Erschienen in:
Marie Vachkovà (Hg.): Beiträge zur bilingualen Lexikographie. Praha: Univ. Karlova, Filozifická Fak, S. 7–24.

KORPUSANALYTISCHE ZUGÄNGE ZU SPRACHLICHEM USUS

Cyril Belica – Kathrin Steyer

Vorbemerkung

Dieser Beitrag soll Prämissen und Prinzipien des Konzepts darstellen, das der korpuslinguistischen Forschung am Institut für Deutsche Sprache in Mannheim zu Grunde liegt (vgl. BELICA/NEUMANN 1997). Uns geht es vor allem darum, deutlich zu machen, dass sich die Korpusarbeit am IDS nicht auf die – unbestritten wichtigen – Arbeiten wie Textakquisition, Pflege und Bereitstellung von Korpora in großer Dimension beschränkt, sondern dass die Entwicklung von Korpusanalysemethoden und die Erforschung ihres Erklärungspotenzials sowohl für die theoretische Linguistik als auch in angewandten Gebieten wie der Fremdsprachendidaktik und Lexikografie im Mittelpunkt unserer Arbeit steht. Unabdingbare Voraussetzungen sind jedoch gut aufbereitete Korpora und Analysemethoden mit einem hohen Qualitätsstandard, die die Differenziertheit und Komplexität der Sprache nicht simplifizieren oder reduzieren.

Folgende Aspekte stehen im Zentrum unseres Konzepts:
- sehr große Korpora als aussagekräftige Stichprobe der Sprache,
- keine A-Priori-Modellierung der Sprache, sondern nachgelagerte Interpretation,
- Analyseparadigma der Korpusbasiertheit,
- Kookkurrenzanalyse als wichtigste Methode zur Strukturierung von sprachlichen Massendaten,
- linguistische Interpretation und lexikografische Anwendung aus kontrastiver Perspektive.

Der Beitrag von Kathrin Steyer und Marie Vachková in diesem Band baut unmittelbar auf unseren Ausführungen auf und wird speziell auf den letzten Aspekt eingehen. Er illustriert, welches empirische Potenzial die hier diskutierten Korpusanalysemethoden, speziell die statistische Kookkurrenzanalyse, für den Bereich Deutsch als Fremdsprache haben. Gerade auf dem Gebiet der Fremdsprachendidaktik und der bi- und multilingualen Linguistik und Lexikografie sieht man besonders deutlich den Nutzen der am IDS entwickelten korpuslinguistischen Methoden und Modelle in anwendungsbezogener Hinsicht.

1. Korpus befragen vs. Korpus analysieren

Die methodischen Herausforderungen an eine am aktuellen Sprachgebrauch orientierte Lexikografie – und demzufolge an eine angewandte Linguistik – resultieren

vor allem aus dem Spannungsfeld zwischen gemeinsamen „Sprachcodes" in einer Sprachgemeinschaft, die unabdingbar sind, damit sich die Sprecher überhaupt verständigen können, und individuell ausgeprägtem Sprachverhalten, das sich von einer Sprechergruppe zur anderen fundamental unterscheiden kann und unterscheidet. Dabei wird u. E. die Fähigkeit des einzelnen – auch des professionellen und auch des muttersprachlichen – Sprechers, Urteile über den wirklich typischen Gebrauch einer Sprachgemeinschaft abzugeben, immer noch überschätzt. Es ist für die einzelnen Sprachbenutzer zwar möglich, in begrenztem Maße zu erkennen, was „falsch" und „richtig", üblich oder unüblich ist, aber ein gesichertes Urteil über den Sprachgebrauch können sie eigentlich kaum fällen: Sie sind – um nur einige Beispiele zu nennen – nicht gleichzeitig in allen Dialektgebieten zu Hause, beherrschen nicht alle Fach- oder Gruppensprachen, haben unterschiedliche Sprachbiografien oder unterschiedliche Bewertungen des Fremdwortgebrauchs in der Muttersprache. Lexikografen stehen nun naturgemäß vor dem Problem, individuelle sprachliche Besonderheiten von wirklich Üblichem unterscheiden zu müssen, was in besonderer Weise für die kontrastive Lexikografie zutrifft. In Wörterbüchern und Lehrwerken soll in der Regel der **Sprachstandard** abgebildet werden. Viele Jahrhunderte wurden deshalb authentische sprachliche Belege gesammelt und in Zettelkästen systematisch dokumentiert. Diese „Zitate des Sprachgebrauchs" haben geholfen, den Blick für das Wesentliche eines sprachlichen Phänomens zu schärfen. Trotz dieser Hilfsmittel finden sich – wie vielfach beklagt – in Wörterbüchern und Lehrwerken aber nach wie vor viele nicht mehr dem aktuellen Gebrauch entsprechende Beispiele und Angaben, die teils über Generationen hinweg tradiert und „weitervererbt" wurden.

Durch die Entwicklung der Computertechnologie und die damit einhergehenden Möglichkeiten, große elektronische Textdatenbanken in Form von Korpora aufzubauen und zu nutzen, treten diese Desiderata besonders deutlich hervor: Es eröffnet sich ein Blick auf eine sehr große Menge an Sprachdaten, was eine völlig neue Herausforderung für die Linguistik bedeutet. Es bietet sich nunmehr die Chance, anhand sehr großer Sammlungen natürlichsprachiger Texte auch solche sprachliche Verwendungsmuster und Strukturen zu erkennen, die sich bisher dem Blick der Sprachteilhaber und auch der Linguisten oft noch entzogen haben. Es bietet sich die Chance, sprachlichen Usus in einer neuen Dimension zu erfassen und zu beschreiben. Gleichwohl wird der Traum weiterhin bestehen bleiben, das gesamte System Sprache in seiner Komplexität und Vielfalt vollkommen erfassbar zu machen. Auch Textkorpora – und seien sie noch so groß – stellen letztlich immer nur einen Ausschnitt, eine Stichprobe aus der Sprache dar. Die auf dieser Basis zu treffenden Aussagen sind folgerichtig nur für das zu Grunde liegende Korpus gültig. Es handelt sich demzufolge auch um eine **korpusbezogene Usualität** sprachlicher Phänomene. Die Herausforderung besteht nun darin, herauszufinden, inwieweit es möglich ist, die Erkenntnisse, die man durch die Analyse eines Sprachausschnitts gewon-

