KATHRIN STEYER

Kookkurrenz

Korpusmethodik, linguistisches Modell, lexikografische Perspektiven

Abstract

Der Beitrag zeigt, wie die im korpuslinguistischen Gesamtkonzept des Instituts für Deutsche Sprache entwickelten und in der praktischen Korpusarbeit konsequent umgesetzten Prinzipien sowie die entsprechenden automatischen Methoden der Korpuserschließung und -analyse für die linguistische Forschung und die Lexikografie fruchtbar gemacht werden können. Im Mittelpunkt steht dabei das Erklärungspotenzial der statistischen Kookkurrenzanalyse, einer automatischen Korpusanalysemethode, die einen sinnvollen Zugang zu sprachlichen Massendaten und damit zu sprachlichem Usus eröffnet. Die Anwendung dieser Methode ermöglicht darüber hinaus die Erfassung, Verifizierung und lexikografische Beschreibung usueller Wortverbindungen auf einer umfassenden empirischen Basis. Es wird grundsätzlich zwischen dem statistisch erhobenen Kookkurrenzpotenzial, also der berechneten lexikalischen Kohäsion zwischen sprachlichen Entitäten, und der nachgelagerten linguistischen Interpretation unterschieden. Die automatische Analyse bringt Kookkurrenzcluster hervor, die nicht nur binäre Relationen zwischen einem Bezugswort und einem Kookkurrenzpartner abbilden, sondern multiple Strukturen konstituieren können. Diese Cluster fungieren als „Bausteine der Kommunikation" und weisen Evidenzen für verschiedenste sprachliche Informationen auf. So können sie semantische und pragmatische Aspektuierungen des Wortgebrauchs, formelhafte Ausprägungen oder auch idiomatische Gebundenheiten indizieren. Schließlich wird in einem Ausblick dargestellt, wie diese Methoden im *elexiko*-Modul ‚Usuelle Wortverbindungen' zur systematischen lexikografischen Erfassung und Beschreibung üblicher Wortverbindungen des Deutschen eingesetzt werden. Ziel ist es, ein korpusbasiertes elektronisches ‚Mehrwortlexikon' für das Deutsche zu erstellen und gleichzeitig neue Einblicke in die Kohäsions- und damit auch in Vernetzungsphänomene des deutschen Wortschatzes zu erlangen.

Vorbemerkung

Der Beitrag ist in zentralen Aspekten ein Resultat der Zusammenarbeit und vieler gemeinsamer Diskussionen mit Cyril Belica, dem Autor der COSMAS-Plattform und Entwickler des Korpusdesigns des IDS.[1] Wir werden

[1] Ich danke in diesem Zusammenhang auch Meike Lauer und Rainer Perkuhn für ihre Unterstützung bei der Erstellung dieses Beitrags. Umfassende Informationen zum IDS-Korpuskonzept und zur COSMAS-Plattform vgl. „Arbeitsgruppe für Korpustechnologie" (2003).

im Folgenden zeigen, wie die in diesem korpuslinguistischen Gesamtkonzept entwickelten und in der praktischen Korpusarbeit konsequent umgesetzten Prämissen und Prinzipien sowie die entsprechenden automatischen Methoden der Korpuserschließung und -analyse (vgl. Punkt 3 und 4) für vielfältige linguistische Perspektiven und für die Lexikografie fruchtbar gemacht werden können. Es wird gleichzeitig deutlich werden, dass mit diesem empirischen Zugang eine strenge Trennung zwischen Linguistik und Lexikografie so nicht mehr haltbar ist. Lexikografie erforderte seit jeher sprachwissenschaftliches Denken, korpusbasierte Lexikografie von heute ist ohne eine systematische Reflexion überhaupt nicht mehr denkbar.

Wortverbindungen werden in der Printlexikografie des Deutschen umfangreich behandelt: Sie fanden Eingang sowohl in die großen einsprachigen Wörterbücher (z. B. GWDS 1999) als auch in Spezialwörterbücher – von Agricolas „Wörter und Wendungen" (1992) über Duden 2 (2001), 11 (2002), 12 (2002) und Schemanns „Deutsche Idiomatik" (1993) bis hin zu Röhrichs „Lexikon der sprichwörtlichen Redensarten" (1991). Nach wie vor fehlt für das Deutsche jedoch eine systematische Erfassung aktuell üblicher Wortverbindungen – und hier vor allem von Kollokationen –, gewonnen aus sehr großen elektronischen Textkorpora, so wie es beispielsweise die großen britischen Wörterbuchunternehmen demonstriert haben.[2] Zur Schließung dieser Lücke will das lexikografische Großvorhaben des IDS *elexiko*, ein im Aufbau befindliches lexikalisch-lexikologisches korpusbasiertes Informationssystem zum deutschen Wortschatz, beitragen.[3] Deshalb erhielten usuelle Wortverbindungen bereits in der konzeptionellen Phase von *elexiko* einen zentralen Stellenwert. Sehr schnell wurde deutlich, dass das Konzept viel weitreichender ist: Aus dem Korpus gewonnene Wortverbindungen sind nicht nur als lexikografischer Beschreibungsgegenstand von Interesse, sondern ermöglichen auch einen wertvollen heuristischen Zugang zu sprachlichem Usus schlechthin. Um dem Anspruch auf Systematik und konsequenten Einsatz der empirischen Analyse von Korpusmassendaten und der entsprechenden lexikografischen Aufbreitung gerecht werden zu können, müssen einige Grundvoraussetzungen vorhanden sein: eine quantitativ ernst zu nehmende Korpusbasis, intelligente Analysemethoden und institutionelle Rahmenbedingungen. Das IDS ist das größte Korpuszentrum in Deutschland und verfügt über eine jahrzehntelange Tradition im Aufbau und in der Auswertung elektronischer Textkorpora, wobei die Korpustechnologie am Institut in den letzten zehn Jahren sowohl quantitativ als auch qualitativ in neue Dimensionen vorgestoßen ist. Das IDS beheimatet die weltweit größ-

[2] Verwiesen sei hier auf die großen Traditionen der englischsprachigen Lexikografie; vgl. u. a. Sinclair 1987, Benson/Benson/Ilson 1997 und ganz aktuell das Oxford Collocations Dictionary von 2002.

[3] Informationen zu *elexiko* unter http://www.elexiko.de

te und aktuellste Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und jüngeren Vergangenheit mit einem Umfang von knapp zwei Milliarden Textwörtern im Jahr 2003.[4] Am IDS wurden und werden komplexe automatische Werkzeuge zur Recherche von sprachlichen Einheiten und zur Analyse von Strukturen in Korpora entwickelt, wobei die Konzepte von Beginn an auf die Verarbeitung von Massendaten ausgerichtet waren. Schließlich bietet sich durch die Möglichkeit, als außeruniversitäre Einrichtung langfristige Studien und Projekte etablieren zu können, die Chance, eine wirklich umfassende und systematische Analyse von sprachlichen Massendaten vorzunehmen.

1. Gratwanderungen

Wortverbindungen werden seit langem vor allem in der Phraseologie, einer inzwischen elaborierten linguistischen Disziplin, und in der etwas jüngeren Kollokationsforschung umfassend behandelt, was die schier unüberschaubare Fülle an Arbeiten und Untersuchungen in diesem Bereich eindrucksvoll zeigt (vgl. Burger 1998). Das Potenzial, das die Mehrwortperspektive vor allem für die semantische und pragmatische Beschreibung von Einzelwörtern bietet, ist auch in den Fokus anderer Forschungsrichtungen der Linguistik gerückt[5], außerhalb der Germanistik vor allem auch im britischen Kontextualismus, in der Korpuslinguistik allgemein oder auch in der französischen Schule der Lexicométrie. Aber es gibt nicht **die** Richtung oder **die** Schule oder gar **die** Wortverbindungstheorie. Diese Heterogenität hat in der Vergangenheit oft auch zu Missverständnissen geführt oder dazu, dass man sich wechselseitig überhaupt nicht wahrgenommen hat, obwohl man über ganz ähnliche Phänomene nachdachte. Wendet man sich in dieser Situation erneut der Erforschung von Wortverbindungen zu, muss man sich auch dieser Verschiedenartigkeit der theoretischen Diskussion stellen und vor allem beweisen, welche neuen Einsichten durch eine alternative empirische Herangehensweise zu erlangen sind.

Genuine Bestandteile unseres Konzepts sind die Entwicklung und die Anwendung mathematisch-statistischer Korpuserschließungsmethoden. Aber auch hier werden wir zu Grenzgängern: Wir tangieren den Bereich der korpusbasierten maschinellen Verarbeitung natürlicher Sprachen, einen Bereich mit eigenen Forschungszielen und einem eigenen Wissenschaftsdiskurs. Im grundsätzlichen Ziel, natürliche Sprache in systematischer Weise mit mathematisch-statistischen Methoden zu analysieren und zu erschließen, stimmen

[4] Zur Zusammensetzung siehe unter http://www.ids-mannheim.de/kt/projekte/korpora.

[5] Besonders verweisen möchten wir auf Heringer (1999), der mit seinem sehr pointiert dargestellten Konzept der distributiven Semantik einen empirischen Zugang zur Bedeutung über die Partnerwörter in der Umgebung eines Wortes zeigt.

wir völlig überein. Wir unterscheiden uns aber in der Art und Weise, wie wir mit den Resultaten umgehen. Die maschinelle Sprachverarbeitung beurteilt die Resultate u. a. daraufhin, ob sie formalisierbar sind und vom Rechner weiterverarbeitet werden können (z. B. für die Spracherkennung oder die maschinelle Übersetzung). Unsere Bewertung der Analyse**resultate** unterliegt nicht diesen strengen Formalisierungskriterien, da sie für Lexikografen, Linguisten und nicht zuletzt auch für die Wörterbuchbenutzer gedacht sind. Obwohl auf beiden Seiten das gleiche Maß an Operationalisierbarkeit von Analyse**methoden** verlangt wird und es partiell ein identisches Repertoire an Termini, Algorithmen und mathematischen Modellen gibt, handelt es sich derzeit noch um zwei verschiedene Perspektiven.

