Erschienen in: Kämper, Heidrun/ Eichinger, Ludwig M. (Hrsg.): Sprachperspektiven.
Germanistische Linguistik und das institut für Deutsche Sprache -
Tübingen: Narr, 2007. S. 465-491.
(Studien zur Deutschen Sprache 40)

Rainer Perkuhn

„Corpus-driven“: Systematische Auswertung automatisch ermittelter sprachlicher Muster

> Eine fundamentale Aufgabe jeder Wissenschaft ist die Schaffung einer Ordnung, das Finden von Mustern in der Menge mannigfaltiger, unübersichtlicher Daten.
>
> (Köhler 2005, S. 4)

1. Vorbemerkung

Als Ihre Waschmaschine vor nicht allzu langer Zeit defekt war, ist Ihnen da nicht aufgefallen, dass auch andere Geräte kurz vorher repariert werden mussten? Und als Sie Ihrem Kollegen davon erzählt haben, meinte dieser nicht, dass er auch seinen Staubsauger ersetzen musste? Und kannte nicht Ihr Kollege jemanden, dem alle Geräte innerhalb von vier Wochen ausgefallen waren? In solchen Situationen ist man schnell mit Aussagen bei der Hand, dass nichts mehr so ist wie früher oder dass die Qualität immer schlechter wird.

Dieses Beispiel lässt sich leicht übertragen auf Themen wie Umweltkatastrophen, Kriminalität, die Jugend von heute oder auch den Gebrauch der deutschen Sprache. Auch hier kommt man schnell zu dem Schluss, auch die deutsche Sprache ist nicht mehr das, was sie mal war. Sie verkommt immer mehr – und auch der Einfluss anderer Sprachen wird immer größer!

Vollkommen unabhängig davon, ob irgendeine der gerade plakativ vorgebrachten Aussagen tatsächlich stimmt oder nicht – hier soll es nicht um den Inhalt der Aussage gehen –, soll an dieser Stelle nur betrachtet werden, wie die Pauschalisierungen als Sonderfall von Aussagen über Typisches zustande kommen. Unser Interesse gilt dem Typischen in der Sprache, aus dem Musterhaftes abgeleitet werden kann, unserer Meinung nach der Schlüssel zum Beherrschen einer Sprache und zum Verstehen der Sprache.[1] Allen Beispielen ist gemeinsam, dass sich das beobachtete Phänomen in gewisser Weise verselbstständigt. Es wird nicht darauf geachtet, welches Ereignis tatsächlich beobachtet werden kann. Es wird auch nicht reflektiert, ob sich die

[1] Die in diesem Beitrag ausgearbeiteten Ideen basieren auf der gemeinsamen Arbeit mit Cyril Belica im Projekt „Methoden der Korpusanalyse und -erschließung“.

Art der Beobachtung im Laufe der Zeit ändert. Und auch das Bewertungssystem, vor dessen Hintergrund die qualitativen Aussagen formuliert werden, ist nicht vor und während der Beobachtung fest vorgegeben.

Bei den pauschalisierenden Aussagen wird häufig unkritisch in Kauf genommen,

- dass nur eine kleine Zahl an Ereignissen in einem kleinen Zeitrahmen tatsächlich beobachtet wurde, ohne diese in ein Verhältnis zur Grundgesamtheit und zu einem zeitlichen Verlauf zu setzen:
  - im Beispiel wurden nur sehr wenige defekte Haushaltsgeräte beobachtet ohne Kenntnis, wie viele Geräte insgesamt tatsächlich über welchen Zeitraum im Gebrauch sind und wie viele Schadensfälle insgesamt zu verzeichnen sind;
- dass anschließend die Beobachtung gezielt auf weitere Ereignisse gelenkt wurde:
  - im Beispiel wurde bewusst nach weiteren Schadensfällen gesucht;
- dass zwischen Beobachtung und Bewertung eine Dynamik entsteht, um ggf. noch mehr Beobachtetes in die Reihe des bisher Bewerteten einordnen zu können:
  - im Beispiel wurde nicht unterschieden, ob es sich um ein Markenfabrikat handelt, wie lange und intensiv das Gerät in Betrieb war und ob nicht vielleicht ein Fehler wegen unsachgemäßer Bedienung vorgelegen hat.

