das Maß *pointwise mutual information* (pointwise MI) angenähert haben. Dass das Umgebungswort *Kirsche* nunmehr schon auf den zweiten Rang aufgerückt ist, ist ein Artefakt der Erhöhung der Häufigkeit der Kombination um 1: 11 statt 10 macht sich bei *Kirsche* stärker bemerkbar als 42 statt 41 bei *Mündung*. Ansonsten hat sich das Spektrum der Werte nur enger zusammengeschoben; die Rangfolge der Ausschläge ist gleich geblieben. In welcher Weise diese oder andere derartige Werte für eine menschliche Interpretation oder eine weitere automatische Auswertung einfließen können, wird in unterschiedlichen Szenarien unterschiedlich gehandhabt.

3 Kookkurrenzanalyse des IDS

Bei der Kookkurrenzanalyse des IDS (Belica 1995) geht es primär um eine erkenntnisleitende Bewertung von Wortkombinationen. Diese werden intern in einer Struktur analog zu Tabelle 3.3 erfasst. Im Standardfall startet ein Nutzer die Analyse selber im Rahmen einer Cosmas II-Sitzung im Anschluss an eine Recherche. Das Verfahren gruppiert dann die Textstellen der Treffermenge anhand der darin entdeckten auffälligen Wortkombinationen und sortiert diese Gruppen nach der gemessenen Auffälligkeit.

Der Nutzer kann hierbei durch die Wahl der Suchanfrage vorgeben, welche Art sprachlicher Einheit untersucht werden soll. In den gängigsten Konstellati-

onen einer Einwortsuchanfrage ist dies entweder eine konkrete Wortform oder eine Grundform, ein sogenanntes *Lemma*, oder als Mischung davon eine Aufzählung mehrerer Formen, etwa eine Teilmenge eines Flexionsparadigmas. Zu den Trefferobjekten kann dann der Kontext frei definiert werden, wie viele Wörter links und rechts betrachtet werden sollen und ob Satzgrenzen den Kontext auf jeden Fall beschneiden. Die Häufigkeiten aller in diesem Kontext beobachteten Umgebungswörter werden dann gezählt und vor dem Hintergrund ihrer Frequenz im Gesamtkorpus statistisch bewertet – allerdings nicht mit dem oben hergeleiteten Maß, sondern mit dem *Loglikelihood Ratio* (Dunning 1993), dessen Formel wesentlich komplizierter aufgebaut ist und hier nicht weiter erklärt werden kann, aber sich für viele Fragestellungen in unserem Umfeld bewährt hat. Im Grunde versucht es aber auch, auf eine etwas andere Art und Weise zu bewerten, wie überraschend es ist, dass Wörter gemeinsam so oft vorkommen im Vergleich zu einem zufälligen Nebeneinanderstehen. Für die Partnerwörter kann der Nutzer wählen, ob Funktionswörter mit betrachtet oder ausgeblendet werden sollen. Dies hat aber keinen sonstigen Einfluss auf die übrigen im Kontext betrachteten Wörter. Des Weiteren kann – unabhängig von der sprachlichen Einheit, nach der gesucht wurde – festgelegt werden, ob die Umgebungswörter getrennt nach Wortformen bewertet werden, oder ob mehrere Formen, die einem gemeinsamen Lemma zugeordnet werden können, zusammen betrachtet und kumuliert bewertet werden sollen. Für diese Vorgaben wird dann die statistische Bewertung vorgenommen, sodass eine Tabellenzeile entsteht, vergleichbar zu dem Beispiel aus dem vorherigen Kapitel. Die Zellen dieser Zeile werden dann nach dem Assoziationsmaß sortiert. In Abhängigkeit von der Zuverlässigkeit der statistischen Bewertung (sozusagen „der Zuversicht in sich selbst") kann der Nutzer entscheiden, wie lang der vordere Abschnitt der sortierten Liste ist, die zunächst festgehalten werden soll. Das Verfahren wiederholt jetzt quasi für jedes Ergebnis eines vorhergehenden Durchlaufs eine etwas modifizierte Bewertung, ob zu den bereits erkannten auffälligen Wortkombinationen noch weitere Wörter hinzutreten, die in der Umgebung der Wortkombination besonders auffällig oft beobachtet wurden. Auch hierbei kann der Nutzer vorgeben, wie sicher die Auffälligkeit eingeschätzt werden soll, sodass unterschiedlich granulare Strukturen entstehen.

Das Gesamtergebnis einer Analyse, alle ermittelten Angaben zur Liste der auffälligen Umgebungswörter, wird in Cosmas II in einer eventuell sehr umfangreichen Tabelle, einem sogenannten *Kookkurrenzprofil*, dargestellt, aus der wir hier nur einen kleinen Ausschnitt zeigen (vgl. Abb. 3.1) zu der Zweierkombination *Elbe ∘ gelegen*, sowie der feiner granularen Dreierkombination *Elbe ∘ gelegen ∘ idyllisch*.

Neben den bereits erläuterten Angaben werden zu den Wortkombinationen stets die Textstellen einblendbar angeboten (verborgen hinter dem ⊞), die den

#	LLR	Häufig	links	rechts	Kookkurrenzen	syntagmatische Muster
⊞ 249	449	5	1	1	**gelegen** idyllisch	60% idyllisch ... an|der ... Elbe [...] gelegen
⊞		127	1	1	gelegen	89% an der Elbe [...] gelegen

Abb. 3.1: Eine ausgewählte Kookkurrenzstruktur zu dem Bezugswort *Elbe*.

analytischen Befund legitimieren. Somit kann das Kookkurrenzanalyseverfahren als ein Gruppier- und Sortierverfahren gedeutet werden, das Textstellen einer Treffermenge mit ähnlichen Kontextmustern zusammenfasst, d. h. nach dem Ordnungssystem, das wir uns im ersten Kapitel gewünscht hatten. Aus den jeweiligen Textstellen einer Gruppe wird als Interpretationshilfe ein syntagmatisches Muster abgeleitet, mit dem versucht wird, die hervorstechende lineare Anordnung relevanter Bestandteile des Musters zu rekonstruieren. Mit all diesen Optionen kann der Nutzer im Grunde vielerlei unterschiedliche Kookkurrenzanfragen an den jeweils zugrunde gelegten Datenbestand formulieren; die Kookkurrenzanalyse ist somit nicht ein einziges Verfahren, sondern steht für eine ganze Sammlung verschiedener Instantiierungen. Ausführlicher beschrieben ist dies in Perkuhn, Keibel & Kupietz (2012: Kap. 8).

Eine grundsätzliche Eigenschaft der Herangehensweise ist jedoch, dass im Zweifelsfall lieber viele Wortverbindungen angeboten werden, ohne dass diese durch gewisse Vorgaben des Nutzers vorweg gefiltert werden – etwa durch syntaktische oder dependentielle Vorgaben und auch nicht durch Angaben wie Wortklassen. Auch wenn aus Sicht vieler Anwendungen gerade dies oft wünschenswert erscheint, nimmt der Ansatz den Nutzer in die Pflicht, selber seine Vorstellungen in qualitative Entscheidungen umzusetzen und die Analyseergebnisse zu interpretieren. Viele der dabei einfließenden menschlichen Bewertungen lassen sich nicht in einem zufriedenstellenden Maß operationalisieren; gerade bei Wortverbindungen lassen sich Beispiele mit einer Genauigkeit von weniger als 0,1% angeben (*Recall* eines kommerziellen Tools bei *werden* plus *sein* als Pronomen). Unabhängig davon sollte es für Forscher auch durchaus lohnend sein, zu schauen, wie die Typen von Wortverbindungen, die sie gerne sehen und untersuchen möchten, sich im Verhältnis zu anderen Wortverbindungen verhalten. Sind alle Verbindungen, die als höher auffällig bewertet wurden, für sie nicht relevant, können diese ignoriert werden. Wenn dies aber nicht durchgängig gilt, wäre ein vorheriges automatisches Ausblenden fatal gewesen.