Ganz generell existiert nach wie vor eine Kluft zwischen der Entwicklung von Korpustechnologien und automatischen Analysemethoden einerseits und der linguistischen Adaption und reflektierten Interpretation andererseits, was u. E. vor allem mit Folgendem zu tun hat: Wenn Linguisten diese interdisziplinäre Herausforderung annehmen, stellen sich spätestens dann bei vielen Zweifel ein, wenn sie mit Resultaten konfrontiert werden, die sie nicht zu deuten wissen, die nicht ihrer Erwartung, ihrer Intuition oder den ihnen vertrauten linguistischen Modellen bzw. Konstrukten entsprechen. Computerspezialisten haben andererseits ein gewisses Misstrauen gegenüber interpretativen Verfahren, wie sie in vielen Zweigen der Linguistik üblich sind. Soll die Zusammenarbeit erfolgreich werden, müssen beide Seiten Barrieren überspringen: Der Linguist muss dem Rechner mehr zutrauen, als er das in der Regel tut. Er muss vor allem seiner eigenen Sprachkompetenz und Sprachintuition misstrauen und sich möglichst unvoreingenommen auf „Rechenprozesse" einlassen. Der Computerspezialist muss dagegen die Grenzen seines Rechners mehr akzeptieren, als er das in der Regel tut. Sprache ist in ihrer Vielfalt und Lebendigkeit nur bis zu einem gewissen Grade formalisierbar, und die deutende und interpretierende Hand des Linguisten bleibt für viele Zwecke letztlich immer unabdingbar.

2. Warum interessieren uns Wortverbindungen?

Wie vielfältig sich Wortverbindungen in Texten darstellen, illustriert folgendes Beispiel aus der „Berliner Zeitung" vom 22.7.2000:

(1)

Jetzt **braucht es** nur noch mutige Politiker. Wo sind sie? [...] Ich muss vorausschicken – **im Allgemeinen** bin ich absolut kein Befürworter dieser sonst so rechtsgerichteten Partei. Aber was hier die CSU macht, ist nicht nur ein **cleverer Zug**, es ist eine Niederlage der rot-grünen Regierungskoalition. Während Berlin über **das Was und Wie** lange redet, machen die Bayern **Nägel mit Köpfen**. Und es sind – um in **diesem Bild zu bleiben** – verchromte Nägel, während **Rot-Grün** noch darüber diskutiert, wie das Verrosten ihrer Nägel hinausgezögert beziehungsweise die **von vornherein** angeroste-

ten Nägel kaschiert werden können. Dass auch Bayerns „blue card" **Mängel aufzeigt** und dass die Bavaren **darauf achten**, nur für sich nützliche und brauchbare Fachkräfte **ins Land zu holen, kann nicht als restriktiv angesehen werden** – dies wird **wohl eher normal** sein als das von vornherein untaugliche Herumgerede von Rot-Grün. Alexander Merbusch, Berlin

(2) Folgende Wortkombinationen bzw. Phrasen können als fixiert angesehen werden:

es braucht
im Allgemeinen
cleverer Zug
das Was und Wie
Nägel mit Köpfen
im Bild bleiben
von vorneherein
Rot-Grün
Mängel aufzeigen
darauf achten
ins Land holen
kann nicht als [restriktiv] angesehen werden
wohl eher normal

Diese Wortverbindungen weisen einen erkennbaren Grad an Fixiertheit auf, der sich jedoch nur schwer aus einer regelhaften, systembedingten Gebundenheit ableiten lässt. Ihre Fixiertheit kann nur im Gebrauch begründet liegen. Es handelt sich um Verbindungskonventionen zwischen Wörtern, um Traditionen des Formulierens und damit um Gebrauchskonventionen (vgl. u. a. Schmidt 1995). Der Grad ihrer Beherrschung spiegelt die Fähigkeit wider, kulturell angemessen zu kommunizieren (vgl. Stubbs 1997). Es handelt sich um die idiomatische Prägung, wie Feilke dieses Prinzip nennt (u. a. 1996, 1998). Es geht um Formelhaftigkeit – so wie sie u. a. Stein (1995) beschrieben hat – und Vorgefertigtes/Vorformuliertes im Sinne von Gülich/Krafft (vgl. u. a. 1998). In diesem Zusammenhang ist besonders die Erkenntnis wichtig, dass es sich bei diesen Gebrauchskonventionen nicht um Abweichungen oder Irregularitäten handelt, sondern um Variation, die nicht als Sonder-, sondern als Regelfall gelten können, wie Stubbs betont (1997, S. 153).

„Es geht nicht nur um feste Wendungen, in denen die einzelnen Wörter festgelegt sind, sondern um zentrale Verwendungen der Wörter. Bei solchen Kombinationen geht es um Wahrscheinlichkeiten, Erwartungen, (sic!) und quantitative Verteilungen. Es geht um Normen des Sprachgebrauchs." (ebd., S. 157)

Der Muttersprachler hat solche Kombinationsmodi im Zuge des Spracherwerbs mitgelernt und kann Wortverbindungen in der Regel ohne Probleme

aktualisieren und auch neue Wortverbindungen decodieren. Einem Fremdsprachenlerner fehlt diese kookkurrenzielle bzw. idiomatische Sprachbiografie. Er muss übliche Wortverbindungen lernen. Dies haben die Lexikografie und vor allem die Fremdsprachendidaktik seit langem erkannt. Einen umfassenden Überblick dazu gibt Bahns (1996, 1997). Verwiesen sei hier natürlich besonders auf die Arbeiten von Hausmann (u. a. 1984, 1985, 1993, 2004), dem das große Verdienst zukommt, die Kollokationen auch im deutschsprachigen Raum aus ihrem Schattendasein an der Peripherie der Phraseologie herausgeholt und ihre Relevanz als Entitäten zwischen Sprachsystem und Sprachgebrauch, vor allem für den Fremdsprachenerwerb, ins allgemeine Bewusstsein gerückt zu haben. Wenn man die besonders typischen sprachlichen Einheiten in der Umgebung eines Wortes erfasst, erfährt man auch etwas über das Wort selbst, über seine Stellung in Text und Diskurs. Diesen Aspekt werden wir in der weiteren Argumentation wiederholt aufgreifen. Wortverbindungen interessieren uns also

- als Gegenstand der Analyse und Beschreibung (Lexikografie/Fremdsprachendidaktik)
- als heuristischer Zugang zu Bedeutungen und aktuellen Verwendungskontexten, d. h. zu sprachlichem Usus.

Die **Methoden**, mit denen Wortverbindungen erfasst und beschrieben werden, sind dabei so vielfältig wie der Beschreibungsgegenstand selbst. Klassische Zugänge sind z. B. kompetenzbasierte Modelle und empirische Verfahren der Beispiel- und Textanalyse, Befragungen oder auch Assoziationstests. Die Entwicklung der Korpustechnologie ermöglicht seit einiger Zeit, diese Erkenntnisse auf eine umfassende empirische Basis zu stellen. Den heuristischen Wert von Korpusdaten bestreitet heute wohl kaum noch jemand.[6] In Textkorpora findet man geronnenes sprachliches Wissen. Sie bilden „Querschnitte eines Diskursuniversums, das virtuell alle Kommunikationsakte einer zu definierenden [...] Sprachgemeinschaft enthält." (Teubert 1999, S. 296) Es ist die Aufgabe des Linguisten, den Skopus des Diskursuniversums so einzugrenzen, dass es auf ein handhabbares Korpus reduziert werden kann (ebd.). (vgl. Punkt 3)

Mit finanziellen Mitteln in teils beachtlicher Dimension werden überall auf der Welt sehr massive Anstrengungen unternommen, solche Korpusdatenbanken aufzubauen. Nicht zuletzt die in den letzten Jahren exorbitant angewachsenen externen Zugriffszahlen auf die IDS-Korpora sind ein ganz konkretes Zeugnis für die Entwicklung. Auf Grund der Vielstimmigkeit in

[6] In letzter Zeit zeichnet sich auch in Deutschland eine deutliche Hinwendung zu Korpora als wichtiger empirischer Datenbasis ab, wie beispielsweise der Sonderforschungsbereich 441 „Linguistische Datenstrukturen: Theoretische und empirische Grundlagen der Grammatikforschung" an der Universität Tübingen zeigt. Zum heuristischen Wert der Korpora vgl. auch Boguraev/Pustejovsky 1996.

diesem Bereich erscheint es uns zunächst notwendig, unsere Position zu Prinzipien der Korpusbasiertheit kurz zu umreißen.

3. Ist das Korpus die Botschaft?

Unter Korpusbasiertheit werden in der Regel sehr verschiedene empirische Prinzipien subsumiert. Nach wie vor dominiert jedoch das traditionelle Paradigma der Korpusbasiertheit, das wir als ‚**Konsultationsparadigma**' bezeichnen. Das Konsultationsparadigma ist durch folgende Vorgehensweise geprägt: Man hat ein Ausgangsproblem und befragt das Korpus.

> „Die A-priori-Hypothese wurzelt fest in der Entscheidung, nach dem Konkreten zu suchen [...]. Als Determinante dieser Vorgehensweise sind somit konkrete, aus der sprachlichen Intuition stammende Erwartungen zu erkennen. Als Ergebnis erhofft man sich die Vertiefung bzw. Abschwächung des Glaubens an die Ausgangshypothese." (Belica 1998, S. 31)

Das Korpus wird in diesem Sinne z. B. konsultiert, um etwas zu erfahren über

- die Existenz eines Phänomens
- die Häufigkeit des Auftretens eines Phänomens
- die Erstdatierung und den historischen Wandel eines Phänomens
- die Begrenzungen für das Auftreten eines Phänomens (Textsorten, Stilebenen, areale Besonderheiten usw.)

Das Korpus dient darüber hinaus in der Regel als Belegsammlung und ‚elektronischer Zettelkasten' im Fall der lexikografischen Anwendung. Für viele Bedürfnisse und Zwecke der linguistischen Forschung ist diese Herangehensweise durchaus sinnvoll und auch vollkommen ausreichend. Problematisch wird sie aber unter folgendem Aspekt: Man sieht nur, was man sehen will bzw. man findet auf diesem Wege nur das, wonach man konkret sucht. Dies bedeutet, dass man letztlich doch vor allem auf die eigene Sprachkompetenz angewiesen ist. Dadurch können sich unerwartete, weil für den individuellen Sprecher noch nicht aktuell präsente, aber durchaus schon usuelle Phänomene oder solche Phänomene, die sich nicht automatisch auf die Ausgangshypothesen beziehen, durchaus unserem Blick verschließen. Unsere Tests zeigen immer wieder, wie sehr die eigene Intuition täuschen kann und wie sehr sie sich von Sprecher zu Sprecher unterscheidet.