nen hat, auf andere Ausschnitte zu extrapolieren. Je größer und vielfältiger die Korpora sind – wir sprechen von *sehr großen Korpora* (very large corpora, vgl. Proceedings 1993) – desto mehr können sie diese verallgemeinernde Funktion erfüllen. Deshalb verfolgt das Institut für Deutsche Sprache in Mannheim seit vielen Jahren das Ziel, eine größtmögliche Quantität in Bezug auf die akquirierten Texte und eine größtmögliche Variabilität in Bezug auf den darin dokumentierten Sprachgebrauch zu erreichen, wie schon Mercer programmatisch formulierte: „More data is better data" (Church – Mercer 1993, 18). Gegenwärtig verfügt das IDS mit etwa zwei Milliarden Textwörtern über die weltweit größte elektronische Sammlung deutschsprachiger Texte für die linguistische Forschung[1]. Die IDS-Korpora geschriebener Sprache stellen damit den derzeit größten Ausschnitt der deutschen Sprache überhaupt dar. Sie enthalten belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten. Sie erlauben die benutzerseitige Komposition beliebiger virtueller Korpora (vgl. Mecolb 1993), die für verschiedenartige Aufgabenstellungen jeweils „repräsentativ" zusammengesetzt sind. Dieser Ansatz, in dem die Frage der Repräsentativität eines Korpus nicht in der Phase der Korpusakquisition, sondern in der Phase der Korpusnutzung behandelt wird, gehört seit Jahren zu den grundlegenden Prinzipien des IDS-Korpuskonzepts[2]. Eine weitere Prämisse ist die konsequente Unterscheidung zwischen der im Korpus aufgezeichneten empirischen „Beobachtung", textexterner Annotation und nachgelagerter Interpretation (z.B. theoriegebundene Modellierung, linguistische Annotation usw.). Das bedeutet, dass wir u.a. zwischen dem Korpus als ultimativer und „unfehlbarer" Quelle für sprachliches Vorkommen einerseits und den subjektiven Interpretationen („Meinungen") über Eigenschaften und Zusammenhänge sprachlicher Vorkommen andererseits strikt trennen. Solche Interpretationen sind nicht das Korpus selbst, auch wenn sie aus praktischen Gründen zusammen mit dem Korpus technisch aufbewahrt werden. Bemühen wir ein Bild aus der Biologie: Das Korpus ist wie eine Sammlung von Schmetterlingen. Man hat sie irgendwo auf der Wiese als Stichprobe aus der Natur entdeckt und eingefangen, holt sie sich ins Laboratorium, um sie dann zu erforschen. Die Ergebnisse dieser Forschungsarbeit, die Bestimmung der Arten und Klassen, ihrer Merkmale und distinktiven Eigenschaften bildet man nach der Analyse der einzelnen Vorkommen in Kategorien und Modellen ab. Die Schmetterlinge selbst bleiben davon unberührt. Während die Stichprobe selbst objektiv ist – den Schmetterling gibt es wirklich – sind die Meinungen darüber, zu welcher Gruppe sie gehören, abhängig von

---

[1] Die Korpora sind über das IDS-eigene Recherchetool COSMAS II zugänglich, siehe http://www.ids-mannheim.de/cosmas2.

[2] Zum Konzept und zu den Aufgaben vgl. auch Perkuhn, Rainer – Belica, Cyril – al Wadi, Doris – Lauer, Meike – Steyer, Kathrin – Weiss, Christian (2005).

der jeweiligen Theorie und können demzufolge auch kontrovers sein. In diesem Sinne ist das Korpus objektiv und „unfehlbar“[3], da es real existierende sprachliche Erscheinungen – als einen Teil der Sprachwirklichkeit – festhält. Morphosyntaktische oder auch semantische Korpusannotationen sind dagegen genau solche subjektiven und damit auch anfechtbaren kategorialen Zuordnungen zu real existierenden sprachlichen Zusammenhängen in Abhängigkeit vom zu Grunde gelegten linguistischen Modell. Verwechselt man gedanklich und methodisch derartige Annotationen mit dem Korpus selbst, so erhält man nicht mehr den „unmittelbaren Blick“ auf die Sprache, sondern Resultate, die nur vor der Folie des zu Grunde liegenden Modells interpretierbar sind.

Die Korpusbasiertheit als empirisches Prinzip ist kaum mehr umstritten. Mit der wachsenden Akzeptanz ist aber auch eine zunehmende Heterogenität der Auffassungen darüber verbunden, was unter diesem Prinzip eigentlich zu verstehen ist. In den meisten Fällen konsultieren Linguisten und Lexikografen ein solches Korpus nach wie vor, um zuvor aufgestellte Hypothesen bzw. eigene Annahmen zu verifizieren oder zu widerlegen[4], zum Beispiel zu folgenden Problemstellungen:

- das Vorkommen bestimmter Phänomene,
- Gebrauchshäufigkeiten,
- Erstdatierung, Herkunft und Wandel,
- Archaisierungen und Neuerungsprozesse im Wortschatz,
- typische Bedeutungs- und Gebrauchsmuster,
- spezifische Gebundenheiten Textsorten, Stilebenen, areale Besonderheiten usw.,
- grammatische Verwendungsspezifika.

Darüber hinaus wird das Korpus in der Regel als **Belegsammlung** im klassischen Sinne benutzt: Man sucht – wie früher in den Zettelkästen – nach besonders aussagekräftigen Beispielen. Eine direkte Betrachtung der auf diese Weise erhaltenen Korpusbefunde – z.B. die intellektuelle Analyse aller Konkordanzen und Belegstellen – funktioniert bis zu einer gewissen Größenordnung auch ohne weitere Unterstützung des Computers. So kann man den folgenden KWIC-Ausschnitt (Key-Word-in-Context) zum Adjektiv *frei* noch allein mit Hilfe der eigenen Kompetenz analysieren und interpretieren (s. rechts oben).

In diesen Konkordanzen vorkommende Wortgruppen wie *freies Geleit, freie Hand bekommen, unter freiem Himmel* oder *auf freiem Fuß* lassen sich auch rein introspektiv als

---

[3] Diese provokant anmutende Formulierung ist ausschließlich in dem o.g. engeren Sinne gemeint. Natürlich sind wir uns sowohl der vielen, teilweise auch prohibitiven Fallstricke der Korpusakquisition als auch der wissenschaftsmethodischen Interpretationsgrenzen der Korpuslinguistik bewusst.

[4] Wir bezeichnen dieses Herangehen als ‚Konsultationsparadigma der Korpusbasiertheit‘ im Unterschied von dem von uns praktizierten ‚Analyseparadigma der Korpusbasiertheit‘.

Korpus: geschr – alle Korpora geschriebener Sprache
Suchanfrage: &frei

KWIC-Übersicht (original/unsortiert)

Anz. Treffer = 436.957
Angezeigter Kontext: 0 Sätze links, 0 Sätze rechts.