In speziell einer Hinsicht ist die Untersuchung von Wortverbindungen der deutschen Sprache besonders schwierig: Es lässt sich kaum vorgeben, in welchen positionellen Beziehungen die Elemente zueinander stehen. Man könnte zwar gewisse Erwartungen in Abhängigkeit von Wortklassen grob schematisch in bestimmte Typen einteilen, zumindest was die Beziehungen innerhalb von Präpositional- oder Nominalphrasen angeht, sofern die Wortklassen verlässlich bestimmt werden konnten. Aber selbst dabei kann nicht ausgeschlossen werden, dass z. B. flektierte Adjektivformen nicht immer nur eine Position vor dem Substantiv besonders auffällig sind, sondern durch eine typische Adjektivreihung auch systematisch weiter davor stehen können (z. B. *blonde* [*glatte/gelockte/gewellte*] Haare). Für andere Konstruktionen, wie die prädikative Verwendung von Adjektiven oder Wortpartnerschaften zu Verben, ist dies ohnehin komplizierter. Der gewählte Ansatz ist, den Kontext als Textfenster definieren zu können. Bei großer Unsicherheit kann dieses Fenster auch vorsichtshalber sehr weit ausgedehnt werden, auf Kosten der Antwortzeit auch durchaus größer als der Standardwert von fünf Wörtern links und fünf Wörtern rechts. Da das statistische Verfahren aber die Vorkommen innerhalb dieses (nunmehr deutlich größeren) Ausschnitts mit dem Vorkommen im Korpus insgesamt vergleicht, benachteiligt der große Kontext Phänomene, für die ein kleiner Kontext ausreichend gewesen wäre. Anstatt verschiedene Kontextdefinitionen selber ausprobieren und vergleichen zu müssen, übernimmt diese Funktionalität ein zusätzlicher Mechanismus, der die anderen Optionen ergänzt: der Autofokus. Wählt der Nutzer diesen mit hinzu, so wird der definierte Kontext als maximaler Kontext interpretiert, zu dem alle denkbaren, zusammenhängenden Unterabschnitte getrennt ausgewertet und die darin beobachteten Vorkommen statistisch bewertet werden. Die Grenzen des am höchsten bewerteten Abschnitts werden mit angezeigt (in Abb. 3.1 die Spalten „links“ und „rechts“, mit der Bedeutung: mit Vorzeichen Minus = vor dem Bezugswort, ohne Vorzeichen = nach dem Bezugswort). Das Gesamtbild der nach dem so ermittelten Auffälligkeitsmaß sortierten Kookkurrenzen wird dadurch gewissermaßen leicht verzerrt. Dies kann man vertreten, da die Phänomene, für die ein kleinerer Kontext besser gewesen wäre, davon profitieren und in dem Sinne an Relevanz aufsteigen. Beim Vergleich der verschiedenen Elemente innerhalb eines Kookkurrenzprofils sollte man aber trotzdem im Hinterkopf behalten, dass diese zum Teil Antworten auf verschiedene Fragen darstellen.

Beim Betrachten eines Gesamtkookkurrenzprofils (wie z. B. in Abb. 3.2) zeigt sich der Effekt des sehr weiten, im Wesentlichen statistischen Kookkurrenzbegriffs: Es finden sich verschiedene Arten von Wortverbindungen, die alle statistisch legitimiert, aber sprachlich teilweise sehr unterschiedlich motiviert sind. Neben syntaktischen Konstruktionen (*östlich [der] Elbe*), Paar-

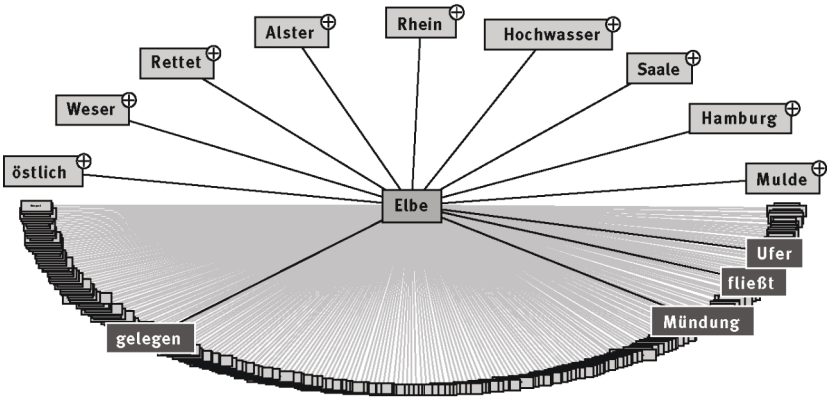

Abb. 3.2: Kookkurrenzprofil des Wortes *Elbe* (CCDB, visualisiert nach Perkuhn 2007a).

formeln oder Ausschnitten aus Aufzählungen, meist von Kohyponymen (*Weser [und] Elbe*) finden sich auch Floskeln oder Namen von Bündnissen oder Ähnliches (*Rettet [die] Elbe*). Einzelne Partnerwörter dazwischen lassen sich über die geografische Verortung der Elbe erklären. Für weitere Mehrwortfügungen wie Kollokationen (im engeren Sinne) und Redewendungen ist unser Beispiel schlecht gewählt; diese ließen sich bei entsprechenden Verben, Adjektiven oder Nomina, die keine Eigennamen sind, analog entdecken. Um das Kookkurrenzverhalten eines Wortes zu erkunden und zur Gänze zu erschließen, bedürfte es also weiterer methodischer Unterstützung (vgl. Perkuhn 2007a, b). Trotzdem ahnt man beim Betrachten schon, dass einige Partnerwörter Hinweise auf die Bedeutung liefern, unabhängig davon, ob es nicht-transparente Fügungen sind, die lange Zeit den Schwerpunkt der Kollokationsforschung ausgemacht haben, oder transparente Kombinationen, die gezielt aus diesem Schwerpunkt herausdefiniert wurden, deren Stellenwert aber gerade für die Fremdsprachendidaktik schon früh gewürdigt wurde – gerade weil sie eben auch zu der Vermittlung von Bedeutung im Kontext beitragen (Bahns 1997: 48). Mittlerweile lässt sich sicher auch erahnen, warum ausgerechnet die fünf Wörter als Umgebungswörter für die Beispiele im ersten Abschnitt gewählt wurden. Vier davon sind auch in dem Kookkurrenzprofil des Wortes *Elbe* vorhanden und in Abbildung 3.2 hervorgehoben, da wir für unseren roten Argumentationsfaden postulieren, dass diese Wörter zu der Bedeutung ‚Fluss' beitragen.

Auch wenn es nach dem vorher Gesagten etwas widersinnig klingt, da wir die Kookkurrenzanalyse als eine Vielzahl von Verfahren beschrieben haben, so haben wir doch eine Standardkonfiguration als prototypisch ausgewählt. Mit dieser wurde bis 2007 eine Datenbank von Kookkurrenzprofilen aufgebaut. In

dieser Kookkurrenzdatenbank CCDB (Belica 2007) werden Methoden entwickelt und experimentell angeboten, die versuchen, die Beziehungen zwischen Wörtern auf Ähnlichkeiten ihrer Profile zurückzuführen.

4 Kookkurrenz(profil)-auswertende Methoden

Nicht nur in der Kookkurrenzdatenbank CCDB, auch in anderen Ansätzen ist ein zentrales Konzept eine zusammenfassende Beschreibung des Kontextverhaltens eines Wortes. Bei diesen ist der Erkenntnismehrwert der Struktur aber untergeordnet unter nachfolgende Auswertungsschritte, die auf dieser Struktur operieren sollen. Die Gesamtheit dieser Struktur wird in Anlehnung an Harris mit *Distribution* bezeichnet:

> The DISTRIBUTION of an element is the total of all environments in which it occurs, i.e. the sum of all the (different) positions (or occurrences) of an element relative to the occurrence of other elements. (Harris 1960: 15 f., Hervorhebung im Original)

Eine Unterscheidung der Verteilung von Instanzen, d. h. von konkreten Formulierungen, die sich zu Kookkurrenzen zusammenfassen lassen, oder von Kookkurrenzen (im Sinne von *Kontexttypen*), die ein Kookkurrenzprofil manifestieren, wird in diesem Zusammenhang nicht vordergründig thematisiert. Das Hauptanliegen dieser Ansätze ist, die Beziehung zwischen zwei Wörtern auf einen Vergleich ihrer Distributionen zurückzuführen:

> If A and B have almost identical environments [...], we say they are synonyms [...]. If A and B have some environments in common and some not [...] we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. (Harris 1970: 786)

Auch wenn wir mit Kookkurrenzprofilen als eine Abstraktion über dem Gesamtkontextverhalten gedanklich schon einen Schritt weiter waren, wollen wir die Idee des Vergleichs analog zu der Struktur in Kapitel 2 noch einmal elementar entwickeln.