Wir verfolgen ein anderes Prinzip der Korpusbasiertheit, das sowohl die IDS-Korpusphilosophie als Ganzes – einschließlich der entsprechenden praktischen Schritte – als auch die Methodenentwicklung maßgeblich bestimmt hat und bestimmt. Wir haben dieses Prinzip der Korpusbasiertheit ‚**Analyseparadigma**' genannt. Die Vorgehensweise im Analyseparadigma ist eine andere als im Konsultationsparadigma: Es werden systematisch große Sprachausschnitte auf der Suche nach usuellen sprachlichen Phänomenen analysiert. Dies geschieht mit dem Ziel der Erfassung des tatsächlichen

Sprachgebrauchs in seiner ganzen Komplexität. Usuelle Phänomene können sehr verschiedener Natur sein: In erster Linie sind wir auf der Suche nach üblichen Wortverbindungen, die uns Auskunft geben über aktuelle Bedeutungs- und Verwendungsmuster von sprachlichen Entitäten. Wir finden mit unserer empirischen Methode aber z. B. auch usuelle grammatische Gebrauchsmuster oder diachrone Besonderheiten heraus. Sprachliche Strukturen, Eigenschaften und Zusammenhänge werden also nicht mittels Introspektion, sondern anhand sprachlicher Massendaten aufgedeckt und beschrieben; sie werden nicht vorher erahnt, sondern erst entdeckt.[7] Die Analyse erfolgt zunächst mit mathematisch-statistischen Methoden und – aus Rechnersicht – ohne Vorannahmen bzw. vorher aufgestellte Hypothesen gemäß dem Prinzip der ‚minimalen Annahme'.[8] Das Korpus wird dabei nicht „prästrukturierend" mit linguistisch bereits interpretierten Metadaten angereichert, so wie das z. B. bei getaggten – meist morpho-syntaktisch annotierten – Korpora der Fall ist. Vielmehr betrachtet der Rechner natürliche Sprache so wie sie ist und sucht dabei nach statistisch relevanten Auffälligkeiten in der Welt der Zeichenketten, die Evidenzen für usuelle Sprachgebrauchsphänomene darstellen. Dies muss natürlich in einem wechselseitigen Prozess von automatischer Erkennung sprachlicher Ereignisse und menschlicher Interpretation dieser beobachteten Ereignisse geschehen, ein wechselseitiger Prozess, der zu einer Optimierung der Rechenprozesse führen soll. Das Korpus ist hier also mehr als eine Belegsammlung; es ist der unmittelbare Analysegegenstand selbst.

Es stellt sich eine weitere Frage: Kann man sich eigentlich für den sprachlichen Usus interessieren und dann aber das Korpus analysieren, das – so groß es auch sein mag – immer nur eine Stichprobe der Sprache darstellt? Ein solches Vorgehen ist dann legitim, wenn die Stichprobe in Bezug auf ein ganz konkretes Untersuchungsziel repräsentativ für die Gesamtheit ist und daher die gefundenen Resultate extrapoliert werden können, also auf die Gesamtheit der Sprache oder des zu beschreibenden Sprachausschnitts übertragbar sind. Da es sich bei der Sprache um ein komplexes, heterogenes und in seiner Gesamtheit nie vollständig erfassbares Phänomen handelt, brauchen wir für unsere Zwecke eine Korpusstichprobe mit der höchstmöglichen Quantität und der höchstmöglichen Variabilität und Vielfalt der Textmerkmale.[9]

---

[7] Dazu Heringer: „Die Interpretation ist geprägt durch das Empiriedilemma, dass man eine Analyse konzipiert, um bestimmte Fragen zu beantworten, dass man aber das Design der Antwort nicht kennt und erst recht nicht die Antwort selbst beurteilen kann. Man bekommt, was man bekommt. Man muss versuchen, das Ergebnis zu verstehen; man muss entdecken, wie man es praktisch nutzen kann; und man muss diese Nutzung verbessern. Der Empiriker ist ein entdeckender Anarchist. Er könnte nichts Neues entdecken, wenn er sich strikt an das Bestehende hält, an die bestehenden Regeln und Methoden, an die bestehenden Kenntnisse." (1999, S. 105).

[8] Zum sinclairschen Prinzip der ‚minimalen Annahme' vgl. auch Belica 1997, Belica 1998.

[9] Eine so umfassende Korpusstichprobe kann dann die Basis für sogenannte ‚virtuelle

4. Kookkurrenz – Auf der Suche nach Strukturen

Es stellt sich nun die Frage, ob wir mit einer solchen Flut an Sprachinformationen, wie sie uns der Rechner zur Verfügung stellt, überhaupt sinnvoll umgehen können. Sehen wir uns dazu ein Beispiel aus dem hochfrequenten Bereich an: Das Wort *Kopf* hat eine Frequenz von über 200000 Treffern. Um sprachliche Informationen zu typischen Verwendungsmustern dieses Lexems zu erhalten, beginnt man, die 10000 KWICs (Kontextzeilen des Suchwortes) der statistischen Zufallsauswahl zu analysieren. Zunächst scheint es so, als habe man alles im Blick. Nach kurzer Zeit verschwimmen jedoch die Konturen; man kann selbst bei dieser reduzierten Trefferauswahl nicht mehr beurteilen, ob es sich um Typisches oder Marginales handelt, ob z. B. eine auffällige und wiederkehrende sprachliche Modifikation einer Redewendung die sprachspielerische Vorliebe eines ganz bestimmten Textautors im Korpus ist oder ob es sich um ein tatsächliches Sprachwandelphänomen handelt. Man braucht also Orientierungs- und Strukturierungshilfen, die eine Interpretation und Bewertung erleichtern. Man braucht statistische Methoden. Das Analyseparadigma der Korpusbasiertheit ist in entscheidendem Maße von der Qualität dieser Methoden abhängig.

Die wichtigste Methode zur Vorstrukturierung stellt dabei die statistische Kookkurrenzanalyse dar, die durch mathematisch-statistische Berechnungen auf der Basis von Wahrscheinlichkeitsannahmen in einer nicht vergleichbaren Schnelligkeit Häufigkeitsbewertungen und Präferenzsetzungen vornimmt, indem sie nach Verteilungen, Auffälligkeiten und signifikanten Zusammenhängen im – aus statistischer Sicht zunächst unstrukturierten Textstrom – sucht.[10] Eines der elaboriertesten Analysewerkzeuge dieser Art stellt die Kookkurrenzanalyse dar, die am IDS entwickelt wurde[11] und seit 1995 auch externen Nutzern über das Internet kostenlos zur Verfügung steht. Am IDS ist sie derzeit in COSMAS II integriert, jedoch auf beliebige Korpora anwendbar. Die COSMAS-Kookurrenzanalyse versucht, Hierarchien

---

Korpora' darstellen. Ausgehend von der Grundüberzeugung, dass es *das* repräsentative Korpus schlechthin nicht geben kann, wurde dieses Konzept von Cyril Belica entwickelt und in der COSMAS-Plattform verwirklicht. Es ermöglicht eine auf das spezielle Analyseinteresse bezogene flexible Zusammenstellung von Teilkorpora. Damit wird die Entscheidung über die Zusammensetzung eines Korpus von der Phase der Akquisition in die Phase der Nutzung verlagert.

[10] Einen sehr guten Einblick in die Welt stochastischer Konzepte, verbunden mit der Darstellung wesentlicher Ansätze auf dem Gebiet der mathematisch-statistischen Kookkurrenzanalysen, gibt Lemnitzer 1997. Vgl. auch Lehr 1996.

[11] Computerprogramm: „Cyril Belica: Statistische Kollokationsanalyse und Clustering. COSMAS-Korpusanalysemodul. © 1995. Institut für Deutsche Sprache, Mannheim". Die Bezeichnung des Analysewerkzeugs als ‚Kollokationsanalyse' wurde in einem frühen Entwicklungsstadium in Anlehnung an die ‚Collocation Analysis' gewählt; vgl. dazu Armstrong 1994. Zur Philosophie, Funktionsweise und Interpretation der COSMAS-Kookkurrenzanalyse vgl. http://www.ids-mannheim.de/kt/projekte/methoden/ka.html und Steyer 2002, 2003a.

von ähnlichen Verwendungen in einer Belegmenge zu entdecken, indem sie Ähnlichkeiten im Kookkurrenzverhalten lexikalischer Entitäten erfasst. Sie versucht, die Kandidaten herauszufinden, die einen statistisch auffälligen lexikalischen Kohäsionsgrad zum Suchwort (wir nennen es im Folgenden ‚Bezugswort') aufweisen. Das bedeutet, diese Wörter kommen im Vergleich zu ihrem Gesamtvorkommen im Korpus auffällig oft in der Umgebung des Bezugswortes vor. Wir definieren ‚Kookkurrenz' also als eine Kohäsionsqualität, die durch mathematisch-statistische Berechnungen ermittelt wird und die dann zu interpretieren ist.[12] Es geht dabei nicht um Frequenzen (um zahlreiches Vorkommen eines Wortes in der Nähe eines Bezugswortes oder um häufiges Miteinandervorkommen einer Wortverbindung), sondern um statistische Auffälligkeiten. Statistisch auffällig kann auch bedeuten, dass ein insgesamt sehr seltenes Wort auffällig oft genau in der Nähe des Bezugswortes vorkommt.[13] Indem die Analyse die ermittelten Kookkurrenzpartner nach dem Grad der lexikalischen Kohäsion zum Bezugswort ordnet, ordnet sie gleichzeitig die Verwendungskontexte, in die diese Kookkurrenzpartner eingebettet sind. Die Kookkurrenzpartner fungieren also als Indikatoren für signifikante Verwendungsmuster, indem sie in eine Hierarchie gebracht werden. Signifikantes wird ins Zentrum gerückt, Unspezifisches marginalisiert. Kookkurrenzen sind somit manifest gemachte Kontextualisierungen.[14] Bezugswort und Kookkurrenzpartner bilden so genannte Kookkurrenzcluster, die in ihrer Gesamtheit wiederum das Kookkurrenzpotenzial eines Bezugswortes konstituieren. Die Basisrelation stellt dabei das Bezugswort und der primäre Kookkurrenzpartner dar, hier verdeutlicht am Bezugswort *Kopf* (angeordnet nach dem Grad der lexikalischen Kohäsion).[15]

(3) Binäre Kookkurrenzcluster mit Bezugswort *Kopf*

Kopf – schüttelt
Kopf – Nägel
Kopf – Dach

---

[12] Unser Terminus ‚Kookkurrenz' ist somit vergleichbar mit ‚collocation' bei Sinclair 1991 und annäherungsweise mit ‚Kovorkommen' im Sinne des statistisch beobachtbaren Ereignisses des gemeinsamen Auftretens von Zeichenketten bei Lemnitzer (1997, S. 120ff.). Schmidt versteht Kookkurrenz in einem etwas engeren Sinne als „Vorkommen der untersuchten Wörter im gleichen Satz und zwar ohne syntaktische Bindung" (1989, S. 177).