Gerade die Medien in unserer Zeit bergen ein großes Potenzial, um die Aufmerksamkeit der Leser zu lenken. Wenn in kurzen Abständen über bestimmte Phänomene geballt berichtet wird, wird der Leser zwar temporär übersensibel für diesen Bereich. Für eine angemessene Bewertung müsste der Leser aber im Grunde auch wissen, wie oft das Phänomen bereits früher aufgetreten ist (oder auch während der verstärkten Berichterstattung undokumentiert bleibt), um eine tatsächliche Veränderung im Gesamtbestand bewerten zu können. Besonders deutlich wird dies, wenn zum Erkennen des Phänomens eine Prüfung oder eine Messung erforderlich ist: Wenn in einer bestimmten Phase Messungen auffällige Werte ergeben, wird gern außer Acht gelassen, dass in anderen Phasen gar keine Messungen durchgeführt wurden, so dass gar keine auffälligen Werte festgestellt werden konnten.

Für eine objektive Einschätzung bedarf es einer Studie über einen längeren Zeitraum mit einem vorher festgelegten und transparenten Bewertungssystem. Damit ließe sich feststellen, nach welchem Zeitraum in welchem Szenario (oder genormt: Anzahl Waschgänge bei einem Mix aus Temperaturen) bei sachgemäßer Bedienung wie viele Schäden durchschnittlich auftreten. Erst eine Änderung dieses Durchschnittswertes im Laufe der Zeit (evtl. auch weiterer statistischer Maße wie der Varianz) erlauben Aussagen über Verbesserung oder Verschlechterung der Qualität.

Eine Korrelationsanalyse kann Zusammenhänge feststellen, ob zwischen Alter bzw. Intensität der Nutzung und Schadenshäufigkeit, evtl. auch zwischen Marke und Schadenshäufigkeit Abhängigkeiten bestehen. Waschmaschinen haben z.B. eine 'Lebenserwartung' von 12 Jahren, danach lohnt es sich nicht mehr, sie reparieren zu lassen. Die Aussage über die Qualität von oben ließe sich nur bestätigen, wenn die Lebenserwartung im Laufe der Zeit abnimmt, dieselben Schäden bereits bei geringer Nutzung auftreten oder die Unterschiede zwischen Marken- und 'Billig'-Gerät abnehmen. Neben der rein quantitativen Bewertung ist häufig noch eine Interpretation der Ergebnisse erforderlich.

Bei der Untersuchung des Gebrauchs der Sprache verhält es sich genauso. Natürlich kann man sich auch in der Linguistik um die Untersuchung von einzelnen Ereignissen kümmern. Um den Gebrauch einer Sprache zu dokumentieren, möchte man aber nicht eine Sammlung von Einzelfällen aufstellen. Vielmehr ist stets das Bestreben, zu generalisieren, und dies funktioniert nur, wenn man von einem Konzept von typischen oder auffälligen Mustern in der Sprache ausgeht (seien es 'Lesarten' in der Lexikografie, 'Valenzangaben', 'Regeln' in einer Grammatik oder 'Sprachhandlungsschemata' in der Soziolinguistik usw.).

Wie kommt man aber zu Typischem oder Auffälligem, ohne der oben diskutierten Versuchung zu erliegen?

Für eine solide Grundlage muss man zunächst möglichst viele authentische Daten in der Breite unabhängig von irgendwelchen Phänomenen sammeln. Desweiteren ist ein Bewertungssystem explizit zu machen und verbindlich festzulegen, mit dem dann im Hinblick auf die Datensammlung als Ganzes diese methodisch ausgewertet werden kann. Erst im Anschluss daran kann man versuchen, die Phänomene zu erläutern und zu beschreiben, evtl. unter Bezugnahme auf bereits bekannte Erklärungsmodelle.