Sollte mit Distribution tatsächlich die absolute Frequenz von Vorkommen von Umgebungswörtern gemeint sein, so ergäben sich für die Wörter *Rhein* und *Elbe* die beiden Säulenabschnitte links in Abbildung 3.3 für die Umgebungswörter, die wir auch in Kapitel 3 schon nicht ganz zufällig ausgewählt hatten. In Kapitel 2 war noch dasselbe Wort sowohl als Bezugswort als auch als Umgebungswort aufgeführt. Welchen Beitrag allerdings eine zweite Instanz desselben Wortes in seinem typischen Kontextverhalten zu dessen „Beschrei-

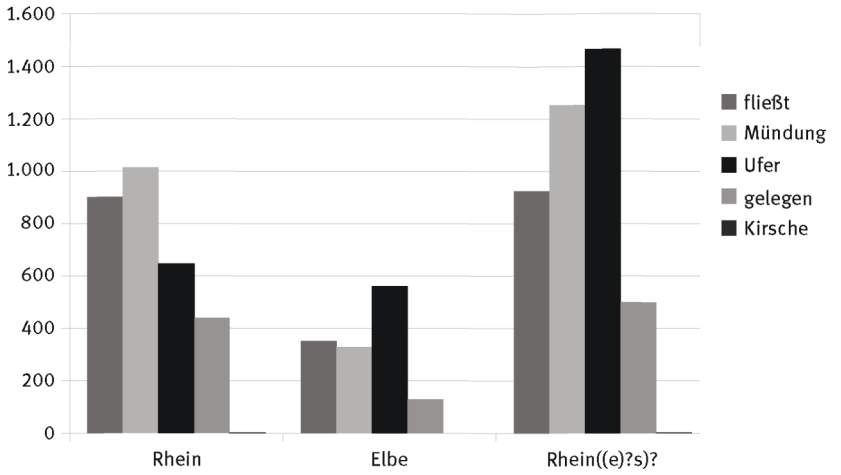

Abb. 3.3: Absolute Häufigkeiten von fünf Umgebungswörtern dreier Bezugswörter.

bung" beitragen kann, ist vollkommen unklar. In manchen anderen Fällen lässt es sich vielleicht plausibel sinnbringend erklären *([von einem] Ufer [zum anderen] Ufer)*; einen neuen Beitrag zum besseren Verständnis des Wortes liefert es aber vermutlich nicht. In dieser und manch anderer Hinsicht sind unsere gewählten Beispiele, insbesondere auch als Eigennamen, etwas speziell – für den eigentlichen Kerngedanken der Argumentation aber sehr gut geeignet.

Wenn es im Folgenden dann aber darum geht, zwei Beschreibungen miteinander zu vergleichen, wird es etwas kritischer, wenn das Vorkommen des Wortes *Rhein* in der Nähe von *Rhein* eine andere Qualität hat als dasselbe Wort in der Nähe von *Elbe*: Während der erste Fall quasi redundant ist, liefert der zweite durchaus einen Hinweis darauf, dass die Wörter paarig oder als Teil einer Aufzählung verwendet werden könnten, was für Kohyponyme durchaus nicht ungewöhnlich ist. Dieser besondere Status der am Vergleich beteiligten Wörter sollte entweder im Vergleichsverfahren, spätestens aber bei der Interpretation berücksichtigt werden. Der Einfachheit halber konzentrieren wir uns im Weiteren nur auf die übrigen Wörter.

Da das Wort *Rhein* viel frequenter ist als das Wort *Elbe*, ließe sich analog zu dem in Kapitel 2 Gesagten schon nicht annehmen, dass die beiden Wörter kongruente Silhouetten ergeben. Denn das wäre das grafische Pendant für den (hier künstlich) kleinen Ausschnitt an Umgebungen, die wir betrachten, zur distributionellen Hypothese. Der nahezu gleichgroße Ausschlag bei dem Partnerwort *Ufer* ist sogar eher störend, da er den Gesamteindruck der Kontur

durchbricht, weil er unverhältnismäßig niedrig liegt. Hintergrund dazu ist ganz einfach, dass die Kombination von *Ufer* typischerweise mit dem Flussnamen im Genitiv vorkommt. Dieser ist bei der Elbe aber nicht markiert *([am] Ufer [der] Elbe*), somit bereits mit berücksichtigt. Beim Rhein ist dies allerdings nicht der Fall *([am] Ufer [des] Rheins*). So, wie die Suchanfrage formuliert war, haben wir den Rhein in der Hinsicht benachteiligt. Nehmen wir Genitivendungen explizit mit hinzu (in Abb. 3.3 mit einem regulären Ausdruck für „*Rhein* oder *Rheins* oder *Rheines*"), stellen wir faire Verhältnisse her. Da nun die Treffermenge noch einmal vergrößert wurde, sind alle Kombinationshäufigkeiten und dementsprechend die Säulen höher. Auch wenn sich jetzt alles auf einem höheren Niveau abspielt, haben wir doch einen sichtbaren Erfolg: Das Umgebungswort *Ufer* ordnet sich nunmehr anders in das Gesamtbild ein und bildet wie bei *Elbe* die höchste Säule. Dass die Säule zum Umgebungswort *Mündung* von der Maßnahme nicht so stark betroffen ist und im Kontrast zu *Elbe* immer noch häufiger als die übrigen Umgebungswörter gebucht ist, liegt einfach daran, dass im Vergleich zu seiner eigenen Mündung (*Mündung [des] Rheins*) viel öfter über andere Flüsse berichtet wird, die in den Rhein münden – und davon gibt es einige (*Mündung [des/der Neckars/Main/Mosel/Ruhr in den] Rhein*). Dieser Effekt ist offensichtlich den Zusammenhängen in der realen Welt geschuldet (und unserer Entscheidung, Eigennamen als Beispiele zu wählen); er verzerrt insbesondere die absoluten Häufigkeiten unabhängig von der (Kern-)Bedeutung des Bezeichneten. Eine vergleichbare Verzerrung können wir uns aber für jedes Wort einhandeln, insbesondere in Abhängigkeit von der Zusammensetzung des zugrunde gelegten Korpus. Je nach Ausrichtung (Textsorte, Textgenre) oder auch durch jeweils aktuellen Einfluss des Tagesgeschehens auf insbesondere Zeitungskorpora, können in bestimmten Diskursen Wörter, Lesarten oder Wortkombinationen (zumindest zeitweise) besonders intensiv genutzt werden.

Zumindest vor den punktuellen Einflüssen kann man sich durch eine breitere und somit meist auch größere Datengrundlage schützen; grundsätzlich gilt aber immer, dass es – wie schon in Kapitel 2 dargelegt – besser ist, mit relativen Häufigkeiten zu argumentieren. Aus diesen lässt sich dann auch der Wert ableiten, der für die Kombination zu erwarten wäre, sowie der Faktor, um wie viel öfter die Kombination beobachtet wurde. In Abbildung 3.4 ist dieser Wert für die Bezugswörter *Elbe*, *Rhein* (im Folgenden nun stets inklusive Genitiv) und *Weichsel* und die bekannten fünf Umgebungswörter dargestellt.

Die Konturen von *Elbe* und *Rhein* haben sich nun wunderbar angenähert. Für Weichsel ahnt man ansatzweise auch eine Ähnlichkeit für die ersten vier Säulen. Da das Wort *Weichsel* aber im Vergleich zu den anderen beiden sehr selten im Korpus vorkommt, schlagen die Säulen etwas überbewertet aus. Im Sinne der gleichen Argumentation wie in Kapitel 2 schaffen wir durch Logarithmisieren eine bessere Vergleichbarkeit, wie sie sich in Abbildung 3.5 zeigt.

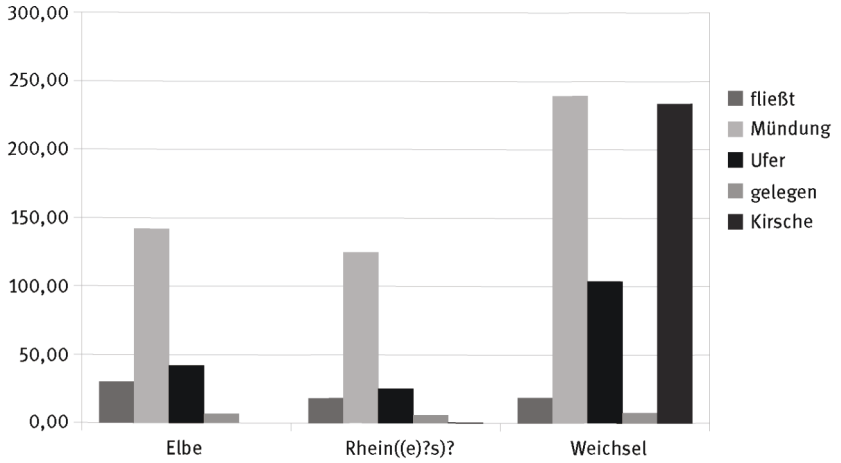

Abb. 3.4: Verhältnis des beobachteten zu dem erwarteten Wert.