[13] So hat das Idiom *über den Löffel balbieren* zum einen eine mit 34 Treffern niedrige Vorkommenshäufigkeit, zum anderen weist das Wort *balbieren* mit 41 Treffern ebenso eine relativ niedrige Frequenz auf. Das Basiselement *Löffel* ist jedoch der kohäsivste Partner von *balbieren* und auch *balbieren* findet sich im oberen Bereich der statistischen Ränge beim Bezugswort *Löffel*.

[14] Vgl. dazu Kilgarriff, der Wortbedeutungen als Abstraktionen von Wortverwendungsclustern (gewonnen aus Korpuszitaten) interpretiert (1997).

[15] Alle folgenden Beispiele wurden mit Hilfe der COSMAS-Kookkurrenzanalyse erhoben.

Kopf – schütteln
Kopf – schüttelte
Kopf – kühlen
Kopf – Hals
Kopf – zerbrechen
Kopf – getroffen

Der binäre Status sagt allein aber noch recht wenig über die tatsächlichen kohäsiven Strukturen aus; es liegen zumeist weitergehende Affinitäten vor, die in einem nächsten Schritt erfasst werden. Die COSMAS-Kookkurrenzanalyse sucht nach weiteren statistisch auffälligen Kookkurrenzpartnern in der Umgebung dieser Cluster.[16] Schließlich treten nicht nur binäre, sondern in vielen Fällen multiple Strukturen zu Tage. Es ergibt sich der statistisch erhobene Kontext von Kookkurrenzclustern, wie folgende Kookkurrenzcluster ausschnittweise verdeutlichen:[17]

(4) Multiple Kookkurrenzcluster mit Bezugswort *Kopf*[18]

Kopf
  schüttelt
    schüttelt ungläubig
    schüttelt verständnislos
    schüttelt fassungslos

  Nägel
    Nägel gemacht
    Nägel machen

  Dach
    Dach Menschen haben verloren

  kühlen
    kühlen bewahren
    kühlen behalten

  getroffen
    getroffen Nagel

[16] Mit der Möglichkeiten, nicht nur binäre Relationen zu erfassen, sondern komplexe und multiple Cluster auffinden zu können, geht die COSMAS-Kookkurrenzanalyse über die meisten Werkzeuge dieser Art hinaus.
[17] Vgl. auch Hausmann in diesem Band, der Kollokationen sowohl als Bineme versteht als auch als Triplestrukturen usw.
[18] In allen folgenden Beispielen sind die primären Kookkurrenzpartner fett gesetzt.

In der Umgebung von *Kopf – schüttelt* weisen also auch die Adjektive *ungläubig, verständnislos* und *fassungslos* eine auffällige Kohäsion auf. Die Relation *Kopf – Nägel* wird durch die Kookkurrenzpartner *gemacht* und *machen* weiter spezifiziert usw. Dieses Clusteringverfahren stellt eine mittels mathematisch-statistischer Berechnungen mögliche sukzessive Kontextspezifikation und Bedeutungsdisambiguierung dar. Die herausgefilterten Kookkurrenzcluster sind als eigenständige Entitäten zu verstehen, die selbst wiederum zum Analyse-Input für weitere Kookkurrenzanalysen werden (vgl. dazu auch Beispiele in 5). In vielen Fällen realisieren sich die wirkliche Bedeutung und der Gebrauch einer sprachlichen Einheit überhaupt erst in Mehrwortstrukturen und weniger auf der Einwortlexem-[19] bzw. der Binemebene.

Dass es sich bei statistisch erhobenen Kookkurrenzphänomenen um qualitativ völlig verschiedene Formen von Affinitäten handelt, wird schon an folgenden Beispielen mehr als deutlich:

schüttelt [verständnislos] [den] Kopf
Nägel [mit] Köpfen [gemacht]
[Menschen haben ihr] Dach [über dem] Kopf [verloren]
kühlen Kopf [bewahren]
[den] Nagel [auf den] Kopf getroffen

Hier gelangen wir an die Schnittstelle: Der Rechner hat alle auffälligen Strukturen, die er erfasst hat, analysiert, attribuiert und geordnet. Er bietet Häufigkeitsbewertungen und Präferenzen, aber keine Interpretation der beobachteten Phänomene. Es handelt sich noch um Resultate reiner Rechenprozesse, die jedoch die entscheidende Aussage bereits enthalten: Eine Entität A verhält sich zu einer Entität B_1, B_2, B_{3-n} auffällig kohäsiv. Wir erhalten Cluster, die sich durch häufigen Gebrauch in einer Sprachgemeinschaft als zusammenhängende Entitäten gebildet haben und so als wichtige Bausteine menschlicher Kommunikation fungieren.

Die Bewertung und Interpretation dieser Kohäsionsphänomene bleibt dem Linguisten vorbehalten. Und dieser hat natürlich Vorannahmen bzw. verfügt über Sprach- und Expertenkompetenz. Er betrachtet die Resultate – anders als der Rechner – nicht voraussetzungslos. Er hat ein Ordnungssystem im Kopf für die Interpretation der statistischen Ergebnisse, z. B. ein ontologisch-referenzielles oder ein semasiologisches. Diese Ordnungssysteme sind jedoch Setzungen, sie können sich von Mensch zu Mensch und von linguistischer Schule zu linguistischer Schule unterscheiden. Daher werden sie – und

[19] In Bezug auf die Ebene der lexikalischen Semantik ist Lexem (Wort) hier im Sinne von Bußmann zu verstehen als kleinster, relativ selbständiger Träger von Bedeutungen, die im Lexikon kodifiziert sind (2002, S. 750).

jetzt kommen wir auf die im Punkt 3 ausgeführte Argumentation zum Analyseparadigma zurück – auch nicht als prästrukturierende Annotationen in die Texte selbst integriert, sondern erst im Nachhinein – quasi als Folie – über die statistisch erhobenen Resultate gelegt (nachgelagerte Interpretation).

5. Linguistische Bewertung – Interpretieren, Ordnen, Verifizieren

Um die unterschiedliche Qualität der gefundenen Kookkurrenzcluster zu verdeutlichen, greifen wir auf tradierte Kategorien der Linguistik zurück, obwohl wir vermuten, dass sie nicht für alle Kookkurrenzphänomene anwendbar sein werden.

5.1 Das Interpretieren von Kookkurrenzclustern

Eine zentrale linguistische Qualifizierung bzw. Interpretation eines Clusters ist die **‚Kollokation'** im Sinne von Hausmanns Basis-Kollokator-Dichotomie. Kollokationsrelationen (z. B. *Kopf – schütteln, Kopf – hochrot, [mit dem] Kopf – nicken*) wären als eine – interpretierte – Teilmenge eines Kookkurrenzpotenzials zu verstehen, die im Bereich der Textproduktion vor allem für Fremdsprachenlerner von besonderer Relevanz ist. ‚Kookkurrenz' ist das statistisch erhobene Potenzial, ‚Kollokation' ein interpretiertes Teilphänomen.

Es gibt darüber hinaus Cluster, die sich nicht oder nicht ausschließlich als Kollokationen qualifizieren lassen, bei denen beispielweise kein direktes (semantisches) hierarchisches Determiniertheitsverhältnis zwischen ihren Elementen besteht. Sie enthalten jedoch relevante sprachliche und außersprachliche Informationen, über die sich der aktuelle Gebrauch sprachlicher Einheiten rekonstruieren lässt. So indizieren folgende Kookkurrenzcluster des Bezugswortes *Wasser* **semantische Aspekte** dieses Lexems, indem sie typische Umfelder des Gebrauchs verdeutlichen.

(5) Kookkurrenzcluster zu *Wasser* (semantische Aspektuierungen)

‚Energieträger/Ressource'

Wasser
- **Strom**
  - **Strom** Gas Abwasser Fernwärme
  - **Strom** Kanal Müll
  - **Strom** Wind
  - **Strom** Versorgung Wärme
  - **Strom** Heizung Leitungen
  - **Strom** Kubikmeter verbraucht Millionen
- **Wind**
  - **Wind** Sonne Biomasse

‚Element'

Wasser
- **Luft**
  - **Luft** Erde Elemente Licht
  - **Luft** Erde Elemente
  - **Luft** Erde Wärme
- **Erde**
  - **Erde** Wind Sonne
  - **Erde** Wind Elemente
  - **Erde** Wind Licht

‚Nahrungsgrundlage'

Wasser
- **Boden**
  - **Boden** Pflanzen aufnehmen
  - **Boden** Pflanzen
  - **Boden** aufnehmen

‚Kochen'

Wasser
- **kochendem**
  - **kochendem** 1/4 übergießen ziehen Minuten lassen voll
  - **kochendem** ziehen Minuten lassen überbrühen
  - **kochendem** Minuten ca gesalzenem

Es gibt des Weiteren Wortschatzbereiche, bei denen **thematisch-diskursive** Kookkurrenzcluster geradezu konstitutiv für den aktuellen Gebrauch sind, z. B. bei politischen Schlüsselwörtern wie *Globalisierung*.

(6) Kookkurrenzcluster zu *Globalisierung* (thematisch-diskursive Aspektuierungen)

Globalisierung
- **Zeitalter** ist Politik
- **Märkte** neue
- **Märkte** Öffnung
- **Zeiten** Kündigungswellen
- **Zeiten** auch gerade
- **Folgen** negativen
- **Folgen** sozialen kulturellen
- **Herausforderungen** begegnen
- **Internationalisierung** Telekommunikation

Liberalisierung Deregulierung Privatisierung
Weltwirtschaft
Thema Herausforderung Chance
Individualisierung Flexibilisierung Digitalisierung
Finanzmärkte
Europäisierung
Ostöffnung EU-Beitritt
Chance Falle
Chance Bedrohung
ist nicht aufzuhalten
Regionalisierung
Angst Wohlstand
Chancen Gefahren
Chancen Risiken
Rationalisierung Arbeitslosigkeit
Neoliberalismus
Fusionen
Vernetzung
Demokratie Kapitalismus
Verlierer Gewinner
schreitet
Strukturwandel

Von besonderer Relevanz sind **formelhafte** Kookkurrenzcluster mit mehr oder weniger kompositioneller und in der Regel nicht-figurativer Bedeutung. Zur Verdeutlichung hier die Kookkurrenzcluster für das Bezugswort ‚voll'.