T86	Frieden, nach der Friedensbewegung: „Nur	freie Menschen sind wirklich friedlich."
T86	Begründung:	Freies Geleit sei nur den direkten
T86	beschlossen, ihre Anhänger in Athen nach	„freiem Gewissen" wählen zu lassen, anstatt zur
T86	Gespräch mit dem Ziel haben, daß auch die	„freie deutsche Presse" den Kampf gegen den
T86	Lappas unter Auflagen	frei Gericht setzt Beugehaft aus / Widersrüche
T86	den Acht-Stunden- Tag, das	freie Wochenende und den Zwei-Monatsraum, in de
T86	Bundesanwaltschaft als operierende Behörde	freie Hand bekommt.
T86	der neuen „sozial-ökologischen" Politik –	frei nach dem Motto: „Der Fortschritt ist eine
T86	bekundeten guten Willen der Metaller	freie Entfaltung bietet.
A99	Uhr (Türöffnung 19.30), der Eintritt ist	frei.
A99	sahen sich danach der Flötist Jürg	Frei und der Pianist in der G-Dur-Sonatine op.
A99	der ihm seit November gerade mal fünf	freie Sonntage liess.
A99	anzuziehen, stehe jedem einzelnen Zöllner	frei.
A99	Wäldern und an Waldrändern sowie nachts im	Freien an der Leine zu führen.
A99	Ihr Ziel ist allerdings darauf beschränkt,	freie Märkte zu schaffen.
A99	So verbindet sich	freie Improvisation mit angeleiteter
A99	des vergangenen Jahres arbeiten die	Freie Liste Gossau (Flig), der Landesring der
A99	von Eagle County, um später einen	freien Kopf zu haben für die verschiedenen
A99	aus Garmisch-Partenkirchen stürzte beim	freien Skifahren in Beaver Creek und erlitt
T00	Er sagt offen, dass er Kapazitäten	frei hat.
T00	Unter	freiem Himmel also weitere technische Details:
T00	und Beck's und ist doch man ganz selbst.	(freie Lyrik aus der Politikredaktion der taz
T00	vergangene Saison kein Spielplanplatz mehr	frei war, kann er dies erst ab heute
T00	Gesellschaft, die noch als Zusammenschluss	freier Geister funktionierte.
T00	hohe Luftfeuchtigkeit und	frei von störenden Gerüchen – hat fast niemand.
T00	Die Täter sind wieder auf	freiem Fuß.

usuelle Wortverbindungen interpretieren[5]. Ob jedoch *freie Geister, frei nach dem Motto, nach freiem Gewissen* sich ebenso als usuelle Wortverbindungen erweisen, lässt sich eigentlich erst durch die Analyse vieler weiterer Kontextstellen feststellen. Der Wortgruppe *Eintritt ist frei* würde man auf der Basis dieses minimalen Sprachausschnitts sicherlich in den Bereich der völlig freien Kombinierbarkeit verweisen, was aber nicht zutrifft. Die Wortverbindung *frei-Eintritt* ist zum einen sehr gefestigt und darüber hinaus syntaktisch restringiert, wie Steyer/Vachková in ihrem Beitrag zeigen. Man muss also alle Vorkommen betrachten, um zu einer gesicherten Aussage kommen zu können. Die Gesamtzahl der Treffer und damit der Kontextzeilen für das Adjektiv *frei* liegt für die IDS-Korpora bei einer Zahl von über 400 000, was einer Seitenzahl von knapp 1 500 entspricht. Hier stößt der Mensch objektiv an Grenzen. Er kann mit wachsender Anzahl von Belegen nicht mehr beurteilen, ob es sich bei einem sprachlichen Phänomen um ein „singuläres Ereignis", und damit um eine okkasionelle Variante (z.B. Abhängigkeit vom Schreibstil eines

[5] Zum Konzept der usuellen Wortverbindungen vgl. Steyer 2000.

Autors) oder um ein usuelles Muster handelt. Er braucht eine automatische Vorgruppierung und -sortierung nach Wichtigem und Unwichtigem, Typischem und Untypischem, Auffälligem und Unauffälligem. Er braucht Hilfen, die eine Ordnung in die ungeordnete Datenwelt bringen, z.B. mathematisch-statistische Analysemethoden. Mit diesen Methoden kann sehr viel mehr an Vorarbeit geleistet werden, als dies die Lexikografen oft noch glauben mögen. Oder um noch einmal das Schmetterlingsbild aufzugreifen: Wenn wir in der Natur zusammengehörende Tiere finden wollen, müssen wir Methoden haben, die diese Zusammengehörigkeit erkennen. Wenn wir also sprachliche Phänomene erkennen wollen, die sich unserem bisherigen Blick verschlossen haben, müssen wir Methoden entwickeln und einsetzen, die genau in diesem Sinne voraussetzungslos nach auffälligen Vorkommen suchen. Um aber nicht missverstanden zu werden: Diese Methoden können – wenn auch auf immer elaborierterem Niveau – letztlich nur Indikatoren für Evidenzen liefern. Die menschliche Interpretation ist in der am aktuellen Sprachgebrauch orientierten Lexikografie derzeit immer noch die letzte Bewertungsinstanz. Probabilistische Methoden erfassen, analysieren und ordnen Strukturen ohne linguistische Vorannahmen oder Hypothesen. Die Resultate stellen – als Ergebnisse reiner Rechenprozesse – Häufigkeitsbewertungen und Präferenzsetzungen, aber keine Erklärungen für die beobachteten Phänomene dar. Der Mensch interpretiert und bewertet die Rechnergebnisse. Er hat – auf Grund seiner Sprach- und Expertenkompetenz – natürlich Vorannahmen. Er hat Ordnungssysteme im Kopf, mit denen er die Ergebnisse der automatischen Analyse systematisiert und einordnet. Es handelt sich immer um iterative Prozesse zwischen Korpusbefragung und menschlicher Interpretation, wobei ein Ziel sein sollte, dass der Computer dem Menschen immer mehr an vorstrukturierender Analyse „abnehmen" kann.

2. Die statistische Kookkurrenzanalyse

2.1 Kookkurrenzprofile und Kontextmuster

Einen besonderen Stellenwert für die Vorstrukturierung von sprachlichen Massendaten hat das automatische Verfahren der **statistischen Kookkurrenzanalyse (KA)**. Das am IDS entwickelte Kookkurrenzanalysemodul ist seit 1995 in das COSMAS-System integriert und gehört zu den komplexesten seiner Art (Belica 1995). Die statistische Kookkurrenzanalyse erfasst die distributionellen Eigenschaften von lexikalischen Strukturen, da diese als vorkommende Zeichenketten im Korpus unmittelbar beobachtbar sind (s.o.). Es geht um die Erfassung von Zeichenketten, die im Vergleich mit ihrem Gesamtvorkommen statistisch überproportional häufig in der Umgebung anderer Zeichenkettenkonfigurationen vorkommen. Die KA erfasst sie und ordnet sie in hierarchischen Kookkurrenzclustern an[6]:

[6] Zur Beschreibung der Funktionsweise der statistischen Kookkurrenzanalyse vgl. www.ids-mannheim.de/kt/misc/tutorial.html.