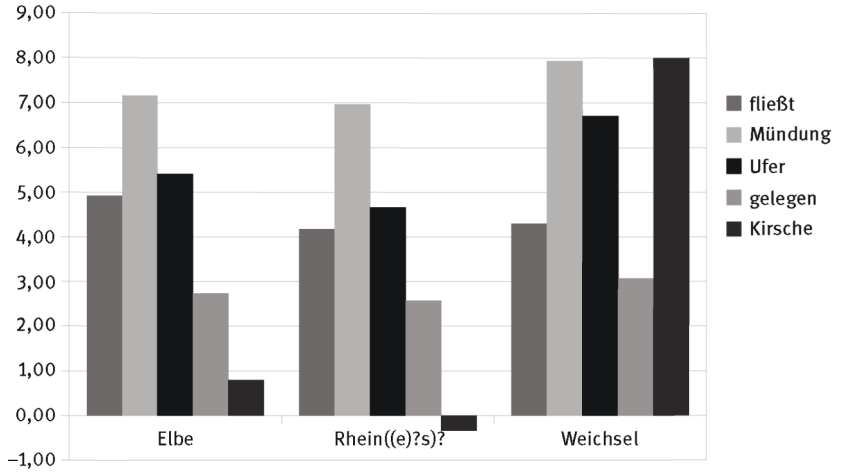

Abb. 3.5: Angaben aus Abbildung 3.4, nur als Zweierlogarithmus, quasi *pointwise mutual information*.

Mit einer gewissen Toleranz zeigen die ersten vier Säulen ein sehr ähnliches Schema. Die fünfte Säule für das Umgebungswort *Kirsche* fällt aus dem Rahmen und zeigt zudem, wie der Trick des letzten Schritts bei kleinen Zahlen plötzlich eine (eigentlich nicht erwünschte) Wirkung zeigt. Der Erwartungswert

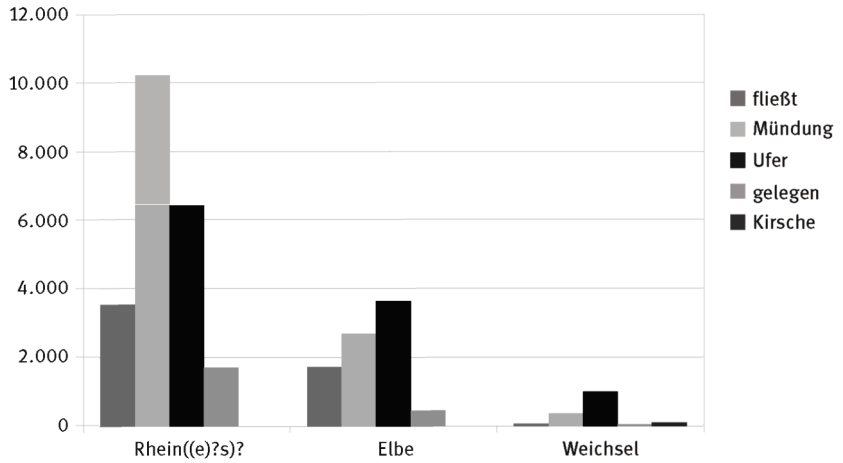

Abb. 3.6: LLR der betrachteten Wortkombinationen.

für *Kirsche* in der Umgebung von *Elbe* lag bei 0,58 und wurde mit tatsächlich beobachteten 0 Vorkommen unterschritten. Durch die Addition von 1 sieht es nun fälschlicherweise so aus, als ob der Erwartungswert leicht überschritten wäre. Der Erwartungswert für *Kirsche* in der Umgebung von *Rhein* lag bei 2,51 und wurde mit tatsächlich beobachtetem 1 Vorkommen unterschritten. In diesem Fall hat der von uns beschönigte Wert von 2 sich nicht so sichtbar niedergeschlagen, da er immer noch unter dem Erwartungswert geblieben ist. Der einzige wirklich beobachtete Treffer ist dabei zwar berechtigt, aber eher kein Stellvertreter für eine systematische, bedeutungsstiftende Verwendung:

> Am ***Rhein*** galt die ***Kirsche*** einst als Markenzeichen. Allein fünf alte Sorten gab es dort in Hülle und Fülle. [RHZ08/JUL.04389 Rhein-Zeitung, 2.7. 2008; Filsen will die Kirsche][1]

Wenn wir einen kurzen Blick auf analoge Darstellungen werfen, die sich aus den quantitativen Angaben der Kookkurrenzanalyse des IDS ableiten lassen, so sehen diese zunächst deutlich weniger beeindruckend aus. Weder das Diagramm, das das Assoziationsmaß LLR aufträgt (Abb. 3.6), noch die Darstellung des inversen Rangs (damit kleine, wichtige Ränge höher abgebildet werden und nicht in der Nähe von 0; Abb. 3.7) zeigen ansatzweise eine schematische Vergleichbarkeit.

1 Die Referenz auf Korpusbelege wird laut bibliographischer Angabe gemäß DeReKo-/Cosmas II-Konvention übernommen.

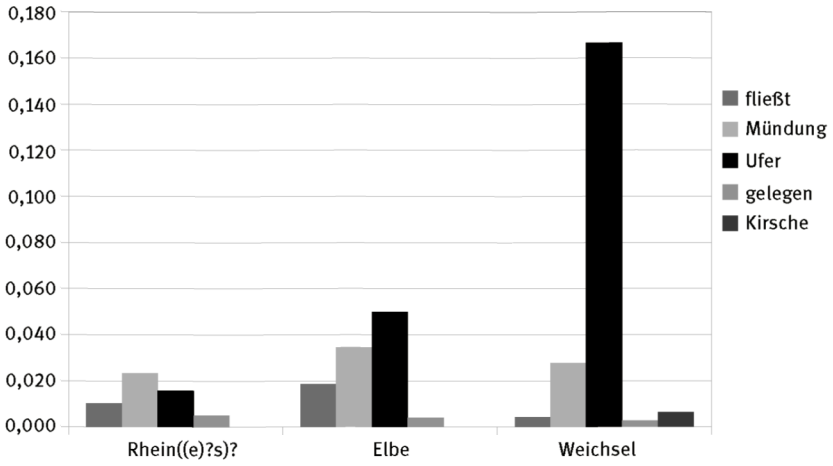

Abb. 3.7: Inverser Rang der betrachteten Wortkombinationen.

Nichtsdestotrotz konnte Belica (2007, 2011) ein Ähnlichkeitsmaß für die Kookkurrenzprofile innerhalb der CCDB entwickeln, das beeindruckende Ergebnisse zeigt – oder vielleicht gerade sogar deswegen, weil ein anderes Modell auf einer abstrakteren Ebene zugrunde gelegt wurde, das die Distribution auf eine weniger starr schematische Weise auswertet.

Für das Ähnlichkeitsmaß werden Werte pro übereinstimmendem Kookkurrenzpartnerwort ermittelt, bei dem sowohl deren pro Analyse berechnete LLR-Werte, ihre Ränge innerhalb und der Umfang des jeweiligen Kookkurrenzprofils, als auch *tf-idf*-analoge[2] Maße einfließen, die bemessen, wie stark der Einfluss des betrachten Kookkurrenzpartners in den übrigen Profilen der Kookkurrenzdatenbank ist – im einfachsten Fall, wie oft er in ihnen vorkommt. Dadurch sollen die Partnerwörter belohnt werden, die nur in ganz wenigen anderen Profilen und auch dort nur mit geringer Auffälligkeit vorkommen. Im Extremfall ist das (gemeinsame) Partnerwort nur in den beiden betrachteten Profilen gebucht und sonst in keinem anderen, sodass es natürlich extrem zu einer Gemeinsamkeit gerade dieser beiden beiträgt. In dem anderen Extremfall, dass das Partnerwort in vielen, womöglich (fast) allen anderen Profilen ebenfalls als besonders auffällig verzeichnet ist, hat es quasi keine diskriminierende

2 *Termfrequenz/inverse Dokumentfrequenz*: Maß aus dem Information Retrieval, das angibt, wie charakteristisch ein Wort für eine Menge von Dokumenten ist (z. B. Schlüsselwörter/Terme für Texte desselben Themas).

Wirkung. Es trägt zu dem dedizierten Vergleich wenig bei und kann durch entsprechende Gewichtung marginalisiert werden. Da die Gewichtungen der einzelnen Parameter sowie auch deren Zusammenführung zu einem Gesamtwert sich aus explorativ-experimentellen Studien abgeleitet haben, lässt sich für die Berechnung keine mathematische Formel angeben, höchstens eine pseudoformale Umschreibung des zuvor Gesagten:

$$Sim(w_1, w_2) = \underset{k_{w_1,x},\, k_{w_2,y} | k_{w_1,x} = k_{w_2,y}}{G} \quad (g(llr(k_{w_1,x}), llr(k_{w_2,y}), rk(k_{w_1,x}), rk(k_{w_2,y}), maxrk(w_1), maxrk(w_2), tfidf(k_{w_1,x}))$$

Formel 3.2: Berechnung des Ähnlichkeitsmaßes in der CCDB.