(7) Kookkurrenzcluster zu *voll* (formelhaft)

voll

Hände

Hände tun noch haben
Hände tun noch derzeit
Hände tun haben derzeit jedenfalls

Lobes

Lobes sind waren auch
Lobes sind jedenfalls

Trend

Trend derzeit liegt

besetzten

besetzten Saal fast

Rechnung
Rechnung aufgegangen

Gange
Gange sind bereits noch

entbrannt
entbrannt bereits

ausschöpfen
ausschöpfen können Möglichkeiten

eingeschlagen
eingeschlagen haben

Geltung
Geltung kommen

funktionsfähig
funktionsfähig noch

Maß
Maß war jetzt

Interessant an diesem Beispiel ist, dass solche Wortverbindungen wie *im Trend liegen, Rechnung aufgehen, entbrannt sein, zur Geltung kommen, funktionsfähig sein* oder *Möglichkeiten ausschöpfen* zwar bekannte Muster sind, dass aber eine formelhafte Ausprägung augenscheinlich auch in besonderer Weise an das Adjektiv *voll* gebunden ist. Auf Grund ihrer (idiomatischen) Unauffälligkeit werden solche usuellen Verbindungen oft nicht erkannt. Mit Hilfe der statistischen Analyse rücken sie mehr als bisher ins Zentrum der Betrachtung von Wortverbindungen.

Die statistische Kookkurrenzanalyse bringt auch Cluster hervor, die auf usuelle **Teilidiome** und **Idiome** hindeuten. Man könnte zunächst meinen, dass sich Idiome auf Grund ihres idiosynkratischen Charakters und der in der Regel nicht auffällig hohen Vorkommenshäufigkeit einer statistischen Analyse eher verschließen – das Gegenteil ist der Fall.

(8) Kookkurrenzcluster zu *Haar* (idiomatisch)

Haar
gutes
gutes lassen kaum

herbeigezogen
herbeigezogen sind

Suppe
Suppe gefunden

Berge
Berge lassen stehen
Berge stehen

geraten
geraten sind

grauen
grauen lassen wachsen

Sumpf
Sumpf eigenen

raufen
gekrümmt

ein
ein wäre um
ein um

Durch die Möglichkeit der Einbeziehung von Funktionswörtern erfasst die COSMAS-Kookkurrenzanalyse auch solche Cluster wie *Haar – ein um wäre* (also *um ein Haar wäre*). Dadurch können wir syntaktische Bindungen erkennen, z. B. typische präpositionale Anschlüsse, eine der häufigsten Fehlerquellen für Fremdsprachenlerner.

Neben der Information, **dass** ein lexikalisches Kohäsionsphänomen vorliegt, besteht der analytische Wert dieser Cluster in ihrer Indikatorenfunktion. Man erhält Evidenzen, die nun in inhaltliche Zusammenhänge zu bringen sind.

5.2 Das Ordnen in Kookkurrenzfelder

Ein wichtiges heuristisches Instrument ist dabei die Erstellung von Kookkurrenzfeldern, die Ordnung von Clustern nach kategorialen Kriterien. So lassen sich Cluster und Komponenten der Cluster systematisieren nach: Wortartenfeldern, Bedeutungsfeldern, thematischen und domänenspezifischen Feldern oder auch nach Feldern, die typische Sprechereinstellungen indizieren oder auf feste, lexikalisierte (z. B. idiomatische) Wortverbindungen hinweisen. Wir greifen zunächst noch einmal das Beispiel *Wasser* auf, das zeigt, wie Kookkurrenzfelder neben den bereits im Wörterbuch kodifizierten Bedeutungen weitere semantische Aspekte hervorbringen, die aus dem aktuellen Gebrauch resultieren.

(9) *Wasser*

Im GWDS 1999 werden folgende Bedeutungen angegeben, die wir hier aus Platzgründen nur punktuell zitieren können:

1. a) [...] *(aus einer Wasserstoff-Sauerstoff-Verbindung bestehende) durchsichtige [...] Flüssigkeit [...]*
   b) [...] *ein Gewässer bildendes Wasser*
2. [...] *Gewässer* [...]
3. [...] *[alkoholische] wässrige Flüssigkeit:* wohlriechende, duftende Wässer; [...]
4. [...] a) *wässrige Flüssigkeit, die sich im Körper bildet* [...] b) *Schweiß* [...] c) *Urin* [...] d) *Tränenflüssigkeit* [...] (S. 4434/35)

Die Kookkurrenzanalyse zu *Wasser* deutet auf signifikante Verwendungsmuster von *Wasser* hin, die im Wörterbuch keine oder nur beiläufige Berücksichtigung finden können. In unseren Analyseresultaten gibt es natürlich auch Evidenzen für die Bedeutungen des GWDS, die wir jedoch in einer anderen Darstellungsform (nicht nummeriert, sondern in Feldern) zeigen.

Wortartenfelder: adjektivisch-adverbial

- kaltes, eiskaltes, gefrorenes, warmes, lauwarmes, heißes, kochendes, siedendes, frisches,
- verseuchtes, verschmutztes, verunreinigtes, sauberes, gereinigtes, abgekocht, belastetes, schweres
- seichtes, klares, glasklares, kristallklares, destilliertes, reines, trübes, flaches, tiefes, knietiefes, knöcheltief, weiches, salzhaltiges, brackiges, türkisblaues
- stilles, gebranntes
- genügend, reichlich
- fließendes, stehendes

Ausgewählte Bedeutungsfelder
(vgl. auch S. 99 f.)

‚Ressource/Energieversorgung'
Strom, Gas, Abwasser, Heizung, Fernwärme, Liter, Kubikmeter, Verbrauch, Energie, Versorgung, verbrauchen, Elektrizität, Sonne, Wind, Biomasse, sparen, sparsamer Umgang mit, Stromleitung

‚bedrohter Umweltfaktor'
sauber, rein, verschmutzt, verdreckt, Verschmutzung, Zugang zu sauberem, verseuchen, Lufthygiene, radioaktiv, Öl, trüb, sparsamer Umgang mit

‚Transportweg'
Schiffahrtsdirektion, Kanal, Boot

‚Natürliche Gewalt/Bedrohung'
unter Wasser gesetzt, Keller volllaufen, abpumpen, ertrinken, überschwemmen, eindringen, (von der Decke) tropfen

‚Lebensraum/Ort'
tauchen, schwimmen, fischen, Boot, Sprung, Fisch, springen

‚Lebensqualität'
fließend warmes, fließend kaltes, Kanalisation, Stromanschluss Wohnen am, Sonne, Wind, Meer, Wärme, freier Zugang zu, Licht, Quell

‚Natur'
Sonne, Wind, Sand, Fels, Ufer, Fisch, glasklar, Quelle, Pflanze, Welle

‚Element'
Luft, Feuer, Erde, Sonne

In anderen Fällen wie im folgenden Beispiel des Verbs *abverlangen* kommt man durch die Kookkurrenzcluster sogar zu einer adäquateren Bedeutungsbeschreibung, als sie beispielsweise im GWDS verzeichnet ist.

(10) *abverlangen*

Das GWDS gibt folgende Bedeutungsparaphrase an:

„**abverlangen** [...] *[mit Dreistigkeit] von jmdm. für eine Gegenleistung fordern, als Preis verlangen [...].*" (S. 120)

Die Kookkurrenzfelder ergeben jedoch eine etwas andere Gewichtung:

‚Abstriche/Verzicht'
Opfer, Zugeständnisse, Kompromisse, Unzumutbares, Leidensfähigkeit, Entbehrungen

‚besondere Leistung/besondere Eigenschaften'
- Konzentration, Höchstleistungen, Kondition, Geschick, Durchhaltevermögen, Virtuosität, Präzision, Kreativität, Sitzfleisch, Akrobatik
- Geduld, Respekt, Einsatz, Disziplin, Flexibilität, Aufmerksamkeit, Fingerspitzengefühl, Leistungsbereitschaft, Mut, Mobilität, Selbstbeherrschung, Selbstverleugnung

‚mit besonderer Anstrengung/besonderem Aufwand/bis an die Grenze gehend'

> große Opfer, schmerzhafte Kompromisse, viel Geduld, höchste Konzentration, ein hohes Maß an [...], Kraftakt, schier Unmögliches, eine erhebliche/gehörige Portion Mut, Übermenschliches, Äußerste, Extremes, vieles, alles

Die Kookkurrenzfelder machen deutlich, dass der Aspekt des Forderns bzw. Erbringens einer wie auch immer gearteten Leistung mit besonders hohem (u. U. bis an die Grenze gehenden) Einsatz von etwas (Kraft, Anstrengung, Energie usw.) den eigentlichen Aussagegehalt dieses Verbs ausmacht.