Im obigen Beispiel hieße das, dass man längerfristig Schadensfälle registrieren und mit möglichst viel Informationen sammeln müsste, z.B. wie oft am Tag/in der Woche benutzt, bei welchen Temperaturen, welche Marke/Preisklasse usw.

Wie viele Daten man braucht, hängt von der Fragestellung ab. Wenn es nur wenige sind, besteht die Gefahr, dass man zufälligerweise in einer Enklave gelandet ist, in der sich alles anders verhält als in der Realität. Und in diese Enklave gerät man schnell, wenn man meint, gezielt Daten erheben zu müssen. Als generelles Fazit gilt: je unvoreingenommener und je mehr, desto besser – aber dadurch auch desto unmöglicher, diese Daten intellektuell und händisch zu beherrschen. D.h., man braucht Analysemethoden, die die Daten vorstrukturieren und so das Erkennen der Phänomene vorbereiten. Mithilfe geeigneter Methoden können aus vielen tausend Beispielen (Ereignissen) wenige hundert Gruppen werden. Viel mehr Gruppen sollten es nicht sein, sonst ist kaum etwas gewonnen gegenüber den unstrukturierten Daten; viel weniger sollten es auch nicht sein, da man sonst den Methoden mehr abverlangt, als sie zu leisten imstande sind: Sie können nicht die Phänomene direkt herausarbeiten, sondern nur Indizien darauf liefern. Aus diesen Indizien muss ein Mensch seine Schlüsse ziehen. Sind die Zielvorgaben für die Methoden zu rigoros, ist es durchaus möglich, dass Indizien nicht ermittelt werden können.

Mit Hilfe der Analysemethoden reduziert sich zwar die Größenordnung der Daten auf einen intellektuell beherrschbaren Bereich, es bleibt ohne methodische Unterstützung aber immer noch schwierig, alle Gruppen gleichberechtigt und vor dem Hintergrund des Gesamtbildes auszuwerten.

Wir beschreiben im Folgenden ein Vorgehensmodell für beliebige Anwendungen, die Aussagen über typische Phänomene in der Sprache herausarbeiten und diese z.B. in Form eines Informationsnetzes präsentieren wollen. Diese Netze sollen interaktiv dynamisch verschiedene Nutzer- und Nutzungsbedürfnisse befriedigen und bestehen aus Informationseinheiten, die im Wesentlichen aus rein automatisch gewonnenen Informationen aus Korpora aufgebaut werden.

Korpora stellen zwar nur einen Ausschnitt der Sprache dar und können nur bedingt 'repräsentativ' sein, sie enthalten ab einer bestimmten Größe aber so viel latentes Wissen über sprachliche Strukturen, dass die auffälligen und typischen sprachlichen Phänomene abgedeckt sind und aufgedeckt werden

können. Je größer Korpora sind, desto mehr seltene Phänomene werden ebenfalls erfasst. Insofern sind sehr große Korpora ('very large corpora'), so wie das DEUTSCHE REFERENZKORPUS des IDS (DEREKO 2005) mit über zwei Milliarden laufenden Textwörtern, die empirische Basis für linguistische Untersuchungen.