Abbildung 3.8 zeigt nach diesem Maß sortiert den ersten Abschnitt des Abgleichs aller Einträge der Kookkurrenzdatenbank mit dem vorgegebenen Wort *Weichsel*. Die Liste lässt sich noch ein ganzes Stück verlängern (in der CCDB interaktiv), zeigt aber schon im ersten Teil, dass wir das Umgebungswort *Kirsche* nicht ohne Grund gewählt haben. Bei dem Wort *Weichsel* handelt es sich um Homonym, das zum einen als Bezeichnung eines Flusses verwendet wird, das zum anderen aber auch als Kurzform für *Weichselkirsche* regional eine Obstsorte bezeichnet. Im Vergleich zur Flussbezeichnung gibt es verhältnismäßig wenig Textstellen für die zweite Lesart. In diesen schlummern aber genügend Hinweise auf sie, sodass sich dies auch im Kookkurrenzprofil zeigt – und sei es in diesem Fall auch besonders über Aufzählungen verschiedener Obstsorten. Die Häufung des Wortes *Kirsche* in seiner Umgebung kann aber auch durchaus auf paraphrasische Verwendung zurückzuführen sein.

> Da wird die Sauerkirsche a.k.a. ***Weichsel*** als ***Kirsche*** verkauft, mit einem schlappen Fruchtanteil von zwei Prozent. [FLT13/MAR.00585 Falter, 27.3. 2013, S. 48; Gewagte und weniger gewagte Limos, die erfrischen oder nicht]

Dass das Wort *Kirsche* in der Umgebung von *Weichsel* oft beobachtet wird, ist also kein Zufall, genauso wenig wie dies bei *Birne*, *Himbeere* oder *Marille* der Fall ist. Die Kookkurrenzanalyse deckt dies jeweils auf; und als gemeinsames Partnerwort trägt es zu dem Ähnlichkeitsmaß bei, so wie weitere gemeinsame Partnerwörter. Ein Teil der Partnerwörter trägt also zu der Ähnlichkeit der Obstbezeichner bei; ein anderer Teil, unter anderem die übrigen vier Umgebungswörter aus unserem Beispiel, sorgen für eine hohe Ähnlichkeit zu den Flussbezeichnern. Die Ähnlichkeiten werden sozusagen lokal gemessen anhand einer opportunistisch-selektiven Strategie.

© Cyril Belica: Modelling Semantic Proximity – Similar Collocation Profiles
Folgende verwandte Kookkurrenzprofile zu Weichsel wurden gefunden (anklickbar, absteigend nach Verwandtschaftsgrad sortiert):
Elbe
Neiße
Donau
Birne
Himbeere
Marille
Oberlauf
Pflaume
Pfirsich
Aprikose
Fluss
Fluß
Rhein
Wolga
Kirsche
...

Abb. 3.8: Ausgabe der Methode *Similar Profiles* der CCDB.

Andere Ansätze, die letztendlich auch auf den Vergleich von Wörtern hinarbeiten, gehen zunächst – und dann auch speziell in diesem Punkt – anders vor. Die Kerngedanken dieser Verfahren (vgl. Deerwester et al. 1990) sind stark durch Leitfragen des *Information Retrieval* geprägt, bei dem es zu Beginn weniger um Wörter in der Umgebung von Wörtern, sondern innerhalb von Dokumenten ging. Ziel war eine Klassifikation nach Textinhalten – insbesondere, um Treffer für Suchanfragen anbieten zu können, in denen der Suchausdruck selber gar nicht vorkommen musste, sondern nur bedeutungsähnliche oder -stiftende Ausdrücke. Dazu wurde für jedes Wort[3] des Vokabulars eine Zeile

3 Üblicherweise „Wortform X Wortform" oder „Lemma X Lemma"; Ansätze mit Mischung „Lemma X Wortform", wie in unserem Beispiel, sind uns nicht bekannt, wäre als Option aber wichtig: „It is important, however, to regard each word separately at first, and not as a member of a paradigm" (Firth 1957: 12).

(wie in unseren Tabellen in Kapitel 2), eine komplette Matrix unmittelbar quantitativer oder daraus abgeleiteter Werte ermittelt (vgl. Tabellen im Anhang). Eine einzelne Zeile dieser Matrix wird in dieser Welt als Vektor bezeichnet; jede Spalte der Matrix wird als eine Dimension eines Vektorraums aufgefasst. Aus der distributionellen Hypothese lässt sich dann darauf aufbauend eine schöne Analogie herstellen, bei der die Anschaulichkeit aber jenseits von drei Dimensionen wieder verlorengeht: Jeder Eintrag in einer Zeile entspricht der Koordinate einer Dimension, sodass der durch eine Zeile beschriebene Vektor (nun auch räumlich) einen Punkt in dem durch die Dimensionen aufgespannten Vektorraum darstellt. Haben wir bisher Ähnlichkeit über die Kongruenz der Säulendiagramme diskutiert, so ist das Pendant hier, dass die Vektoren zweier ähnlicher Wörter Punkte beschreiben, die in dem Vektorraum nahe beieinander liegen. Die etwas schwächere Fassung davon dämpft schon die Erwartung, dass man globale Skalierungen für alle Wörter und Dimensionen finden werden kann, und vergleicht nur noch die Richtung der Vektoren, indem die Winkel zwischen ihnen gemessen werden.

Eine große Herausforderung dieser Verfahren, die längst auf Kookkurrenzen übertragen wurden, ist die Erstellung und das Auswerten der Matrix für das gesamte Vokabular. Bei einem geschätzten Umfang unseres Korpusvokabulars von ca. 20.000.000 Token ist selbst bei frequenten Wörtern (z. B. 100.000 Treffer mal 10 Wörter im Kontext) und maximaler Diversität nur ein Bruchteil der Spalten mit Werten gefüllt. Verständlicherweise hat man sich viel damit beschäftigt, diese Matrix auf das zu reduzieren, was wirklich wichtig ist: Hochfrequente, teilweise auch niedrigfrequente Wörter werden ausgeblendet, Wörter, die stets gemeinsam in die gleiche Richtung weisen, zusammengefasst. Die resultierenden Verfahren zeigen gute Ergebnisse für Synonyme, haben aber große Schwierigkeiten bei Polysemie oder Homonymie.

> Despite their usefulness, most VSMs [vector space models, Anm. d. Verf.] share a common problem that each word is only represented with one vector, which clearly fails to capture homonymy and polysemy. (Huang et al. 2012: 873)

Auf unser Beispiel übertragen zeigt sich an *Elbe* und *Rhein*, dass genügend gemeinsame Umgebungswörter eine hohe Ähnlichkeit nahelegen (nur im Ansatz illustriert durch unsere ausgewählten vier Fluss-Umgebungswörter; vgl. Abb. 3.5). Der Einfluss des Umgebungsworts *Kirsche* ist hier vernachlässigbar klein. Beim Vergleich von *Elbe* oder *Rhein* mit *Weichsel* ist dessen ausgeprägtes Vorhandensein dort aber nicht mehr von der Hand zu weisen. Solange *Kirsche* mittelbar oder unmittelbar zu einer Dimension des Vektorraums beiträgt, zeigt der Vektor für *Weichsel* in eine deutlich andere Richtung als für die beiden anderen Flüsse. Die Teilähnlichkeit zu einer Lesart lässt sich kaum aus der

Distanz oder dem Winkel der Vektoren herauslesen. Wird *Kirsche* (und Konsorten) hingegen als Störgeräusch weggeblendet, rutscht *Weichsel* dichter an die Flüsse; für die Gesamtbetrachtung wird allerdings eine Lesart, womöglich auch ein Teil der sie konstituierenden Umgebungswörter, generell ignoriert.