5.3 Das Verifizieren der statistischen Resultate für lexikografische Angaben

Kookkurrenzcluster und -felder stellen also wichtige heuristische Hilfsmittel dar, mit denen man Strukturen bzw. auffällige, typische und schließlich musterhafte usuelle Verwendungskontexte in sprachlichen Belegmengen erkennt. Dass diese Evidenzen zu einem wichtigen lexikografischen Hilfsmittel werden können, lässt sich an folgenden Aspekten der lexikografischen Beschreibung von Wortverbindungen zeigen (vgl. Burger 1998, Steyer 2000):

- Aktuelles Vorkommen von Wortverbindungen
- Semantischer Status
- Invarianz und Modifikation
- Externe Valenz
- Pragmatische Verwendungskontexte (Modalisierungen und Bewertungsmuster)

Aktuelles Vorkommen. Betrachtet man die einschlägigen Wörterbücher und phraseologischen Abhandlungen, so gibt es mittlerweile so etwas wie ein vererbtes kollektives Beispielgedächtnis von Forschern und Lexikografen, dessen Tradierung sich teilweise über Jahrzehnte zurückverfolgen lässt: Jemand sollte doch dahin gehen, *wo der Pfeffer wächst*, weil er *etwas auf dem Kerbholz hat*, sonst muss er zu schnell *den Löffel abgeben* oder er wird über denselbigen *balbiert*, aber er braucht deswegen nicht *die Flinte ins Korn zu werfen*, weil er weiß, *wo Barthel den Most holt*. Viele dieser Phraseologismen und Idiome bedürfen einer empirischen Überprüfung am aktuellen Sprachgebrauch.[20] Die COSMAS-Kookkurrenzanalyse erlaubt eine Verifizierung von Wortverbindungsphänomenen z. B. hinsichtlich ihrer Usualität oder auch in Bezug auf noch nicht kodifizierte bzw. sehr formelhafte Wortverbindungen. Einen besonderen Stellenwert nehmen Wortverbindungen ein, die zwar bereits usualisiert und in den Korpora repräsentiert sind, die Sprecher auch völlig adäquat verstehen können, die sie aber individuell –

[20] Einen wirklich großen Fortschritt stellt in dieser Hinsicht der Duden 11 (2002) dar.

nur mit Hilfe ihrer Sprachintuition – spontan nicht aktualisieren würden. Folgende in den IDS-Korpora signifikante Kookkurrenzcluster deuten z. B. auf solche aktuell üblichen (‚schwebenden') Wortverbindungen hin:

(11) Aktuelles Vorkommen

Kopf – Schere	[Schere im Kopf]
Auge – Dollarzeichen	[Dollarzeichen in den Augen]
frei – Kopf	[frei im Kopf sein]
voll – krass	[voll krass]
wirklich – nicht	[nicht wirklich]
Erde – Scheibe	[Die Erde ist eine Scheibe]
Hund – tot	[toter Hund]

Gerade Beispiele wie *Die Erde ist eine Scheibe*[21] und *toter Hund* können als Beleg für eine überindividuelle Repräsentation von sprachlichem Usus in Korpora gelten, ohne dass bereits eine direkte Rückwirkung auf die Sprachkompetenz jedes einzelnen Sprechers und damit auch jedes einzelnen Sprachwissenschaftlers erkennbar sein muss.

Semantischer Status. Neben dem Erkennen der Üblichkeit kann die Kookkurrenzanalyse Aufschluss bringen über den Status der Komponenten einer Wortverbindung. So wurde bei der auch in unserem Korpus signifikanten Kollokation *schütteres Haar* zumeist angenommen, dass das Adjektiv *schütter* nur in Bezug auf *Haar* erklärbar sei. Um dies zu verifizieren, haben wir eine Reziprokanalyse vorgenommen und neben *Haar* auch den Kookkurrenzpartner *schütter* einer Analyse unterzogen (*schütter* fungiert nun als Bezugswort). Es stellt sich heraus, dass das Adjektiv selbst ein ganzes Spektrum an eigenen Kookkurrenzpartnern aufweist.

(12a) Semantische Aspekte der Komponenten von *schütteres Haar*

Kookkurrenzpartner von *schütter*

- Haar, Haupthaar, Haarwuchs, Bart, Haarpracht, Haupt, Fell, Locken
- Zuschauerkulisse, Reihen, Kulisse, Applaus, Aufmarsch, Beifall, Wäldchen, Vegetation, Ränge, Häuflein, Grün, Kronen, Wald

Es liegt demnach eine eigenständige Bedeutung, nämlich „spärlich vorkom-

[21] Die Wendung *Die Erde ist eine Scheibe* wird sehr häufig genau in dieser satzwertigen Form verwendet und zumeist in eine koordinierende Struktur mit einer UND-Einleitung eingebettet. Etwas ist X und die Erde ist eine Scheibe. Sprecher drücken damit die Widersinnigkeit, Unglaubwürdigkeit oder das Anachronistische eines Sachverhalts aus. Mit *toter Hund* bezeichnen Sprecher den Sachverhalt, dass etwas „von gestern" ist und ihm keine Bedeutung mehr beigemessen wird.

mend" vor, und *schütter* ist in der lexikografischen Praxis (siehe GWDS) auch als eigenes Lemma anzusetzen.

Anders dagegen verhält es sich bei *sich die Haare raufen*: ‚angesichts eines aufgetretenen Problems oder Mißgeschicks entsetzt sein'. Auf Grund ihrer Bildlichkeit erscheint diese Wortverbindung transparenter, als sie es laut Korpus ist. So könnte man sich eine Bedeutung von *raufen* im Sinne von ‚zerwühlen', ‚ziehen' oder ‚zupfen' vorstellen, so wie es das GWDS in etwa angibt: „ [...][mhd. roufen, ahd. rouf(f)en, urspr. = (sich an den Haaren) reißen]: **1.** *(aus der Erde) herausziehen, [aus]rupfen [...]* (landsch.:)" Unsere Testanalysen haben für diesen Bedeutungsaspekt keine auffälligen Befunde erbracht. Viel dominanter sind die Kookkurrenzfelder, die auf folgende zwei Fälle hinweisen: die idiomatische Bindung an *Haar* (auch solche Kookkurrenzpartner wie *Mähne, Bart* werden nur in idiomatischer Bedeutung gebraucht) und die auch im GWDS angegebene zweite Bedeutung ‚miteinander kämpfen'.

(12b) Semantische Aspekte der Komponenten von *sich die Haare raufen*

Kookkurrenzpartner von *raufen*

- miteinander, saufen, Hochzeit, Hunde, wild, verzweifelt, beißen, streiten, grölen, betrunken, Männer, Buben, Fans, Schulhof, boxen, kämpfen, gerungen, toben, springen, rauchen, trinken
- EU-Fördergebiete, Pole-Position, Lehrstellen, Budget

Solche Reziprokanalysen können also auch dazu dienen, Bedeutungserweiterungen oder neue Lesarten von Kookkurrenzpartnern des jeweiligen Bezugslemmas zu erkennen.

Invarianz und Modifikation. Die Bestimmung des invarianten Kerns einer Wortverbindung und in Zusammenhang damit das Erkennen von Modifikationsanfälligkeiten oder Modifikationsresistenzen spielt in der Idiomatografie eine wichtige Rolle. Erkenntnisse zu invarianten Kernen dienen vor allem zum Bestimmen von Normalformen, die in der lexikografischen Praxis immer noch die formale Basis für das Artikelstichwort darstellen. Ein traditioneller Weg ist die Generierung von morpho-syntaktischen und lexikalisch-semantischen Restriktionen (vgl. dazu vor allem Fleischer 1997). Wir können mit unseren Methoden die Invarianz aus der Typik ableiten. Invarianz kann dabei auf verschiedenen Ebenen des Sprachsystems vorliegen, sie kann aber ebenso auf der höheren Abstraktionsebene kategorialer Festigkeit angesiedelt sein:

(13) Invarianz von *~hinter dem Ofen hervorlocken* (KWIC-Übersicht)

T92 eigentlich **keinen Hund hinter dem Ofen hervorlocken**.
U97 nehmers wohl **keinen Hund hinter dem Ofen hervorlocken**"
R97 werden damit **keinen Hund hinter dem Ofen hervorlocken**. Die

E00 würde **keinen müden Hund hinter dem Ofen hervorlocken**. Dazu braucht
U98 **auch keinen Menschen mehr hinter dem Ofen hervorlocken**. Auf den
B98 schließlich **niemanden mehr hinter dem Ofen hervorlocken**. Aus der
M99 **kein** [sic!] **Nachfolger mehr hinter dem Ofen hervorlocken**.
B98 eting **kaum noch jemanden hinter dem Ofen hervorlocken**. Der erneute
T92 ne kann so leicht **nichts hinter dem Ofen hervorlocken**. Und doch:
T99 USA **noch weniger Leute hinter dem Ofen hervorlocken** als eine
Z97 Variété!" **noch irgendwen hinter dem Ofen hervorlocken** zu können.
M00 gungen **locken nur wenige hinter dem Ofen hervor**. Die Formel „global

Invariant ist auf der lexikalischen Ebene nur *hinter dem Ofen* und das Verb *hervorlocken* (in seinen verschiedenen flektierten Formen). Das Lexem *Hund* stellt zwar das prototypische Basiselement dar, es ist aber austauschbar, z. B. durch ‚Mensch'. Die Negation ist gleichfalls ein kategoriales Basiselement, das allerdings nicht in allen Fällen durch die Quantitätsverneinung *kein* repräsentiert ist. Die invariante Form lautet demzufolge:

[NEG +(fak NOMEN)+ hinter dem Ofen hervorlocken] Gerade bei der Bestimmung invarianter Kerne sehen wir perspektivisch eine wichtige Schnittstelle zwischen kognitiven Modellen, z. B. im Sinne von Dobrovol'skij (1995 und in diesem Band), und korpusbasierten Zugängen.

Externe Valenz.[22] Die Bestimmung des invarianten Kerns hat aber nicht nur Auswirkungen auf die Bestimmung der internen Struktur der konventionalisierten Normalform bzw. Nennform, sondern auch auf die externen Valenzangaben (vgl. dazu auch Keil 1997, S. 63 ff.). Der invariante Kern wird zur Entität, die durch weitere Ergänzungen spezifiziert wird. Bisher wurde dies vor allem durch den Einsatz solcher Phrasenmusterangaben wie [*jemand* tut *etwas*] realisiert. Durch die statistische Analyse der Umgebungen eines Kookkurrenzclusters ist eine Indizierung und eine sehr umfassende Auffüllung mit den – nach einem Signifikanzkriterium ermittelten – aktuell gebräuchlichen Partnern möglich; diese Auswahl bleibt dann nicht mehr dem individuellen Sprachgefühl des einzelnen Sprechers allein überlassen.

(14a) Externe Valenz von [*sich*] *über Wasser halten*

Nominale Kookkurrenzpartner

[mit]
- Gelegenheitsjobs, Gaunereien, Diebstählen, Prostitution, Einbrüchen, Betteln
- Finanzspritzen, Matcheinnahmen, Krediten, Subventionen, Geschäften

(14b) Externe Valenz von *an den Haaren herbeigezogen*

[22] vgl. Burger 1998, S. 176 f.

Nominale Kookkurrenzpartner

- Vorwurf, Begründung, Handlung, Geschichte, Kritik, Argument, Vergleich

Die Verallgemeinerungen der zu Grunde liegenden Muster lassen sich damit konkreter und spezifischer fassen, als das mit der bisher üblichen abstrahierten Form möglich ist.