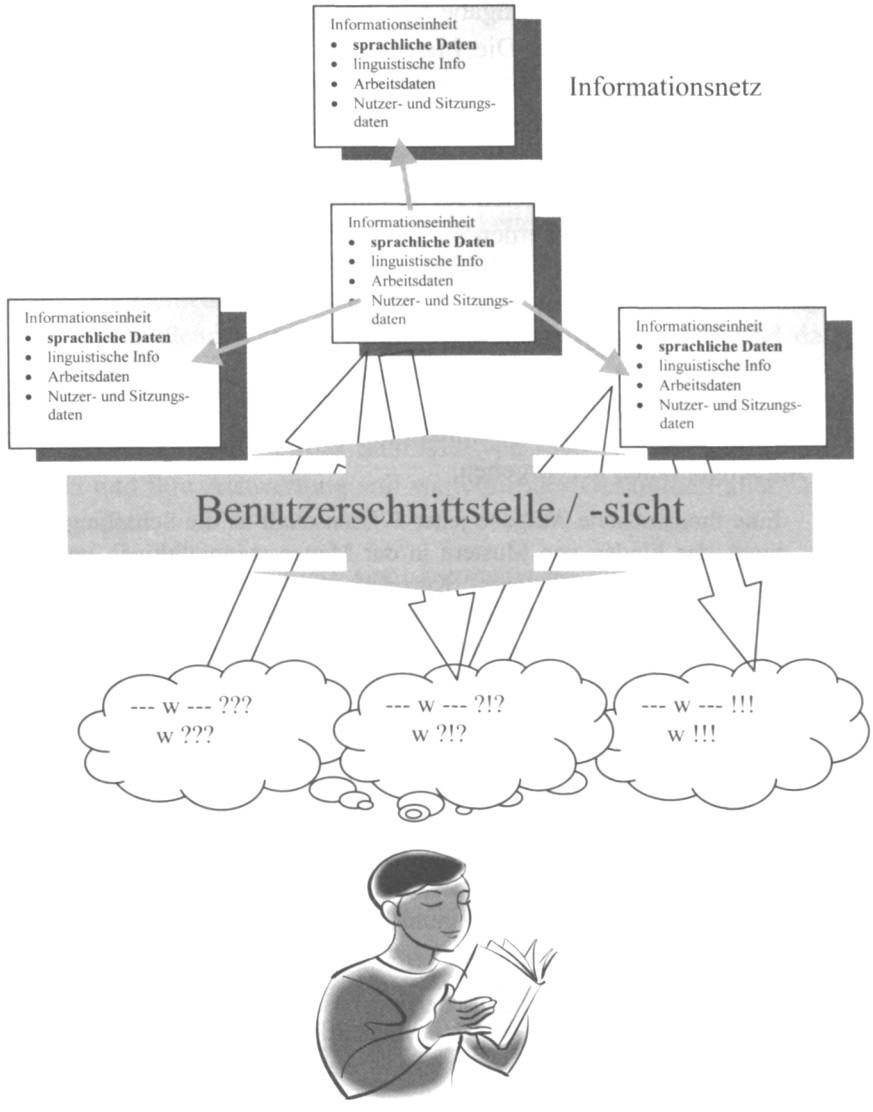

Abb. 1: Schematische Darstellung der Nutzung eines Informationsnetzes

Unser Hauptaugenmerk liegt auf den sprachlichen Daten als Gruppierungen von Originaltextausschnitten, so wie sie mathematische Verfahren (in unserem Fall die Kookkurrenzanalyse) auf der Grundlage von sehr großen Korpora berechnen. Die Verfahren ermöglichen Vorstrukturierungen des Untersuchungsgegenstandes, die ermittelten Gruppierungen sind eigenständige Objekte der Interpretation und weiterer Bearbeitungsschritte. Als Arbeitshypothese wollen wir die Angabe von linguistischen Informationen so minimal wie möglich halten. Die Information soll als Rechtfertigung dienen können, warum diese Informationseinheit eine relevante Informationseinheit darstellt. Unser zweites, hier nicht näher ausgeführtes Anliegen besteht in der Ausarbeitung des Konzepts der Relevanz, das letztendlich dazu führt, dass in bestimmten Situationen für bestimmte Nutzer bestimmte Informationseinheiten angeboten werden sollen bzw. auf welche Einheiten ausgewichen werden soll.

2. Korpuslinguistische Maxime

Korpuslinguistische Aussagen sind Aussagen über sprachliche Strukturen,[2] die aus der Analyse von massenhaftem, authentischem, möglichst originärem Sprachgebrauch hervorgehen.

> Eine fundamentale Aufgabe jeder Wissenschaft ist die Schaffung einer Ordnung, das Finden von Mustern in der Menge mannigfaltiger, unübersichtlicher Daten. Klassifikations-, Korrelations-, Mustererkennungs- und andere induktiv-heuristische Verfahren dienen hauptsächlich dem Zweck, neue, zuvor nicht bekannte Phänomene und Zusammenhänge zu entdecken, zumal wenn, wie in der Korpuslinguistik, die Daten wegen ihrer **schieren Masse** mit dem Intellekt **nicht** einmal **gesichtet** werden könnten. Tatsächlich beruhen viele Erkenntnisse auf empirischen Generalisierungen, die nachträglich deduktiv verankert [...] wurden. (Köhler 2005, S. 4f., Hervorhebung vom Autor).