5 Distributionen kondensiert – Word Embeddings

Die gesamte Wort-Kontext-Information in einer vollständigen Matrix aller denkbaren Kombinationen vorzuhalten, ist nicht nur technisch unhandlich; es spricht auch vieles dafür, dass wir Menschen bei der kognitiven Verarbeitung den gleichen Gehalt auf eine wesentlich kompaktere Repräsentation herunterbrechen:

> If all knowledge [...] were stored [...], a person's knowledge of word A would be given by all the contexts in which A had occured. However, most theorists assume [...] that the cognitive representation of a word is some abstraction or generalisation derived from the contexts that have been encountered. (Miller & Charles 1991: 5)

Seit einiger Zeit hat sich ein Forschungsstrang diese These auf die Fahne geschrieben (Lapesa & Evert 2014; Levy & Goldberg 2014). Er knüpft zwar gedanklich an viele Punkte der Vektorraummodelle an, versucht aber vor allem, diese kompakte Repräsentation über neuronale Netze zu lernen. Insbesondere der Methodenapparat *word2vec*, der in der Forschergruppe um Mikolov (Mikolov et al. 2013) entwickelt wurde, und Ableger davon haben der Richtung einen gewaltigen Schub gegeben und viel Hoffnung geweckt.

Stark vereinfacht kann man sich das neuronale Netz vorstellen wie in Abbildung 3.9: Es gibt eine Eingabe- und eine Ausgabeschicht (*input/output layer*), die jeweils das gesamte Vokabular abbilden. Interessant ist aber vor allem die mittlere, versteckte Schicht (*hidden layer*), die wesentlich weniger Knoten enthält als die beiden anderen; in einigen Studien werden kleine Hunderterzahlen (z. B. 300) genannt. Jeder Knoten dieser Schicht enthält so viele Felder, wie es Wörter gibt. Beim Training über die Korpuskontextinformationen lernen diese Felder Gewichtungen, die dahingehend ausgerichtet sind, möglichst gut die Kombinationshäufigkeit mit den Wörtern der Ausgabeschicht vorherzusagen. In einer Modellvariante soll das Bezugswort anhand der Kontexte vorhergesagt werden (CBOW, *continous bag of words*); in der oben dargestellten Variante soll der Kontext anhand des Bezugswortes vorhergesagt werden (*SkipGram*, eigentlich für die Beschreibung der vorgegebenen Textfenster-/Abstandsmetrik[4]).

4 Vergleichbar mit einer Textfensterdefinition ohne Autofokussierung, unter Umständen aber mit einer Ausdehnung der Kontextbreite bei Filterung von Stoppwörtern.

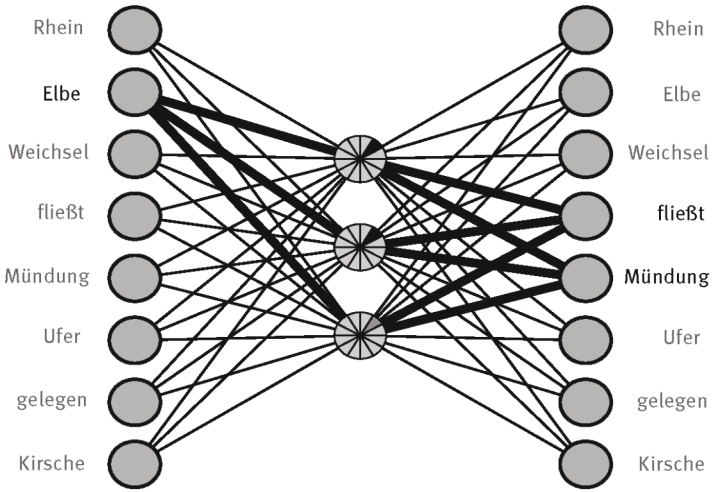

Abb. 3.9: SkipGram-Modell zum Erlernen von Word Embeddings.

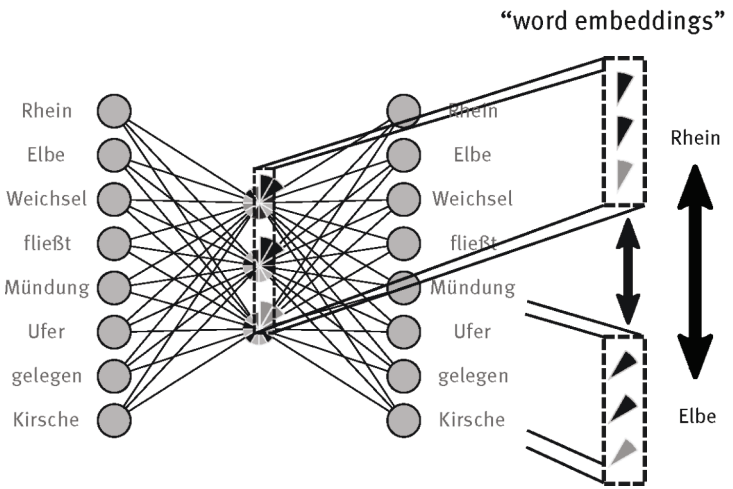

Abb. 3.10: Gelernte *Hidden Layer*-Gewichte als *Word Embeddings* zum Vergleich.

Liest man nach dem Training jeweils die Felder aus, die einem Wort zugeordnet sind, erhält man wieder einen Vektor, der zwar wesentlich kleiner ist als der Ausgangsvektor für alle seine Kombinationen, der aber immer noch eine große Vorhersagekraft für das Kontextverhalten hat. Diese gelernten Vektoren, *word embeddings* genannt, dienen nun zum Vergleich: Eng verwandte Wörter

sollten auch ähnliche Vektoren haben, die sich wiederum räumlich deuten und eventuell auf zwei Dimensionen verflachen lassen, um diese Beziehung auch zu visualisieren.

Wofür die Knoten der Zwischenschicht allerdings stehen, ist bisher noch weitestgehend unklar. Obwohl quasi subsymbolisch entstanden, kursiert die Hoffnung, dass sie quasi als Merkmalsbündel operieren, was stichprobenartig durch algebraische Betrachtungen der Vektoren genährt wird: So könnte etwa für die *word embeddings* für *gehen*, *rennen* und *schnell* die Gleichung aufgestellt werden: v(*rennen*) = v(*gehen*) + v(*schnell*). Studien zum Deutschen stehen allerdings noch aus, für Auswertungen englischer Korpora werden entsprechende Beispiele in der Literatur genannt. Eine systematische Untersuchung liegt jedoch noch nicht vor. Interne experimentelle Studien deuten allerdings an, dass auch diese Ansätze die Herausforderung der Polysemie nicht besser in den Griff bekommen. Sie erben gewissermaßen die Last, ein mehrdeutiges Wort in einem einzigen Vektor darzustellen, der dann in einem einheitlich dimensionierten und skalierten Raum angeordnet werden soll. Vielleicht ist auch gerade die Vorhersagerichtung und das Berücksichtigen negativer Beispiele, was sich für Bedeutungsnähe als besonders geeignet erwiesen hat, für Polysemie gerade nachteilig: Wenn die meisten (auch absolut gezählten) Kontexte auf *Weichsel* als Fluss hinweisen, ist die Vorhersage von *Kirsche* aus *Weichsel* womöglich exotischer, als andersherum. Wenn dann noch bei den meisten Flüssen als negativ dazugelernt wird, dass das Wort *Kirsche* in der Umgebung quasi nicht vorkommt, spricht die hohe Zahl an Kombinationen dagegen, dass *Weichsel* ein Fluss sein könnte.

6 Fazit – Kookkurrenz, Distribution und Polysemie/Homonymie

In den beiden Welten, in der der Kookkurrenz und in der der Distribution, steht die Auswertung der nahen sprachlichen Kontexte – vollkommen zurecht – im Mittelpunkt. Kookkurrenzbasierte Ansätze wie in der CCDB können den Methodenapparat für das Ermitteln des typischen Kontextverhaltens und den des (nachfolgenden) Vergleichs jeweils getrennt gestalten und so auch auf etwas speziellere Konstellationen eingehen. Durch gewissermaßen lokal angepasste Skalen können Ähnlichkeiten zwischen *Weichsel* und *Rhein*, aber auch zwischen *Weichsel* und *Kirsche* zunächst getrennt festgestellt werden. Erst auf der Grundlage dieser Mosaiksteinchen entsteht ein Gesamtbild in einem weiteren Schritt. Distributionelle Ansätze legen den Grundriss des Gesamtbildes durch

eine global-einheitliche Dimensionierung und Skalierung von vornherein fest. Jede sprachliche Einheit muss sich unter dieser Vorgabe an irgendeiner Stelle eindeutig verorten. Bei einem mehrdeutigen Wort lässt sich dieser innere Konflikt nicht so auflösen, dass räumliche Nähe und semantische Nähe miteinander einhergehen. Wird eine Lesart bevorzugt, rückt zwar die Position zu deren Pol; die anderen Lesarten werden aber ignoriert. Werden mehrere Lesarten berücksichtigt, überlagern sich deren Vektoren (oder *Embeddings*), sodass sich die Position in die Zwischenräume verschiebt. Die Überlegungen, wie man diese Herausforderung in den Griff bekommt, zielen darauf, mehrere Darstellungen der Wörter (sozusagen pro Lesart eine) abzuleiten. Dazu werden dem Verfahren die (Anzahl der) Lesarten vorgegeben, die es über ein Training entsprechend klassifizierter Kontexte auseinanderhalten soll. Dies kann im einfachen Fall die Zuordnung Kontext-Lesart sein. Andere Vorschläge gehen noch einen Schritt weiter und bieten als Informationsressource stattdessen eine Klassifikation der Texte an – nach Themen oder Diskursen ausgerichtet –, davon ausgehend, dass diese sogenannten *globalen Kontexte* die Manifestation der Lesarten bedingen. Kookkurrenzen können dann nach auffälliger Verteilung über ihre globalen Kontexte zumindest schwerpunktmäßig damit verbundenen Lesarten zugeordnet und die Vektoren entsprechend differenziert werden. In dem der CCDB (und Kookkurrenzanalyse des IDS allgemein) zugrunde liegenden korpuslinguistischen Paradigma wird aber die Vorgabe von Lesarten sehr kritisch gesehen, ähnlich wie es Kilgarriff formuliert hat:

> Where 'word senses' have a role to play in a scientific vocabulary, they are to be construed as abstractions over clusters of word usages. (Kilgarriff 1997: 112)

Ähnlich verhält es sich mit der Kraft im Hintergrund, die für ihre Manifestation verantwortlich zeichnet, und die – nicht ganz zufällig – auch im CCDB-Jargon *globale Kontexte* genannt wird. Unser Ordnungssystem, und hier denken wir speziell auch an das kognitive, mag an vielen Stellen in unserem Weltwissen verwurzelt sein. Abgesehen vom unmittelbar selbst Erlebten leiten wir aber auch dies großenteils aus sprachlich Vermitteltem ab und bauen es ständig aus. Insofern ist auch dieses Ordnungssystem teilweise ein Abbild unseres Spracherlebens. Anstatt zu beanspruchen, dies vorgeben zu können, plädieren wir dafür, dass die jeweils wirkenden globalen Kontexte aus der Sicht des betrachteten Wortes auf die Sprache heraus zu interpretieren sind. Dazu bietet die CCDB ein Verfahren an, das die Liste ähnlicher Profile eines Wortes als eine *self-organizing map* kart(ograf)iert. Die Elemente der Liste werden paarweise mit derselben Ähnlichkeitsmetrik wie oben verglichen und dann nach diesem Maß gruppiert und räumlich arrangiert.

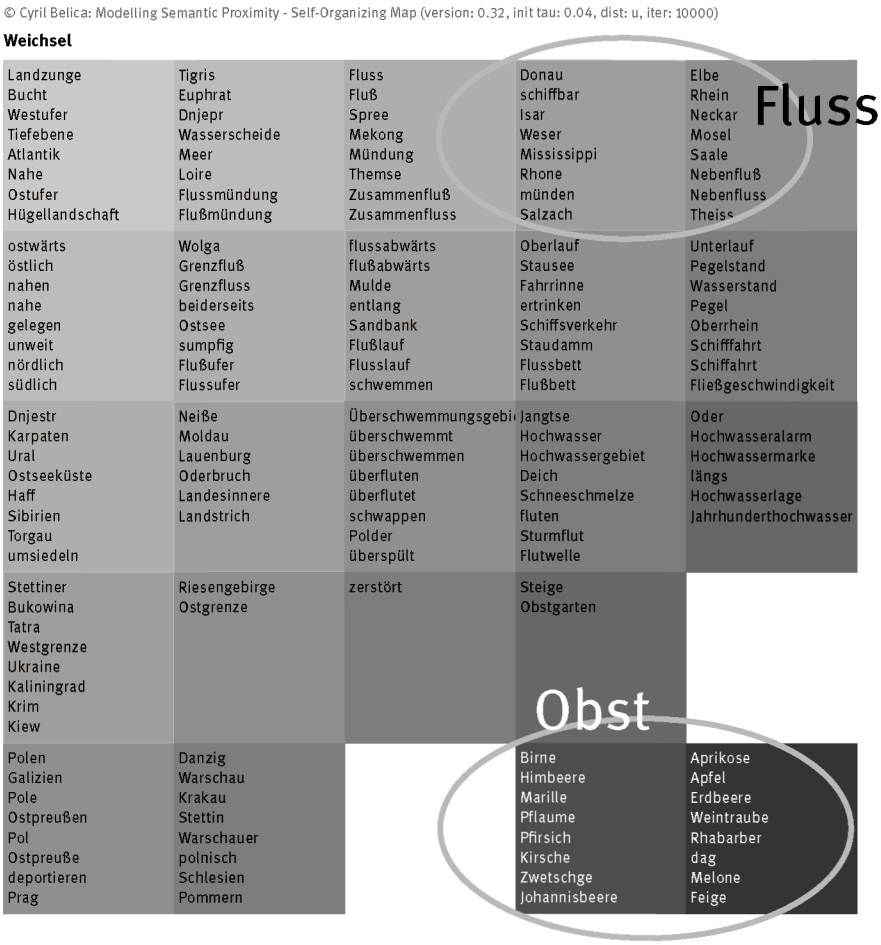

Abb. 3.11: Mit globalen Kontexten annotierte SOM des Wortes *Weichsel* aus der CCDB.

In der Karte (vgl. Abb. 3.11) zeigt sich die Stärke dieses Ansatzes, dass das Erstellen und das Vergleichen der Kookkurrenzprofile zunächst nach lokalen Prinzipien erfolgt, bevor danach eine räumliche Anordnung angestrebt wird. Und das, obwohl die ermittelten Kennzahlen (vgl. Abb. 3.6 und Abb. 3.7) keine leichte, unmittelbare Vergleichbarkeit suggeriert hätten. Erste interne Studien mit auf *embeddings* basierenden Darstellungen zeigen keinen Bereich für die Lesart ‚*Obst*', auch wenn es vom Ansatz an dieser Stelle theoretisch auch wieder möglich wäre, wenn die Ähnlichkeiten zu beiden Lesarten, wenngleich stark reduziert, erfasst worden wären. Vielleicht erliegt man in den Vektorraumvorstellungen aber auch gerade der Versuchung, dass die Näheanalogie,

die sich bei enger Bedeutungsverwandtschaft erzielen lässt (als Kontur des Säulendiagramms oder als Punkte im Vektorraum) auf andere Form-Bedeutungsbeziehungen übertragen lässt. Inwieweit dies tatsächlich möglich ist, mit wie vielen und welchen externen oder intrinsischen Zusatzinformationen, das wird noch Gegenstand vieler zukünftiger Untersuchungen in beiden Bereichen sein.

Literatur

al-Wadi, Doris (1994): *COSMAS – Ein Computersystem für den Zugriff auf Textkorpora. Version R.1.3–1.* Benutzerhandbuch. Mit einem Geleitwort von Prof. Dr. Gerhard Stickel. Mannheim: Institut für deutsche Sprache.

Bahns, Jens (1997): *Kollokationen und Wortschatzarbeit im Englischunterricht.* Tübingen: Narr.

Bański, Piotr, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Piotr Pęzik, Carsten Schnober & Andreas Witt (2013): KorAP: The new corpus analysis platform at IDS Mannheim. In Zygmunt Vetulani & Hans Uszkoreit (Hrsg.), *Human language technologies as a challenge for computer science and linguistics*, 586–587. Proceedings of the 6th Language and Technology Conference. Poznań: Fundacja Uniwersytetu im. A.

Belica, Cyril (1994): *A German lemmatizer.* Final Report MLAP93–21/WP2. Luxemburg.

Belica, Cyril (1995): *Statistische Kollokationsanalyse und -clustering.* Korpuslinguistische Analysemethode. http://corpora.ids-mannheim.de (letzter Zugriff 1. 8. 2017).

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 1. 8. 2017).

Belica, Cyril (2011): Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In Andrea Abel & Renata Zanin (Hrsg.), *Korpora in Lehre und Forschung*, 155–178. Bozen-Bolzano: Freie Universität.

Belica, Cyril & Rainer Perkuhn (2015): Feste Wortgruppen/Phraseologie I: Kollokationen und syntagmatische Muster. In Ulrike Haß & Petra Storjohann (Hrsg.), *Handbuch „Wort und Wortschatz"*, 201–225 (= Handbücher Sprachwissen 3). Berlin, Boston: de Gruyter.

Berry-Rogghe, Godelieve/Geneviève L.M. (1973): The computation of collocations and their relevance in lexical studies. In Adam J. Aitken, Richard W. Bailey & Neil Hamilton-Smith (Hrsg.), *The computer and literary studies*, 103–112. Edinburgh: University Press.