Abbildung I

Basis: **"frei"**, Analysetyp **0**

+	-2	1	54160	**Eintritt** ist Kollekte	155	96%	Der Eintritt [...] ist [...] frei [es wird eine] Kolle
+	-2	1	54160	Eintritt ist frei.pd	27	96%	Der Eintritt [...] ist frei.pd
+	-2	1	54160	Eintritt ist	2929	86%	Der Eintritt [...] ist [...] frei
+	-2	1	54160	Eintritt Kollekte	197	93%	Der Eintritt [ist] frei [es wird eine] Kollekte
+	-2	1	54160	Eintritt frei.pd	31	100%	Der Eintritt [ist] frei.pd
+	-2	1	54160	Eintritt	6227	71%	Der Eintritt [ist] frei
+	1	1	37539	**Verkauf** Abonnement Theatergemein	5	100%	Abonnement F grün Theatergemeinde grün
+	1	1	37539	Verkauf Abonnement	153	98%	Abonnement ... freier Verkauf
+	1	1	37539	Verkauf Miete	154	98%	Miete ... und freier Verkauf
+	1	1	37539	Verkauf Theatergemeinde	81	92%	Uhr Theatergemeinde [...] freier Verkauf
+	1	1	37539	Verkauf	3683	92%	freier [...] Verkauf
+	1	1	26352	**Fuß** gesetzt angezeigt	2	50%	freien Fuß gesetzt ... angezeigt
+	1	1	26352	Fuß gesetzt Kaution	46	100%	gegen Kaution auf freien Fuß gesetzt worde
+	1	1	26352	Fuß gesetzt	713	99%	wieder auf freien Fuß gesetzt
+	1	1	26352	Fuß angezeigt	556	99%	wurde auf freiem Fuß [...] angezeigt
+	1	1	26352	Fuß Kaution	92	60%	gegen Kaution auf freien Fuß gesetzt worde
+	1	1	26352	Fuß	2487	58%	wieder auf freiem Fuß angezeigt
+	-1	1	12446	**Plätze** noch sind wenige	70	80%	Es sind [...] noch [...] wenige Plätze frei
+	-1	1	12446	Plätze noch sind	643	77%	Es sind [...] noch [einige] Plätze frei
+	-1	1	12446	Plätze noch wenige	94	86%	Es sind|nur noch [...] wenige [...] Plätze [...] fr
+	-1	1	12446	Plätze noch	1133	67%	sind noch [einige] Plätze [...] frei
+	-1	1	12446	Plätze sind wenige	74	82%	Es sind [noch] wenige Plätze frei
+	-1	1	12446	Plätze sind	753	75%	Es sind [noch einige] Plätze [...] frei
+	-1	1	12446	Plätze wenige	105	89%	sind|nur noch wenige [...] Plätze [...] frei
+	-1	1	12446	Plätze	1871	59%	sind|noch|einige Plätze [...] frei
+	1	1	11994	**Universität** Berlin Künste	1	100%	Künste ... Freien Universität Berlin
+	1	1	11994	Universität Berlin	559	86%	an der Freien Universität [...] Berlin
+	1	1	11994	Universität Künste	5	40%	Künste die Freie Universität
+	1	1	11994	Universität	1536	77%	an der Freien [...] Universität Berlin ...
+	1	1	10341	**Himmel**	1209	92%	unter freiem [...] Himmel
+	1	1	9171	**Lauf** gelassen Phantasie	13	84%	ihrer Phantasie freien Lauf gelassen
+	1	1	9171	Lauf gelassen	96	79%	freien Lauf gelassen
+	1	1	9171	Lauf lassen Phantasie	50	82%	ihrer Phantasie [...] freien Lauf [...] lassen
+	1	1	9171	Lauf lassen	336	88%	ihrer ... freien [...] Lauf [zu] lassen
+	1	1	9171	Lauf Phantasie	96	96%	ihrer Phantasie [...] freien Lauf lassen
+	1	1	9171	Lauf	1023	95%	freien [...] Lauf lassen

Diese statistisch ermittelten Kookkurrenzcluster – wir nennen die Gesamtheit dieser Cluster für eine Bezugseinheit (z.B. für ein Bezugswort) das **Kookkurrenzprofil**[7] – stellen nicht nur einfach usuelle Wortverbindungen einer Sprache dar, sondern elementare Konstituenten der Sprache, Syntagmen, die durch massenhaften Gebrauch entstanden sind und als Bausteine wiederum eingesetzt werden. Kookkurrenzcluster liefern also als holistische Entitäten Evidenzen für rekurrente Muster auf allen Ebenen des Sprachsystems[8]. Der Nutzen solcher Kookkurrenzprofile für die bi- und multilinguale Lexikografie, für die Fremdsprachendidaktik und die Übersetzungspraxis scheint schon in der Natur der zu erkundenden Sache begründet: Es werden usuelle Wortverbindungen einer Sprache erkannt und extrahiert. Über die Relevanz von Wortverbindungen aus fremdsprachlicher Perspektive ist vielfach geschrieben worden[9].

Diese Methode hat aber sehr viel weitreichendere Implikaturen als die – unbestritten sinnvolle – Extraktion von lexikalischen Kookkurrenzen. Sie macht vor allem **systematisierte Kontextmuster** transparent. Diese Kontextmuster stellen für uns den eigentlichen Schlüssel bei der Lösung von Interferenzproblemen und vorliegenden Asymmetrien von Kontrastsprachen dar. Sie erfüllen distinktive Funktionen hinsichtlich der situativen Verwendung von Einheiten in der Fremdsprache und sind mit den klassischen Lesartenkonzepten vor allem der einsprachigen lexikalischen Semantik und Lexikografie eigentlich nicht kompatibel. Dieses Potenzial der feinen ‚Kontextdisambiguierung' durch Kookkurrenzcluster scheint uns vor allem für ein mittleres bis gehobenes Sprachniveau (bis hin zu den ausgefeilten stilistischen Bedürfnissen von professionellen Übersetzern) von besonderer Bedeutung zu sein. Der Einsatz dieser Methoden kann zum Erreichen jener Qualität beitragen, die Stubbs als „kulturell angemessenes Kommunizieren" beschrieben hat:

„Bei solchen Kombinationen geht es um Wahrscheinlichkeiten, Erwartungen und quantitative Verteilungen. Es geht um Normen des Sprachgebrauchs. Hier wird die Verbindung zwischen Idiomatik und kultureller Kompetenz deutlicher: man muß Versprachlichungen kulturell festgelegter Konzepte (z.B. Sprechakte), also landeskundlicher Informationen, beherrschen, um sich idiomatisch ausdrükken zu können. Durch den Gebrauch von idiomatischen Redewendungen werden gemeinsame Wertvorstellungen abgerufen. Nur wer die idiomatischen Ausdrücke

---

[7] In einer IDS-internen korpuslinguistischen Analyse-, Evaluierungs- und Experimentierplattform – der CCDB-Kookkurrenzdatenbank – sind derzeit etwa 100 000 Kookkurrenzprofile gespeichert. Ein Teil der Plattform ist online zugänglich über http://corpora.ids-mannheim.de/ccdb.

[8] Zu linguistischen Interpretationen und lexikografischen Anwendungen der Korpusanalysemethoden vgl. Steyer u. a. 2002, 2004.

[9] Hier sind natürlich vor allem die Arbeiten von Hausmann zu Kollokationen aus kontrastiver und fremdsprachendidaktischer Perspektive (u. a. 1984, 2003) zu nennen, für das Sprachenpaar Deutsch-Französisch vgl. auch Steyer 1998, Steyer – Teubert 1998.

und die damit verbundenen Wertungen oder Vorstellungen in einer Kultur kennt, kann also Texte verstehen, bzw. erfolgreich kommunizieren." (1997, 157)

Wir können dies hier nur punktuell illustrieren. Differenzierte Untersuchungen bleiben weiteren Forschungen der Autoren vorbehalten.