> In a corpus-driven approach the commitment of the linguist is to the **integrity** of the **data as a whole**, and descriptions aim to be **comprehensive** with respect to corpus evidence. (Tognini-Bonelli 2001, S. 84, Hervorhebung vom Autor).

Die korpuslinguistische Herausforderung besteht darin, zwei Punkte miteinander in Einklang zu bringen, die auf den ersten Blick miteinander unvereinbar zu sein scheinen:

---

[2] D.h. Aussagen über sprachliche Einheiten, ihre Eigenschaften und Relationen.

1) die Datenmassen sollen möglichst umfangreich sein („More data is better data", vgl. Church/Mercer 1993), so umfangreich, dass sie „mit dem Intellekt nicht einmal gesichtet werden könnten" (Köhler 2005, S. 4),
2) trotzdem ist der Korpuslinguist bei seinen Untersuchungen und mit seinen Aussagen dem gesamten Datenbestand verpflichtet, seine Erkenntnisse sollen insbesondere vor dem Hintergrund der Gesamtheit aller Daten ihre Gültigkeit bewahren.

Um dieses scheinbare Dilemma aufzulösen, bedarf es entsprechender Methoden und Vorgehensmodelle. Gegenstand der Methoden ist es, mit Hilfe automatischer, mathematisch-statistischer Verfahren Strukturen in den Datenmassen zu ermitteln und so zu visualisieren, dass – je nach Perspektive der Fragestellung – entsprechende Auffälligkeiten hervortreten. Die in den berechneten Strukturen zum Ausdruck kommende Kohäsion oder Affinität kann ganz unterschiedlicher Natur sein – diese zu beschreiben ist Gegenstand der anschließenden Interpretation. Die Vorgehensmodelle sind dafür verantwortlich, dass der korpuslinguistische Anspruch auch nach dem Betrachten der Analyseergebnisse weiter aufrechterhalten wird: Auch das Weiterbearbeiten der Ergebnisse, das Sammeln von (evtl. vielen 'kleinen') Erkenntnissen und ihre Auswertung soll weiterhin der Gesamtheit aller Daten verpflichtet sein.

Ohne eine geeignete, zumindest semi-automatische Unterstützung ist dies kaum möglich. Wollte ein Linguist z.B. 100.000 Vorkommen eines Wortes untersuchen und bräuchte er nur sechs Sekunden für das Lesen eines Gebrauchskontextes, für das Einschätzen/Bewerten einer darin manifestierten Besonderheit und für das Notieren der Beobachtung, so müsste er 165 Stunden (knapp eine Woche) ununterbrochen arbeiten.[3] Bereits nach einigen wenigen hundert Beispielen wird der Überblick verloren gehen, welche Beobachtungen überhaupt und mit welcher Auffälligkeit zu verzeichnen waren – selbst ein gutes Buchhaltungssystem könnte nicht garantieren, dass die Gleichheit oder Ähnlichkeit von Beobachtungen erfasst wird, schlimmstenfalls bleiben als Ergebnis 100.000 nebeneinander stehende, unzusammenhängende Notizen. Idealerweise sollten die 100.000 Vorkommen aufgrund gleicher oder ähnlicher Musterhaftigkeit in wenigen hundert Gruppen zu-

[3] 100.000 Vorkommen eines Wortes sind bei unserer Datengrundlage ein durchaus realistischer Wert, die 6 Sekunden zur Bearbeitung sind vor allem auf Dauer sicher nicht einzuhalten.

sammengefasst werden (z.B. 250 Gruppen zu durchschnittlich je 400 Vorkommen), so dass annähernd die Übersicht über die Gruppen bewahrt bleiben, aber auch der Zusammenhalt innerhalb einer Gruppe schnell und trotzdem umfassend erkannt werden kann.[4]

3. Die einzelnen Komponenten des Vorgehensmodells

Das Vorgehensmodell ist grundsätzlich für die linguistische Interpretation sprachlicher Assoziationen konzipiert. Es wurde vor dem Hintergrund dieser Anwendung spezifiziert und innerhalb des linguistischen Paradigmas erforscht und entwickelt.