Pragmatische Verwendungskontexte. Bei vielen Kookkurrenzpartnern von Kookkurrenzclustern handelt es sich nicht allein um die Ausfüllung externer Valenzen, vielmehr indizieren sie Domänengebundenheiten, typische Sprechereinstellungen und -bewertungen oder typische Kontextualisierungen im Diskurs und können somit für komplexe Textinterpretationen genutzt werden. So findet man Kookkurrenzpartner, die Metakommentare sind, die Mündlichkeit indizieren oder Distanzierungssignale darstellen wie z. B. *so genannt.* Zwei Beispiele sollen zur Illustration solcher pragmatischer Verwendungskontexte genügen:

(15) Modalisierungen von [*sich*] *über Wasser halten*

Kookkurrenzcluster
kaum noch, gerade noch, nicht mehr, können, mühsam, mehr recht als schlecht, notdürftig, halbwegs, irgendwie, einigermaßen

Über die statistische erhobenen Kontexte der Wortverbindungen lassen sich auch diskursspezifische Sprechereinstellungen identifizieren, wie das Beispiel der Kookkurrenzen zu *harter Hund* zeigt:

(16) Bewertungsmuster zu *harter Hund*

Verbale Kookkurrenzpartner
gelten (*als harter Hund gelten*), verschrieen (*als harter Hund verschrieen sein*), raushängen (*den harten Hund raushängen lassen)*, haben Ruf (*den Ruf als harter Hund haben*)

Adjektivische Kookkurrenzpartner
wild, eisern, unerbittlich, konsequent

Nominale „synonymische" Kookkurrenzpartner
Schleifer, Polterer, Wahrheitsfanatiker, Heißmacher, Schlitzohr, Profi, Quälix

Nominale „antonymische" Kookkurrenzpartner
Softietyp, Weichspüler, Kumpeltyp

Häufung von adversativen Konstruktionen

Er ist ein **harter** Hund, aber niemals unfair.
Ich habe gehört, er ist ein **harter** Hund, der aber immer Erfolg hat.
Der Trainer selbst ist zwar ein **harter** Hund, aber rücksichtslos loyal.

6. Anwendung und Perspektiven

Die Methoden der Korpuserschließung, insbesondere der statistischen Kookkurrenzanalyse, werden derzeit vor allem im *elexiko*-Modul ‚Usuelle Wortverbindungen' systematisch angewendet. Im Projekt werden signifikante Kookkurrenzcluster des Deutschen, so wie sie in den IDS-Korpora vor allem im hochfrequenten Bereich repräsentiert sind, mit mathematisch-statistischen Methoden schrittweise herausgefiltert, geordnet, systematisiert, interpretiert und schließlich lexikografisch in Form von Kookkurrenzangaben für die *elexiko*-Lemmastrecke aufbereitet. Kookkurrenzangaben enthalten nach formalen und linguistischen Kriterien in Gruppen geordnete und markierte (multiple) Clusterangaben einschließlich ausführlicher Korpusbelege. Durch diese Systematisierungen und durch die darauf aufbauende – wiederum automatische – Kookkurrenzanalyse der Kookkurrenzcluster wird eine immer tiefer gehende Kontextspezifizierung und damit ein Erkennen usueller Bedeutungen und Verwendungskontexte mittels statistischer Evidenzen erreicht. In den Kookkurrenzangaben sind verschiedene metasprachliche Kommentare vorgesehen, so z. B. eine explizite Markierung des Kernbereichs der Kollokationen. Die Kennzeichnung, ob es sich bei einem Cluster um eine Kollokation im engen Sinne handelt oder nicht und – wenn ja – welche Komponente Basis, welche Kollokator ist, soll in zweifacher Hinsicht von Nutzen sein: Durch die expliziten Markierungen kann dieser Teilbereich der Kookkurrenzen später gesondert automatisch erstellt und online abrufbar gemacht werden, um ihn vor allem den nichtmuttersprachlichen Lernern und dem gesamten Bereich ‚Deutsch als Fremdsprache' zur Verfügung zu stellen. Gleichzeitig stellt dieser Teilbereich eine in quantitativer Hinsicht sehr ertragreiche empirische Basis für die Kollokationsforschung dar. In diesen Artikeln sind des weiteren sowohl pragmasemantische Gebrauchskommentare vorgesehen als auch Angaben zu Invarianz, typischen Verwendungsmustern, externen Valenzen, zum Status der Basiselemente usw. Ziel ist es, einen Mehrwortstandard des Deutschen auf der Basis sehr großer elektronischer Korpora zu erarbeiten, auf den nicht nur sprachinteressierte Laien, sondern auch Fachkollegen, vor allem in der Auslandsgermanistik, zurückgreifen können. Diese Arbeiten sind eng verzahnt mit den Vorhaben im Bereich der Korpustechnologie und der Weiterentwicklung der Korpusanalysemethoden am IDS, speziell mit der Weiterentwicklung der statistischen Kookkurrenzanalyse. Eine wichtige gemeinsame Analyse-, Experimentier- und Evaluationsplattform stellt die (nur hausintern zugäng-

liche) COSMAS-Kookkurrenzdatenbank (CCDB)[23] dar, die das rein statistisch erhobene Kookkurrenzpotenzial von Lemmata nach vordefinierten Parametern auf einer fixen Korpusbasis zu einem bestimmten Zeitpunkt abbildet und – je größer sie wird – einen immer umfassenderen Einblick in die Kohäsionsphänomene des deutschen Wortschatzes und damit auch in seine internen Vernetzungen bietet. Der zunehmenden Komplexität der erfassten Phänomene versuchen wir durch eine ständige und wechselseitige Reflexion der Methoden und ihrer Interpretationsspielräume gerecht zu werden.

Vor allem in Hinblick auf eine angemessene linguistische Interpretation der beobachteten statistischen Phänomene gibt es derzeit noch sehr viele ungelöste Fragen, z. B.:

- Wie ist das zum Teil divergierende Kookkurrenzverhalten in Bezug auf unterschiedliche grammatische Formen eines Lemmas (beispielsweise Singular vs. Plural oder die Dominanz der Partizipform innerhalb eines Flexionsparadigmas) zu interpretieren?
- Wie geht man mit ganz offensichtlichen Verdrängungsmechanismen um, also zum Beispiel mit frequenten und unspezifischen Vorkommen von Numeralia, Eigennamen oder Maßeinheiten? Inwieweit kann man durch eingrenzende oder ausschließende Suchanfragen diese Verdrängungen vermeiden, ohne dass man wiederum „prästrukturierend" wirkt?
- Wie ist der statistisch unspezifische Bereich zu bewerten?[24]
- Kann man mit der Erweiterung des Analysefokus (der Kookkurrenzspanne) koreferenziellen Mechanismen oder gar textuell-diskursiven Kohärenzen und Isotopien auf die Spur kommen?
- Welche Zusammenhänge gibt es zwischen Kookkurrenzen und Wortbildungsphänomenen, speziell Komposita (eine typische Frage an das Deutsche)?

Wir sind dabei mit den ‚Mühen der empirischen Ebene' und mit einer außerordentlichen Heterogenität der Befunde konfrontiert. Dies verstehen wir jedoch als Herausforderung und als Chance, sprachlichem Usus näher zu kommen. Es gibt bisher sehr wenig linguistische Erfahrung im Umgang mit Korpora solcher Quantität. Deshalb befinden wir uns derzeit noch in einer

[23] Informationen zur CCDB unter http://www.ids-mannheim.de/kt/projekte/methoden/ka.html/#CCDB

[24] Bisher haben wir nur Hypothesen in Bezug auf den statistisch unspezifischen Bereich. Diese Hypothesen bedürfen einer systematischen empirischen Überprüfung. Wir können nicht in jedem Fall beurteilen, ob es sich wirklich um rein Okkasionelles handelt oder nicht auch um Typisches, das nur noch nicht erfasst wurde. Die Gründe können z. B. in der Verdrängung durch statistisch stärkere sprachliche Phänomene liegen oder darin, dass Vertreter derselben Klasse statistisch unterschiedlich auffällig werden und damit statistisch unterschiedlich ins Gewicht fallen.

experimentellen Phase, einer Phase des reflektierten Sammelns, Systematisierens und Kategorisierens. Wir versuchen erst noch, in das Universum der Kookkurrenz und – ganz allgemein – der inneren sprachlichen Korpuszusammenhänge einzudringen und die Strukturen zu entschlüsseln. Und wir können derzeit nur ahnen, dass wir nach einer Phase der umfassenden Analyse von sprachlichen Massendaten vor ganz neuen – möglicherweise auch theoretisch neuen – Herausforderungen stehen werden. Ein Forschungsziel unseres kooperativen Vorgehens ist es daher auch, Einblicke in Kohäsionsphänomene zu gewinnen, die über die rein lexikalischen Affinitäten hinausgehen und auf einer abstrakteren, kategorialen Ebene liegen, die damit interne Gesetzmäßigkeiten erhellen und uns einen Zugang zur kognitiven Ebene unserer Sprachverarbeitung ermöglichen. Das zwingt uns auch, einen neuen Blick auf tradierte Modelle und Ansätze zu werfen. Natürliche Sprache konstituiert sich nicht in klar strukturierten Rastern, etwa auf der rein syntaktischen, lexikalischen oder semantischen Ebene des Sprachsystems. Das, was eine sprachliche Einheit im tatsächlichen Gebrauch ausmacht, was sie im Vergleich zu anderen als etwas Besonderes erscheinen lässt, was ihre Funktion in der Kommunikation bestimmt, unterscheidet sich – betrachtet man die Korpusbefunde – oft von Entität zu Entität. Die Kookkurrenzanalyse bringt uns auf die Spur dieser im Korpus verankerten distinktiven Gebrauchseigenschaften sprachlicher Einheiten.

Es sollte deutlich geworden sein, dass die Kookkurrenzanalyse es ermöglicht, rekurrente Muster zu erkennen und Fragen zu formulieren bzw. zu beantworten, die zum Teil erst durch die Analyse hervorgetreten sind. John Sinclair beschreibt dies sinngemäß so: Eine intensive Auseinandersetzung mit Korpustexten löst nicht automatisch die Probleme ihrer Beschreibung; aber es wird dadurch viel klarer, welche Probleme überhaupt zu lösen sind. Heute müssen wir die Evidenzen noch mit Vorsicht nutzen, aber nutzen müssen wir sie. Denn Sprache sieht deutlich anders aus, wenn man viel von ihr auf einmal betrachtet (1991, u. a. S. 100).

Literatur

„Arbeitsgruppe für Korpustechnologie“ (2003): Homepage unter http://www.ids-mannheim.de/kt

Armstrong, Susan (Hg.) (1994): Using Large Corpora. Cambridge/Massachusetts/London.