Im Folgenden werden zwei Zugänge zur Interpretation von Kookkurrenzprofilen etwas näher erläutert: die analytisch-synthetische Konstitution heuristischer Kookkurrenzfelder zum einen und die Interpretation von Clusterhierarchien und syntagmatischen Mustern zum anderen.

2.2 Heuristische Kookkurrenzfelder

Im Mittelpunkt dieser Herangehensweise steht die analytische Betrachtung, linguistische Interpretation und Systematisierung **aller** ermittelten Kookkurrenzpartner im Kookkurrenzprofil einer sprachlichen Bezugseinheit (Wort- oder Wortgruppe). Das für diesen Beitrag ermittelte Kookkurrenzprofil des Adjektivs *frei* enthält über 200 primäre Kookkurrenzpartner, z.B. *Eintritt, Verkauf, Fuß, Plätze, Universität, Himmel, Lauf, Wahlen, Fahrt, Markt, ist, gesetzt, angezeigt.*

Es geht nun darum, nach Gemeinsamkeiten zwischen diesen einzelnen Kookkurrenzpartnern zu suchen und diese zu entsprechenden Gruppen (Kookkurrenzfeldern) zusammenzufassen. Solche Kookkurrenzfelder stellen ein heuristisches Mittel dar, um Evidenzen für die aktuelle Bedeutung und den aktuellen Gebrauch der Bezugseinheiten deutlich zu machen. Diese Gemeinsamkeiten können alle Ebenen des Sprachsystems betreffen, also sowohl morphologischer als auch lexikalisch-semantischer, pragmatischer oder grammatischer Natur sein. Theoretischer Hintergrund ist die schon durch John Rupert Firth, dem Begründer des britischen Kontextualismus und Vater der modernen Korpuslinguistik, und später auch im Distributionalismus vertretene These, dass typische Umgebungspartner und die Beziehung zwischen Wörtern sehr viel über das einzelne Wort selbst aussagen: „You shall know a word by the company it keeps" (Firth, 1957)[10]. Mit den heutigen Möglichkeiten der systematischen Analyse sprachlicher Massendaten erlangt diese These eine völlig neue Dimension.

Die Kriterien für das Ordnen in Kookkurrenzfelder hängen dabei immer vom jeweiligen Analyseinteresse, von der Relevanz für den konkreten Anwendungszusammenhang und vom zu Grunde gelegten linguistischen Modell ab. Eine zentrale Frage ist z.B., ob die Kookkurrenzen aus muttersprachlicher oder fremdsprachiger Perspektive betrachtet und interpretiert werden. Natürlich denkt man zunächst an die Systematisierung der Kookkurrenzpartner nach morpho-syntaktischen Kriterien (also z.B. alle

[10] In seiner Nachfolge u. a. Sinclair (1991) und Stubbs (1997). Auf die sprachtheoretische Herausforderung dieser Perspektive kann an dieser Stelle nicht eingegangen werden. Verwiesen sei vor allem auf die Forschungsrichtungen der *Corpus Pattern Analysis* (vgl. Hanks 2004) und der *Corpus Driven Linguistics* (vgl. Tognini-Bonelli 2001), in deren Kontext auch die aktuellen korpuslinguistischen Forschungen am IDS einzuordnen sind.

nominalen, verbalen Partner von frei), wie es in der Kollokationsforschung in vielen Fällen praktiziert wird und was auch schon viele automatische Tools leisten:

Nominale Kookkurrenzpartner von *frei*

- *Eintritt, Verkauf, Fuß, Plätze, Universität, Himmel, Lauf, Wahlen, Fahrt, Markt, Meinungsäußerung, Wildbahn, Träger, Weg, Marktwirtschaft, Wähler, Demokraten, Natur, Hand, Zugang, Uhr, Training, Volksbühne* usw.

Verbale Kookkurrenzpartner von *frei*

- *gesetzt, erfunden, übersetzt, gewählt, geworden, bewegen, wählen, herumlaufen, leben, entscheiden, bestimmen* usw.

Unsere empirischen Pilotuntersuchungen haben aber immer wieder gezeigt, dass solche wortartenbezogenen Zuordnungen nur in ganz speziellen Fällen wirklich aussagekräftig sind, da die Kookkurrenzpartner innerhalb eines solchen Feldes größtenteils völlig verschiedene syntagmatische Gebrauchsmuster mit teils sehr unterschiedlichen syntaktischen Realisierungen und pragmatischen Funktionen indizieren. Für das Adjektiv frei bilden vor allem folgende Kookkurrenzfelder relevante Aspekte des aktuellen Gebrauchs ab:

- Semantische Felder
- Thematische und domänenspezifische Felder
- Onymische Felder
- Modifizierende Felder

Semantische Felder u. a.

- ‚kostenlos': *Eintritt, Zugang, Zutritt*
- ‚nicht besetzt': *Platz, Zimmer, Stelle, Betten, Parkplatz*
- ‚nicht eingeschränkt/reglementiert': *Verkauf, Zugang, Fahrt, Training, Wahlen, Arztwahl, Aussicht, Personenverkehr*
- ‚kreativ': *Technik, Training, Improvisation, entfalten, Stil*

Thematische und domänenspezifische Felder

- ‚Grundrechte': *Meinungsäußerung, Wahlen, Bürger, Presse, Entfaltung, Religionsausübung, Wort, Berichterstattung*
- ‚Wirtschaft': *Marktwirtschaft, Wettbewerb, Warenverkehr, Aktionäre, Wirtschaftsverband, Handel, Kapitalverkehr, Strommarkt, Unternehmertum*
- ‚Berufe': *Journalist, Autor, Schriftsteller, Publizist*

Onymische Felder

- Parteien/Institutionen/Vereine/Verbände: *Freie Universität Berlin, Freie Wähler, Freie Evangelische Gemeinde, Sender Freies Berlin, Freie Presse, Freie Wohlfahrts-*

pflege, Freie Demokraten (FDP), Bund freier Bürger, Bund freier Waldorfschulen, Freie Hansestadt Hamburg, Nationalkomitee Freies Deutschland, Bundesverband Freier Tankstellen

Modifizierende Felder

- *völlig, weitgehend, genügend, relativ, möglichst, ganz, grundsätzlich, gänzlich, absolut, vollkommen, total, bedingt, ziemlich, nahezu, teilweise, endgültig*

Auf diesem methodischen Weg der Gruppierung von Kookkurrenzclustern lassen sich auch Inventare usueller fester Wortverbindungen (Idiome, Teilidiome, Slogans, kommunikative Formeln usw.) erstellen, hier nur auszugsweise zum Bezugswort *frei*:

Fuß	[*auf freien Fuß setzen*]
Lauf	[*freien Lauf lassen*]
Himmel	[*unter freiem Himmel*]
Meinungsäußerung	[*Recht auf freie Meinungsäußerung*]
Fahrt Bürger	[*freie Fahrt für freie Bürger*]
Träger	[*Freie Träger*]
Wildbahn	[*in freier Wildbahn*]
Weg	[*Weg frei machen*]
Hand	[*freie Hand lassen*]
Fall	[*im freien Fall*]
Motto	[*frei nach dem Motto: „…"*]
Stücken	[*aus freien Stücken*]
Erfunden	[*frei erfunden*]
Bewegen	[*frei bewegen*]
Wirtschaft	[*freie Wirtschaft*]
Frank	[*frank und frei*]
Fair	[*frei und fair*]
Bahn	[*Bahn frei! Freie Bahn*]
Ring	[*Ring frei!*]
Geleit	[*freies Geleit*]
Radikale	[*freie Radikale*]
Logis	[*freie Kost und Logis*]
Betten	[*freie Betten*]
Haus	[*frei Haus*]
Fahrt	[*freie Fahrt*]
Manege	[*Manege frei!*]
Kräfte	[*freies Spiel der Kräfte*]
Kopf	[*frei im Kopf sein*]

Empfangbar	[*frei empfangbar*]
Rede	[*freie Rede*]
Minute	[*jede freie Minute*] usw.

Relevant werden alle diese Systematisierungen – z.B. für ein zweisprachiges Wörterbuch – vor allem in semantischer Hinsicht, wenn es um das Erkennen von Teilbedeutungen oder speziellen thematischen Verwendungsaspekten sprachlicher Einheiten geht, für die es in der Fremdsprache keine Äquivalente gibt oder die so noch nicht in den Wörterbüchern kodifiziert sind.

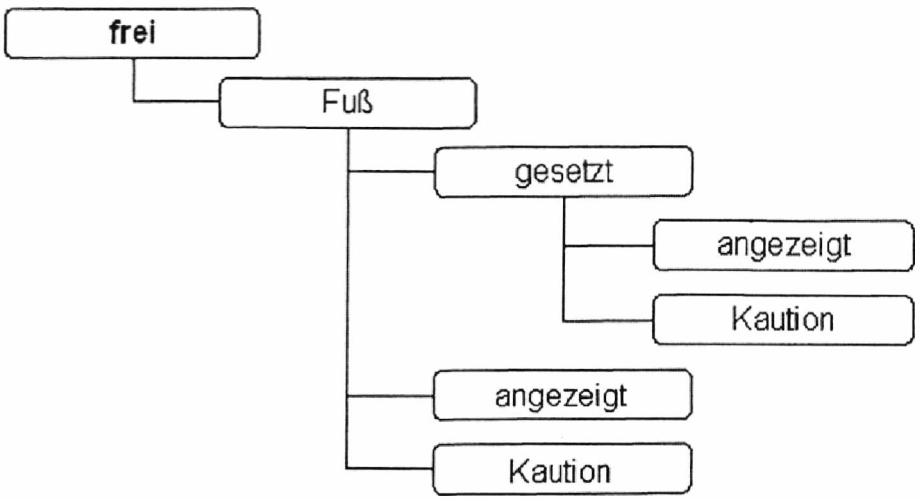
2.3 Clusterhierarchien und syntagmatische Muster

Der Betrachtungsgegenstand ist aus dieser Perspektive nicht mehr die Menge aller lexikalischen Kookkurrenzpartner in ihrem Verhältnis zur Bezugseinheit, sondern jedes einzelne Kookkurrenzcluster als holistische Einheit. Es interessiert also nicht mehr vordergründig die **Bezugseinheit** als solche, z.B. was das Adjektiv *frei* bedeutet, sondern es interessiert, welche Gebrauchssyntagmen mit *frei* konstituiert werden und welche Funktionen sie im Sprachgebrauch erfüllen. Die KA liefert dazu Informationen auf mehreren Ebenen: Sie ermittelt und visualisiert Determiniertheitsverhältnisse der Komponenten untereinander in Clusterhierarchien (Bezugswort – primäre Kookkurrenzpartner – sekundäre usw. Partner) und ihre syntaktischen Realisierungen in Form von syntagmatischen Mustern:

Abbildung II

	1	1	26352	**Fuß gesetzt angezeigt**	2	50%	freien Fuß gesetzt ... angezeigt
+	1	1	26352	Fuß gesetzt Kaution	48	100%	gegen Kaution auf freien Fuß gesetzt worden
+	1	1	26352	Fuß gesetzt	713	99%	wieder auf freien Fuß gesetzt
+	1	1	26352	Fuß angezeigt	556	99%	wurde auf freiem Fuß [...] angezeigt
+	1	1	26352	Fuß Kaution	92	60%	gegen Kaution auf freien Fuß gesetzt worden
+	1	1	26352	Fuß	2487	58%	wieder auf freiem Fuß angezeigt

Hinter dieser Abbildung verbirgt sich zunächst folgende Clusterhierarchie:



Man kann anhand der Darstellung erkennen, dass ein deutliches Vorkommen der Relation *frei – Fuß* vorliegt. Das Lexem *Fuß* ist primärer Kookkurrenzpartner. Es gibt darüber hinaus aber weitere Auffälligkeiten in der Umgebung dieser Relation: Die flektierten Verben *angezeigt* und *gesetzt* und das Nomen *Kaution* weisen ebenfalls eine auffällige Affinität zu *frei – Fuß* auf. Diese Clusterhierarchien dienen in erster Linie zum Erkennen konzeptueller Zusammenhänge zwischen lexikalischen Einheiten. Sie sagen aber noch nichts über die natürlichsprachigen Realisierungen aus, über die Position der Komponenten im Syntagma und über deren typischen Abstand, im Unterschied zur Sinclairschen Analysespanne (*span*) (1991, 170). Einen ersten Anhaltspunkt dafür bietet die Fokusangabe zur typischen Stellung des Kookkurrenzpartners zum Bezugswort (siehe „1 1" in obiger Abbildung II). Man kann bereits anhand dieses Wertes Modifikationsanfälligkeiten beurteilen (was u. a. sehr hilfreich für die Bestimmung von Normal- oder Grundformen von Wortverbindungen sein kann). Kookkurrenzcluster mit einer Fokusangabe von „1 1" – wie in unseren Beispiel – weisen einen solch engen Grad syntaktischer Gebundenheit auf, dass man von einer ausgeprägten Modifikationsresistenz sprechen und deshalb von einem hohen Festigkeitsgrad ausgehen kann. Cluster mit einer Spanne von zum Beispiel „– 5 + 5" sind dagegen modifikationsanfälliger und damit syntaktisch weniger fest.