Für die Illustration der einzelnen Komponenten bzw. Schritte des Vorgehensmodells haben wir im Folgenden die arbiträre Domäne 'Descartes im öffentlichen Diskurs' gewählt, um den generischen Charakter des Modells hervorzuheben. Die inhaltlichen Aspekte des Modells sind hier eher zweitrangig und sollen nicht von dem eigentlichen Anliegen dieses Abschnitts, der Art des Modellierens, ablenken. Anhand der Erläuterungen sollte es leicht fallen, das Vorgehensmodell auf beliebige Domänen zu übertragen. So kann sich jeder Leser sein Forschungsgebiet wählen und die beschriebenen Schritte an einer konkreten Aufgabenstellung nachvollziehen.

3.1 Visualisierung

Unser Ansatz schließt an ein beliebiges Verfahren an, das aufgrund ähnlicher Eigenschaften mehrere tausend Ereignisse hierarchisch zu wenigen hundert Gruppen, im Folgenden Cluster genannt, zusammenfassen kann.[5] Diese Anzahl an Clustern ist vielleicht intellektuell beherrschbar, aber nicht so ohne weiteres im Überblick darstellbar. Alles gleichberechtigt nebeneinander präsentieren zu wollen, legt eine kreisförmige Darstellung nahe, die keine Rangfolge impliziert. Der Anspruch führt allerdings dazu, dass entweder

[4] 400 Gebrauchskontexte würden nach der Überschlagsrechnung immer noch ca. 40 Minuten beanspruchen, die in der Gruppe abgebildete Musterhaftigkeit führt aber dazu, dass das Charakteristische der Gruppe wesentlich schneller erkannt wird – und die 'Notizen' müssten nicht mehr zu 400 Einzelvorkommen, sondern nur noch einmal für diese Gruppe vermerkt werden.

[5] Die genauen Zahlen hängen von der Datengrundlage, dem betrachteten Phänomen und verschiedenen Einstellungen bestimmter Parameter bei der Analyse ab, sind aber durchaus realistische Werte für 'Standardanwendungen'.

eine so große Fläche beansprucht wird, dass die Übersichtlichkeit verloren geht, oder dass die normalerweise begrenzte, zur Verfügung stehende Fläche (Seitenformat oder Bildschirmgröße) gerecht aufgeteilt werden muss. Dies bedingt wiederum eine so starke Verkleinerung, dass letztendlich nichts zu erkennen ist. Von den drei konfligierenden Ansprüchen (1) alles (2) gleichberechtigt und (3) erkennbar darstellen zu wollen, eröffnet eine Abschwächung der letzten beiden Aspekte eine dynamische Lösung über eine interaktive Bildschirmdarstellung. Nur ein Ausschnitt wird deutlich abgebildet (der Fokus), der Rest (die Peripherie) wird nahezu unkenntlich angedeutet. In der folgenden Grafik wird ein Cluster als Kästchen dargestellt und mit der Eigenschaft beschriftet, mit der es sich von den anderen unterscheidet. Bei der Visualisierung wird das Prinzip gleich zweifach angewandt: zum einen radial (von innen nach außen), zum anderen zirkulär (kreisförmig).

In dem Beispiel wurde das Kontextverhalten des Bezugwortes 'Descartes' mathematisch analysiert und visualisiert. Die Beschriftung der Kästchen entspricht den Wortformen, die signifikant mit dem Bezugswort kookkurrieren, d.h. öfter im Kontext vorkommen als es rein zufällig sein dürfte. Der innere Ring zeigt die primären Partner, die besonders auffällig kookkurrieren und von dem Verfahren zuerst erkannt wurden. Weiter außen kommen zu den bisher dokumentierten Partnern weitere hinzu. Zusammen gelesen weisen diese auf komplexere kookkurrierende Wortkombinationen hin.