Bodmer, Franck (2005): COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport* (3). Mannheim: 2–5.

Bodmer Mory, Franck (2014): Mit COSMAS II »in den Weiten der IDS-Korpora unterwegs«. In Institut für Deutsche Sprache (Hrsg.), *Ansichten und Einsichten*, 376–385. 50 Jahre Institut für Deutsche Sprache. Redaktion: Melanie Steinle & Franz Josef Berens. Mannheim: Institut für Deutsche Sprache.

Brückner, Tobias (1983): Programmdokumentation REFER. *LDV-Info* (2). *Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung. Mannheim*: Institut für deutsche Sprache, 1–26.

Brückner, Tobias (1988/1989): Recherchesystem für Verben. *LDV-Info* (7). *Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung*. Mannheim: Institut für deutsche Sprache, 41–67.

Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer & Richard Harshman (1990): Indexing by Latent Semantic Analysis. *Journal of the American Society for information science* 41, 391–407.

Dunning, Ted (1993): Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 61–74.

Fechner, Gustav T. (1860/1907): Elemente der Psychophysik. Leipzig: Breitkopf.

Firth, John R. (1957): A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis*, 1–32. Oxford: Blackwell.

Harris, Zellig S. (1960): *Structural linguistics*. Chicago, London: Phoenix.

Harris, Zellig S. (1970): Distributional structure. In Zellig S. Harris, *Papers in structural and transformational linguistics*, 775–794. Dordrecht: Reidel. (Erstveröffentlichung: Word 1954, 10, No. 2–3, 146–162).

Huang, Eric H., Richard Socher, Christopher D. Manning, & Andrew Y. Ng (2012): Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics*, 873–882. Long Papers-Volume 1: Association for Computational Linguistics.

Institut für Deutsche Sprache (2017a): *Cosmas II*$_{web}$ 2.2.1 (Release Januar 2017). Mannheim: Institut für Deutsche Sprache. https://cosmas2.ids-mannheim.de/cosmas2-web/faces/home.xhtml (letzter Zugriff 5. 7. 2017).

Institut für Deutsche Sprache (2017b): *Deutsches Referenzkorpus* . Archiv der Korpora geschriebener Gegenwartssprache 2017-I (Release vom 8. 3. 2017). Mannheim: Institut für Deutsche Sprache. PID: 10932/00-0373-23CD-C58F-FF01-3. www.ids-mannheim.de/DeReKo.

Keibel, Holger & Cyril Belica (2007): CCDB: A corpus-linguistic research and development workbench. In *Proceedings of corpus linguistics 2007*. Birmingham: University of Birmingham. http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf (letzter Zugriff 1. 8. 2017).

Kilgarriff, Adam (1997): I don't believe in word senses. *Computers and the Humanities XXXI*, 91–113.

Lapesa, Gabriella & Stefan Evert (2014): A large scale evaluation of distributional semantic models: Parameters, interactions and model selection In *Transactions of the Association for Computational Linguistics*, vol. 2, 531–545. https://transacl.org/ojs/index.php/tacl/article/download/457/91 (letzter Zugriff 1. 8. 2017).

Levy, Omer & Yoav Goldberg (2014): Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* https://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf (letzter Zugriff 1. 8. 2017).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean (2013): Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, George A. & Walter G. Charles (1991): Contextual correlates of semantic similarity. *Language and Cognitive Processes VI*, 1–28.

Perkuhn, Rainer (2007a): Systematic exploration of collocation profiles. In *Proceedings of the 4th Corpus Linguistics Conference (CL 2007)*, Birmingham. Birmingham: University of Birmingham.

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

Perkuhn, Rainer (2016): Collocation(s) in German minds. In Begoña Sanromán Vilas (Hrsg.), *Collocations cross-linguistically. Corpora, dictionaries and language teaching*, 167–192. Helsinki: Société Néophilologique.

Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. (= UTB 3433). Paderborn: Fink.

Perkuhn, Rainer, Cyril Belica, Holger Keibel, Marc Kupietz & Harald Lüngen (2015): Valenz und Kookkurrenz. In María José Domínguez Vázquez & Ludwig M. Eichinger (Hrsg.), *Valenz im Fokus. Grammatische und lexikographische Studien. Festschrift für Jacqueline Kubczak*, 175–196. Mannheim: Institut für Deutsche Sprache.

Teubert, Wolfgang & Cyril Belica (2014): Von der linguistischen Datenverarbeitung am IDS zur "Mannheimer Schule der Korpuslinguistik". In Institut für Deutsche Sprache (Hrsg.), *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Redaktion: Melanie Steinle & Franz Josef Berens, 298–319. Mannheim: Institut für Deutsche Sprache.

Anhang

Tab. A.3.1: Absolute Häufigkeiten der Umgebungswörter (obere Zeile) zu den Bezugswörtern (erste Spalte).

	Rhein	**Elbe**	**Wechsel**	**fließt**	**Mündung**	**Ufer**	**gelegen**	**Kirsche**	**total**
Rhein	2.166	1.501	58	901	1.013	646	437	1	32.7216
Elbe	1.501	325	91	352	328	560	127	0	82.344
Rhein((e)?s)?	2.303	1.541	61	923	1.253	1.468	498	1	35.8927
Weichsel	58	91	28	16	41	102	11	10	6.109
fließt	901	352	16	188	130	59	20	0	126.253
Mündung	1.013	328	41	130	43	99	92	0	24.917
Ufer	646	560	102	59	99	662	380	0	143.440
gelegen	437	127	11	20	92	380	25	1	207.198
Kirsche	1	0	10	0	0	0	1	28	6.227

Tab. A.3.2: „Einfache Erwartungswerte“ der Umgebungswörter (obere Zeile) zu den Bezugswörtern (erste Spalte).

	Rhein	Elbe	Wechsel	fließt	Mündung	Ufer	gelegen	Kirsche
Rhein	17,99	49,53	25,80	19,39	110,47	12,24	5,73	0,44
Elbe	49,53	42,62	160,84	30,10	142,14	42,16	6,62	0,00
Rhein((e)?s)?	17,43	46,36	24,74	18,11	124,57	25,35	5,95	0,40
Weichsel	25,80	160,84	667,09	18,44	239,49	103,50	7,73	233,73
fließt	19,39	30,10	18,44	10,49	36,74	2,90	0,68	0,00
Mündung	110,47	142,14	239,49	36,74	61,58	24,63	15,84	0,00
Ufer	12,24	42,16	103,50	2,90	24,63	28,61	11,37	0,00
gelegen	5,73	6,62	7,73	0,68	15,84	11,37	0,52	0,69
Kirsche	0,44	0,00	233,73	0,00	0,00	0,00	0,69	642,05

Tab. A.3.3: Werte aus obiger Tabelle (mit Häufigkeit um 1 erhöht) logarithmisiert.

	Rhein	Elbe	Wechsel	fließt	Mündung	Ufer	gelegen	Kirsche
Rhein	4,17	5,63	4,71	4,28	6,79	3,62	2,52	–0,20
Elbe	5,63	5,42	7,35	4,92	7,16	5,40	2,74	0,79
Rhein((e)?s)?	4,12	5,54	4,65	4,18	6,96	4,67	2,58	–0,33
Weichsel	4,71	7,35	9,43	4,29	7,94	6,71	3,08	8,01
fließt	4,28	4,92	4,29	3,40	5,21	1,56	–0,49	0,18
Mündung	6,79	7,16	7,94	5,21	5,98	4,64	4,00	2,52
Ufer	3,62	5,40	6,71	1,56	4,64	4,84	3,51	–0,01
gelegen	2,52	2,74	3,08	–0,49	4,00	3,51	–0,89	0,46
Kirsche	–0,20	0,79	8,01	0,18	2,52	–0,01	0,46	9,38

Tab. A.3.4: Ergebnisse der Kookkurrenzbewertungen der drei Bezugswörter mit fünf Umgebungswörtern (Kookkurrenzanalyse Cosmas II in DeReKo-2017-I).

		fließt	Mündung	Ufer	gelegen	Kirsche
Rhein((e)?s)?	Rang	98	43	64	195	–
	1/Rang	0,01	0,02	0,02	0,01	
	LLR	3.520	10.222	6.415	1.686	–
Elbe	Rang	54	29	20	248	–
	1/Rang	0,02	0,03	0,05	0	
	LLR	1.709	2.673	3.637	453	–
Weichsel	Rang	242	36	6	376	158
	1/Rang	0	0,03	0,17	0	0,01
	LLR	69	360	992	48	98