Aussagekräftiger sind dann aber die syntagmatischen Einbettungen von Bezugseinheiten und Kookkurrenzpartnern in ihrer natürlichsprachigen syntaktischen Realisierung. Mit Hilfe des hier beschriebenen Korpusanalysemoduls können solche **syntagmatischen Muster** (vgl. Belica 2004) mit rein mathematisch-statistischen Mitteln aufgespürt und extrahiert werden, ohne den Apparat der regelbasierten Phrasenstrukturanalyse auf der Basis morpho-syntaktisch annotierter Korpora zu Grunde legen zu müssen. Diese syntagmatischen Muster stellen z.B. Indikatoren für übliche präposi-

tionale Anschlüsse und typische syntaktische Strukturen dar. Beispiele für berechnete syntagmatische Muster zu *frei* enthält die folgende Abbildung:

Abbildung III

96% Der Eintritt [...] ist [...] **frei** [es wird eine] Kollekte erhoben
96% Der Eintritt [...] ist **frei.pd**
86% Der Eintritt [...] ist [...] **frei**
93% Der Eintritt [ist] frei [es wird eine] Kollekte
100% Der Eintritt [ist] **frei.pd**
71% Der Eintritt [ist] **frei**
100% Abonnement F grün **Theatergemeinde** grün freier Verkauf
98% Abonnement ... **freier Verkauf**
98% Miete ... und **freier Verkauf**
92% Uhr **Theatergemeinde** [...] **freier Verkauf**
92% **freier** [...] Verkauf
50% **freien Fuß gesetzt** ... **angezeigt**
100% **gegen Kaution auf freien Fuß gesetzt** worden
99% **wieder auf freien Fuß gesetzt**
99% wurde **auf freiem Fuß** [...] **angezeigt**
60% **gegen Kaution auf freien Fuß gesetzt** worden
58% **wieder auf freiem Fuß** angezeigt
80% Es sind [...] **noch** [...] **wenige Plätze frei**
77% Es sind [...] **noch** [einige] **Plätze frei**
86% Es sind|nur **noch** [...] **wenige** [...] **Plätze** [...] **frei**
67% sind **noch** [einige] **Plätze** [...] **frei**
82% Es sind [noch] **wenige Plätze frei**
75% Es sind [**noch einige**] **Plätze** [...] **frei**
89% sind|nur **noch wenige** [...] **Plätze** [...] **frei**
59% sind|noch|einige **Plätze** [...] **frei**
100% Künste ... **Freien Universität Berlin**
86% an der **Freien Universität** [...] **Berlin**
40% Künste die **Freie Universität**
77% an der **Freien** [...] **Universität Berlin** ...
92% unter **freiem** [...] **Himmel**
84% ihrer Phantasie **freien Lauf gelassen**
79% **freien Lauf gelassen**
82% ihrer Phantasie [...] **freien Lauf** [...] **lassen**
88% ihrer ... **freien** [...] **Lauf** [zu] **lassen**
96% ihrer Phantasie [...] **freien Lauf lassen**
95% **freien** [...] **Lauf lassen**

Die syntagmatischen Muster zu *frei – Fuß*:

... freien Fuß gesetzt... angezeigt
gegen Kaution auf freien Fuß gesetzt worden
wieder auf freien Fuß gesetzt
wurde auf freiem Fuß [...] angezeigt
gegen Kaution auf freien Fuß gesetzt worden
wieder auf freiem Fuß angezeigt

lassen eine präpositionale Invarianz (Präposition: *auf*) in zwei möglichen Kasusrealisierungen (Dativ und Akkusativ) erkennen. Des Weiteren tritt die Präpositionalphrase *gegen Kaution* auffällig hervor und lässt auf eine Formel *gegen Kaution auf freien Fuß* setzen schließen, eine für Nichtmuttersprachler nicht unbedingt einfach ableitbare feste Konstruktion. Auf die semantischen Ausdifferenzierungen zwischen den Subclustern *frei – Fuß – angezeigt vs. frei – Fuß – gesetzt* und den Wert dieser auf diese Weise ermittelten syntagmatischen Muster aus kontrastiver Sicht werden Steyer/Vachková noch einmal genauer eingehen.

Die datengeleitete Ermittlung typischer syntagmatischer Strukturen führt dazu, dass man nicht nur Resultate erhält, die der klassischen linguistischen Auffassung von Phrasenstrukturen entsprechen, sondern dass auch in großer Zahl ‚idiosynkratisch' scheinende Wortverbindungen hervortreten, die nicht unbedingt regelgeleitet kombiniert sind, die jedoch in genau dieser Konstruktion auffällig häufig, also **usuell**, verwendet werden. Für das Adjektiv *frei* betrifft dies z.B. die Gruppe der asyndetischen Konstruktionen, bei denen *frei* dem Nomen unflektiert nachgestellt wird[11]. Diese fixe syntaktische Struktur kommt zu 100 % in den entsprechenden Clustern vor.

Eintritt frei, Weg frei, Zimmer frei, Ring frei, Bahn frei, Start frei, Feuer frei, Bühne frei.

2.4 Umgebungsmuster von Kookkurrenzclustern

Die KA bietet die Möglichkeit, durch weitere spezifizierende Umgebungsanalysen einer Wortverbindung externe Valenzen und Ergänzungen bzw. typische Verwendungsmuster von Wortverbindungen zu erkennen und weiter zu spezifizieren:

Nominale Kookkurrenzpartner von ***freien Lauf lassen***

- *Phantasie, Kreativität, Gefühl, Emotion, Gedanken, Freude, Bewegungsdrang, Begeisterung, Spieltrieb, Assoziation, Temperament,*

[11] Konstruktionen wie solche vom Typ asyndetischer Appositionen wurden von Schmidt in mehreren Studien (u.a. 1998) im Kontext der Analyse von Formulierungstraditionen untersucht, wobei die Konstruktionen mit dem nachgestellten und unflektierten Adjektiv *frei* in unserem Beispiel wohl eher als – um das Verb *sein* reduzierte – Ellipsen zu interpretieren sind.

- *Unmut, Frust, Wut, Zerstörungswut, Aggression, Tränen, Enttäuschung, Zorn, Empörung, Hass,* usw.

Diese externen Valenzangaben dienen wiederum als heuristischer Zugang zu aktuellen Bedeutungen und Verwendungen einer Wortverbindung:

Semantische Kookkurrenzfelder zu **unter freiem Himmel**

‚veranstalten, stattfinden'
- *Gottesdienst, Versammlung, Veranstaltungen, Spektakel, Kino, Konzertfilme, Fest, Bieranstich, Prozession, Unterricht, Demonstration, Steinzeitkunst*
- *stattfinden, genießen, Bewirtung, wo Speisen unter... zubereitet werden, auf Motivsuche, Bierchen, gemütliches Beisammensein,*
- *bei lauen Temperaturen, bei schönem Wetter, im Sommer, bei Regen*

‚übernachten, leben'
- *schlafen, übernachten, lagern, campieren, nächtigen, die Nacht verbringen, leben*
- *in Zelten, im Schlafsack,*
- *bei Minustemperaturen,*

‚verbotenerweise'
- *Versammlungen, Glücksspiel, unbewacht, NATO-Munition, radioaktiver Müll einfach so*

‚obdachlos'
- *Zehntausende Flüchtlinge*

Gerade auf dem Gebiet der Erforschung und Beschreibung solcher Umgebungsmuster von Wortverbindungen können die automatischen Analysemethoden einen großen Erkenntnisfortschritt auch im Bereich der Mikrostrukturen von Texten bringen. Neben der – wie eben angedeutet – Disambiguierung von pragma-semantischen Aspekten betrifft dies insbesondere auch die Problematik der Invarianz, Variation und Modifikation.