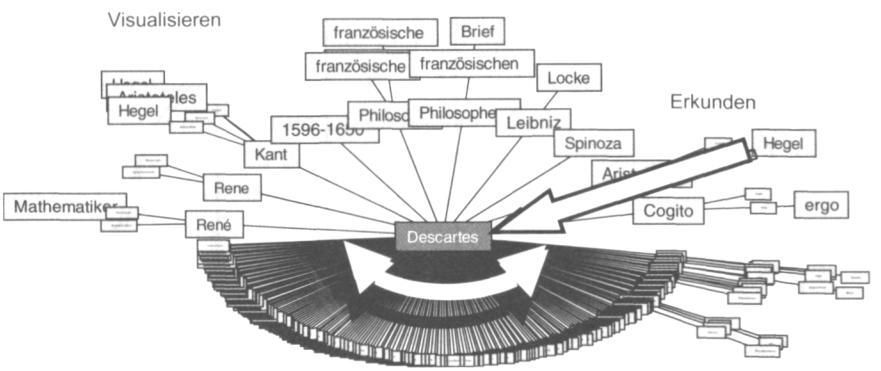

Abb. 2: Visualisierung einer berechneten Struktur (Pfeile deuten Interaktionsmöglichkeiten an)

Die radiale Fokussierung entspricht der hyperbolischer Ansätze (vgl. Lamping/Rao/Pirolli 1995): Das deutlich erkennbare Zentrum entspricht dem oberen Bereich, der äußere Ring dem unteren Rand einer Halbkugel. Der Ausgangspunkt der Clusterhierarchie wird zunächst in den Mittelpunkt gestellt, die primären Cluster auf dem ersten Ring, weitere Cluster auf weiteren Ringen um diesen herum. Die Verbindungen zwischen den Clustern stellen die hierarchischen Beziehungen dar; die Abstände zwischen den Ringen sind im Prinzip gleich groß, werden allerdings in den dreidimensionalen Raum der Halbkugel projiziert und wirken dadurch in der zweidimensionalen Abbildung gestaucht, so 'als ob man von oben auf die Halbkugel schauen würde' (vgl. Abb. 3).

Die zirkuläre Fokussierung wählt einen kleinen übersichtlichen Bereich (im Beispiel zehn primäre – initial die auffälligsten – Cluster), die sich sozusagen bevorzugt in der oberen Hälfte ausbreiten dürfen, die restlichen Cluster werden auf die untere Hälfte verteilt. Diese Art der Fokussierung ist zusätzlich notwendig, da ohne sie die Darstellung von mehreren hundert primären Clustern zu dem gleichen Erscheinungsbild führen würde wie in der unteren Hälfte. Welcher Ausschnitt der Grafik sich im Fokus befindet, kann hinsichtlich beider Aspekte vom Benutzer verändert werden. Elemente der äußeren Ringe können in das Zentrum gerückt werden, Elemente der unteren Hälfte sind nach Drehung in der oberen Hälfte deutlich erkennbar.

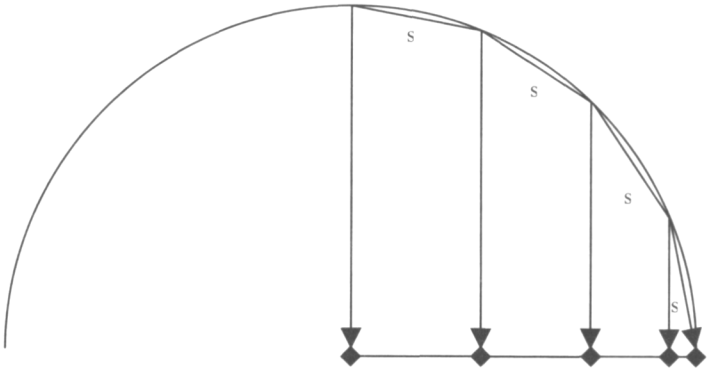

Abb. 3: Hyperbolische Projektion

Dadurch, dass sich ein Benutzer zu jedem Kästchen auch die Daten anzeigen lassen kann, die zu diesem Cluster geführt haben, ermöglicht diese Form der Darstellung über die Vorstrukturierung tatsächlich den Blick auf viele tausend Daten.