Aarts, Jan/Belica, Cyril/Cloeren, Jan/Gross, Maurice/Moulin, André/Neumann, Robert/Sinclair, John/van Sterkenburg, P. G. J. (1993): MECOLB Project Proposal. MLAP Call 1993: Exploratory Actions for the Language Industry. Feasibility and Validation Study. Luxembourg.

Bahns, Jens (1996): Kollokationen als lexikographisches Problem. Eine Analyse allgemeiner und spezieller Lernerwörterbücher des Englischen. (= Lexicographica 74). Tübingen.

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

Belica, Cyril (1997): Korpuslinguistik als Arbeitsfeld der LDV: Korpora und ihre Methoden. In: Unterlagen zur Evaluation der Zentralen Arbeitsstelle Linguistische Datenverarbeitung am Montag, 10. März 1997. IDS. (unveröffentlicht).

Belica, Cyril (1998): Statistische Analyse von Zeitstrukturen in Korpora. In: Teubert, Wolfgang (Hg.): Neologie und Korpus. (= Studien zur deutschen Sprache 11). Tübingen. S. 31–42. (übersetzte u. überarb. Fassung von Belica 1996).

Boguraev, Branimir/Pustejovsky, James (Hg.) (1996): Corpus Processing for Lexical Acquisition. Cambridge/Massachusetts.

Burger, Harald (1998): Phraseologie. Eine Einführung am Beispiel des Deutschen: (= Grundlagen der Germanistik 36). Berlin.

Bußmann, Hadumod (2002): Lexikon der Sprachwissenschaft. Dritte, aktualisierte und erweiterte Aufl. Stuttgart.

Cowie, A. P. (1998): Phraseology. Theory, Analysis, and Applications. Oxford.

Dobrovol'skij, Dmitrij (1993): Datenbank deutscher Idiome. Aufbauprinzipien und Einsatzmöglichkeiten. In: Földes, Czaba (Hg.): Germanistik und Deutschlehrerausbildung. Szeged/Wien. S. 51–67.

Dobrovol'skij, Dmitrij (1995): Kognitive Aspekte der Idiom-Semantik. Studien zum Thesaurus deutscher Idiome. (= Eurogermanistik 8). Tübingen.

Dobrovol'skij, Dmitrij (2004): Idiome aus kognitiver Sicht. In: Steyer, Kathrin (Hg.): Den Nagel auf den Kopf treffen. Wortverbindungen mehr oder weniger fest. (= Jahrbuch des Instituts für Deutsche Sprache 2003). Berlin/New York. S. 117–143.

Dodd, Bill (Hg.) (2000): Working with German corpora. Birmingham.

Dunning, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational Linguistics, Vol 19, 1.

Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt a. M.

Feilke, Helmuth (1998): Idiomatische Prägung. In: Barz, Irmhild/Öhlschläger, Günther (Hg.): Zwischen Grammatik und Lexikon. (= Linguistische Arbeiten 390). Tübingen. S. 69–80.

Fleischer, Wolfgang (1997): Phraseologie der deutschen Gegenwartssprache. 2., durchges. u. erg. Aufl., Tübingen.

Gülich, Elisabeth/Krafft, Ulrich (1998): Zur Rolle des Vorgeformten in Textproduktionsprozessen. In: Wirrer, Jan (Hg.): Phraseologismen in Text und Kontext. (= Phrasemata I). Bielefeld. S. 11–38.

Hausmann, Franz Josef (1984): Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: Praxis des neusprachlichen Unterrichts. Jg. 31. S. 395–406.

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, Henning/Mugdan, Joachim (Hg.): Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.–30.6.1984. (= Lexicographica 3). Tübingen. S. 118–129.

Hausmann, Franz Josef (1993): Was ist und was soll ein Kontextwörterbuch? (Einführung). In: Ilgenfritz, Peter/Stephan-Gabinel, Nicole/Schneider, Gertrud: Langenscheidts Kontextwörterbuch Französisch-Deutsch. Ein neues Wörterbuch zum Schreiben, Lernen, Formulieren. Berlin/München/Wien/Zürich/New York. S. 5–9.

Hausmann, Franz Josef (2004): Was sind eigentlich Kollokationen? In: Steyer, Kathrin (Hg.): Den Nagel auf den Kopf treffen. Wortverbindungen mehr oder weniger fest. (= Jahrbuch des Instituts für Deutsche Sprache 2003). Berlin/New York. S. 309–334.

Heringer, Hans Jürgen (1999): Das höchste der Gefühle. Empirische Studien zur distributiven Semantik. Tübingen.

Keil, Martina (1997): Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex). (= Sprache und Information 35). Tübingen.

Kilgariff, Adam (1997): „I don't believe in word senses." http://www.itri.bton.ac.uk/~Adam.Kilgarriff/publications.html#1997 (auch erschienen in: Computers and the Humanities 31 (2), S. 91–113.

Lehr, Andrea (1996): Kollokationen und maschinenlesbare Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze. (= Reihe Germanistische Linguistik 168). Tübingen.

Lemnitzer, Lothar (1997): Akquisition komplexer Lexeme aus Textkorpora. (= Reihe Germanistische Linguistik 180). Tübingen.

Schmidt, Hartmut (1989): Sprachgeschichte zwischen Wort und Text. Über die Notwendigkeit einer historischen Wortkombinationsforschung. In: Sprachwissenschaft in der DDR – Oktober 1989. Vorträge einer Tagung des Zentralinstituts für Sprachwissenschaft am 31.10. und 1.11 1989. (= Linguistische Studien, Reihe A, Arbeitsberichte 208). Berlin. S. 170–186.

Schmidt, Hartmut (1995): Wörter im Kontakt. Plädoyer für historische Kollokationsuntersuchungen. In: Gardt, Andreas/Mattheier, Klaus J./Reichmann, Oskar (Hg.): Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien. (= Reihe Germanistische Linguistik 156). Tübingen. S. 127–143.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

Stein, Stephan (1995): Formelhafte Sprache. Untersuchungen zu ihren pragmatischen und kognitiven Funktionen im gegenwärtigen Deutsch. (= Sprache in der Gesellschaft. Beiträge zur Sprachwissenschaft 22). Frankfurt a. M./Berlin/Bern/New York/Paris/Wien.

Steyer, Kathrin (1998): Kollokationen als zentrales Übersetzungsproblem – Vorschläge für eine Kollokationsdatenbank Deutsch-Französisch/Französisch-Deutsch auf der Basis paralleler und vergleichbarer Korpora. In: Bresson, Daniel (Hg.): Lexikologie und Lexikographie Deutsch-Französisch. (= Cahiers d'Études Germaniques 35). Aix-en-Provence. S. 95–113.

Steyer, Kathrin (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 28, S. 101–125.

Steyer, Kathrin (2002): Wenn der Schwanz mit dem Hund wedelt. Zum linguistischen Erklärungspotenzial der korpusbasierten Kookkurrenzanalyse. In: Haß-Zumkehr, Ulrike/Kallmeyer, Werner/Zifonun, Gisela (Hg.): Ansichten zur deutschen Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag. (= Studien zur deutschen Sprache 25). Tübingen. S. 215–236.

Steyer, Kathrin (2003a): Korpus, Statistik, Kookkurrenz. Lässt sich Idiomatisches „berechnen"? In: Burger, Harald/Häcki, Buhofer, Annelies/Gréciano, Gertrud (Hg.): Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie. (= Phraseologie und Parömiologie 14). Baltmannsweiler. S. 33–46.

Steyer, Kathrin (2003b): Idiomatik hypermedial. Zur Repräsentation von Wortverbindungen im Informationssystem „Wissen über Wörter". In: Palm Meister, Christine (Hg.): EUROPHRAS 2000. Akten der Internationalen Tagung zur Phraseologie 15.-18. Juni 2000 in Aske, Schweden. Tübingen. (im Druck).

Stubbs, Michael (1997): ‚Eine Sprache idiomatisch sprechen'. Computer, Korpora, Kommunikative Kompetenz und Kultur. In: Mattheier, Klaus J. (Hg.): Norm und Variation. (= forum Angewandte Linguistik 32). Frankfurt a. M./Berlin/Bern/New York/Paris/Wien. S. 151–167.

Teubert, Wolfgang (1999): Korpuslinguistik und Lexikographie. In: Deutsche Sprache 27, S. 292–313.

Viehweger, Dieter (1989): Probleme der Beschreibung semantischer Vereinbarkeitsrelationen im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Franz Josef/ Reichmann, Oskar/Wiegand, Herbert Ernst/Zgusta, Ladislav (Hg.): Wörterbücher. Ein internationales Handbuch zur Lexikographie. (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.1). Berlin/New York. S. 888–893.

Wörterbücher

Agricola, Erhard (Hg.) (1992): Wörter und Wendungen. Wörterbuch zum deutschen Sprachgebrauch. Unter Mitwirkung v. Herbert Görner und Ruth Küfner. Überarb. Neufassung der 14. Aufl., 1. Aufl. der Neufassung. Mannheim/Leipzig/Wien/Zürich.

Benson, Morton/Benson, Evelyn/Ilson, Robert (1997): The BBI Dictionary of English Word Combinations. Revised edition. Amsterdam/Philadelphia.

DUDEN 2 (2001): Das Stilwörterbuch. 8., völlig neu bearb. Aufl. Herausgegeben von der Dudenredaktion. Mannheim/Leipzig/Wien/Zürich.

DUDEN 11 (2002): Redewendungen. 2., neu bearb. und aktualisierte Aufl. Herausgegeben von der Dudenredaktion. Mannheim/Leipzig/Wien/Zürich.

DUDEN 12 (2002): Zitate und Aussprüche. 2., neu bearb. und aktualisierte Aufl. Herausgegeben von der Dudenredaktion. Mannheim/Leipzig/Wien/Zürich.

GWDS (1999): Das große Wörterbuch der deutschen Sprache in zehn Bänden. 3., völlig neu bearb. und erw. Aufl. Herausgegeben von der Dudenredaktion. Mannheim/Leipzig/Wien/Zürich.

Oxford Collocations Dictionary for students of English (2002): Oxford.

Röhrich, Lutz (1991): Das große Lexikon der sprichwörtlichen Redensarten. Freiburg/Basel/Wien.

Schemann, Hans (1993): PONS Deutsche Idiomatik. Die deutschen Redewendungen im Kontext. Stuttgart/Dresden.

Sinclair, John (Hg.) (1987): Collins COBUILD English Language Dictionary. London/ Glasgow (Collins). Stuttgart (Klett).