3. Schlussbemerkung

Die mit Hilfe mathematisch-statistischer Methoden ermittelten Korpusbefunde werden die Kompetenz und die Spracherfahrung des Lexikografen nie völlig ersetzen, aber sie können in einem ungleich größerem Maße zu einer am wirklichen Sprachgebrauch orientierten Sprachbeschreibung beitragen. Diese Wirkung kann sich umso mehr entfalten, je subtiler und elaborierter die Korpusanalysemetho-

den werden. Damit wachsen aber gleichzeitig die methodisch-methodologischen Anforderungen in Hinblick auf die linguistische Interpretation und eine noch direktere lexikografische bzw. didaktische Nutzbarkeit der Ergebnisse. Die Herausforderung korpusbasierten Arbeitens besteht nach unserer festen Überzeugung darin, die Analysemethoden einer ständigen Reflexion zu unterziehen, sich also stets zu fragen, was man mit ihnen erklären kann und was auch nicht. Auch für die Auslandsgermanistik und den Bereich ‚Deutsch als Fremdsprache' eröffnet sich hier ein spannungsreiches Forschungs- und Anwendungsfeld.

Literatur

BELICA, C. (1995): *Statistische Kollokationsanalyse und Clustering. Korpusanalysemodul.* Institut für Deutsche Sprache. Mannheim. (http://corpora.ids-mannheim.de)

— (2004): *Methoden der Korpusanalyse und -erschließung.* Vortrag. IDS-Kolloquium am 18. 5. 2004. Mannheim.

BELICA, C. – NEUMANN, R. (1997): *Das wissenschaftliche Konzept der Zentralen Arbeitsstelle Linguistische Datenverarbeitung. Unterlagen zur Evaluation.* Institut für Deutsche Sprache. Mannheim. (Unveröffentlicht).

CHURCH, K. – MERCER, R. (1993): Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *CL* 19:1. 1–24.

FIRTH, J. (1957): A Synopsis of Linguistic Theory 1930–1955. In: *Studies in Linguistic Analysis.* Philological Society. Oxford. Reprinted in Palmer, F. (ed. 1968). Selected Papers of J. R. Firth. Longman. Harlow.

MECOLB (1993): *Project Proposal. MLAP Call 1993. Exploratory Actions for the Language Industry.* Feasibility and Validation Study. Luxembourg.

HANKS, P. (2004): The Syntagmatics of Metaphor and Idiom. In: *International Journal of Lexicography.* Volume 17. Number 3. 245–274.

HAUSMANN, F. J. (1984): Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: *Praxis des neusprachlichen Unterrichts.* Jg. 31. 395–406.

— (2004): Was sind eigentlich Kollokationen? In: *Wortverbindungen – mehr oder weniger fest.* Hrsg. von K. Steyer. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York. 309–334.

PERKUHN, R. – BELICA, C. – AL WADI, D. – LAUER, M. – STEYER, K. – WEISS, C. (2005): Korpustechnologie am Institut für Deutsche Sprache. In: *Tagungsband der Internationalen Konferenz „Korpuslinguistik deutsch: synchron – diachron – kontrastiv".* 20.–23. 3. 2003. Universität Würzburg.

PROCEEDINGS (1993): *Proceedings of the First Workshop on Very Large Corpora: Academic and Industrial Perspectives.* Columbus. Ohio.

SCHMIDT, H. (1998): Traditionen des Formulierens: Apposition, Triade, Alliteration, Variation. In: *Das 20. Jahrhundert. Sprachgeschichte – Zeitgeschichte.* Hrsg. von H. Kämper und H. Schmidt. Jahrbuch des Instituts für Deutsche Sprache 1997. 86–117.

SINCLAIR, J. M. (1991): *Corpus, Concordance, Collocation.* Oxford.

STEYER, K. (1998): Kollokationen als zentrales Übersetzungsproblem – Vorschläge für eine Kollokationsdatenbank Deutsch-Französisch/Französisch-Deutsch auf der Basis paralleler und vergleichbarer Korpora. In: *Lexikologie und Lexikographie Deutsch-Französisch.* Hrsg. von D. Bresson. Cahiers d'Études Germaniques 35. Aix-en-Provence. 95–113.

— (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: *Deutsche Sprache* 2/00. 101–125.

— (2002): Wenn der Schwanz mit dem Hund wedelt. Zum linguistischen Erklärungspotenzial der korpusbasierten Kookkurrenzanalyse. In: *Ansichten der deutschen Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag.* Hrsg. von Haß-Zumkehr, U. – Kallmeyer, W. – Zifonun, G. Studien zur Deutschen Sprache 25. Tübingen. 215–236.

— (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: *Wortverbindungen – mehr oder weniger fest.* Hrsg. von K. Steyer. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York. 87–116.

STEYER, K. – TEUBERT, W. (1997): Deutsch-Französische Übersetzungsplattform. Ansätze, Methoden, empirische Möglichkeiten. In: *Deutsche Sprache* 4/97. 343–359.

STUBBS, M. (1997): „Eine Sprache idiomatisch sprechen": Computer, Korpora, Kommunikative Kompetenz und Kultur. In: *Norm und Variation.* Hrsg. von K. Mattheier. Forum Angewandte Linguistik 32. Frankfurt a. M./Berlin/Bern/New York/Paris/Wien. 151–167.

TOGNINI-BONELLI, E. (2001): Corpus linguistics at work. Studies in Corpus linguistics 6. Amsterdam.

CONTRIBUTIONS TO THE STUDY OF GERMAN USAGE A CORPUS-BASED APPROACH

Summary

This paper outlines some basic assumptions and principles underlying the corpus linguistics research and some application domains at the Institute for German Language in Mannheim. We briefly address three complementary but closely related tasks: first, the acquisition of very large corpora, second, the research on statistical methods for automatically extracting information about associations between word configurations, and, third, meeting the challenge of understanding the explanatory power of such methods both in theoretical linguistics and in other fields such as second language acquisition or lexicography. We argue that a systematic statistical analysis of huge bodies of text can reveal substantial insights into the language usage und change, far beyond just collocational patterning.

KORPUS A ZKOUMÁNÍ NĚMECKÉHO JAZYKOVÉHO ÚZU

Resumé

Příspěvek prezentuje základní předpoklady a principy, z nichž vychází výzkum korpusů a jeho některé aplikace v Institutu německého jazyka v Mannheimu (Institut für Deutsche Sprache in Mannheim). Stručně charakterizuje tři navzájem úzce související cíle: získávání velmi obsáhlých korpusů, výzkum statistických metod určených pro automatizovanou extrakci informací o společném výskytu slov v možných konfiguracích a konečně interpretační hodnotu takových metod jak pro lingvistiku teoretickou, tak pro ostatní oblasti, jakými je například výuka cizích jazyků či lexikografie. Autoři studie ukazují, že tyto metody mají daleko širší užití než odhalení kolokačních struktur: Systematická statistická analýza velkých textových korpusů může zásadním způsobem objasnit navíc jazykový úzus a jeho proměny.