In unserem Beispiel sind die Daten die strukturierten Gebrauchskontexte des Wortes 'Descartes'. Folgendes Cluster ging aus der Analyse hervor, da die Zeichenfolge '1596-1650' signifikant mit dem Wort 'Descartes' kookkurriert. Es umfasst die Textstellen, in denen beide vorkommen, sowie die quantitativen Analyseergebnisse.

Partnertoken	Intervall	LLR	Anz.	Ratio	syntagmatische Muster
"1596-1650"	[1,5]	418	17	100%	Rene|René Descartes [... Philosoph] 1596-1650

KWICs

```
LIM us hatte Gassendis Zeitgenosse Rene Descartes ( 1596-1650 ) die Atomistik we
M98 thematikers und Naturforschers René Descartes (1596-1650). Wahrscheinlich we
V98  des französischen Philosophen Rene Descartes (1596-1650) versteigert worden
n91 Philosophen und Mathematikers Rene` Descartes (1596-1650) begab er sich in d
N92 tiker (geb. 1922). Todestage: Rene` Descartes, franz. Philosoph (1596-1650),
N93 schen Wissenschaftsbegriffes, Rene` Descartes (1596-1650), nach Leiden einge
N95 tiker (geb. 1950). Todestage: Rene' Descartes, franz. Philosoph (1596-1650);
N95 tiker (geb. 1950). Todestage: Rene` Descartes, franz. Philosoph (1596-1650);
N96 lliten "Luna-10". Geburtstage: Rene Descartes, frz. Philosoph und Naturwisse
N96 inister, 2.80 F. - 30. 3.: SM "Rene Descartes 1596-1650", 4.40 F. -
N00                             Rene` Descartes, frz. Philosoph (1596-1650);
BIH cartes im Windelband gelesen - René Descartes, latinisiert Renatus Cartesius
BIH tiker und Naturwissenschaftler René Descartes (Renatus Cartesius, 1596-1650)
BIH e Form des Philosophierens.  - René Descartes (1596-1650), frz. Philosoph,
Z99  "Ich denke, also bin ich." S+ René Descartes (1596-1650) +S
B03 ekannter Brief des Philosophen René Descartes (1596-1650) ist in der Staatsb
T05  den französischen Philosophen René Descartes (1596-1650) offiziell rehabili
```

Abb. 4: Inhalt eines Clusters

Auch wenn die äußeren Ringe und die untere Hälfte undeutlich abgebildet werden, können deren Elemente in den Fokus gerückt werden, sie können aber auch in den undeutlichen Bereichen einen besonderen Status zugewiesen bekommen, gemäß dem sie hervorgehoben darzustellen sind. Somit erfüllt die Überblicksdarstellung – bei all der Unschärfe vieler Bereiche – zwei sehr wichtige Funktionen: (1) Bei jedem Element wird dem Betrachter stets vergegenwärtigt, dass er sich nur in einem Ausschnitt des untersuchten 'Universums' befindet und dass (und evtl. auch wie) der Ausschnitt in den Rest eingebettet ist. (2) Hegt der Betrachter den Verdacht, dass er auf eine besondere Erscheinung gestoßen ist, kann er diese 'Indizien' hervorheben (auch im unscharfen Bereich) und somit eine Spur dieses Phänomens und evtl. seines Erkenntnisgewinns dokumentieren.

In unserem Beispiel fallen einige Cluster auf, die aufgrund von Jahresangaben (oder Bereichen) hervorgetreten sind und die untereinander einen engen zeitlichen Zusammenhang nahe legen. Vielleicht fallen dem Bearbeiter auch einfach nur vierstellige Ziffernfolgen auf, die er erst später weiter differenzieren kann.

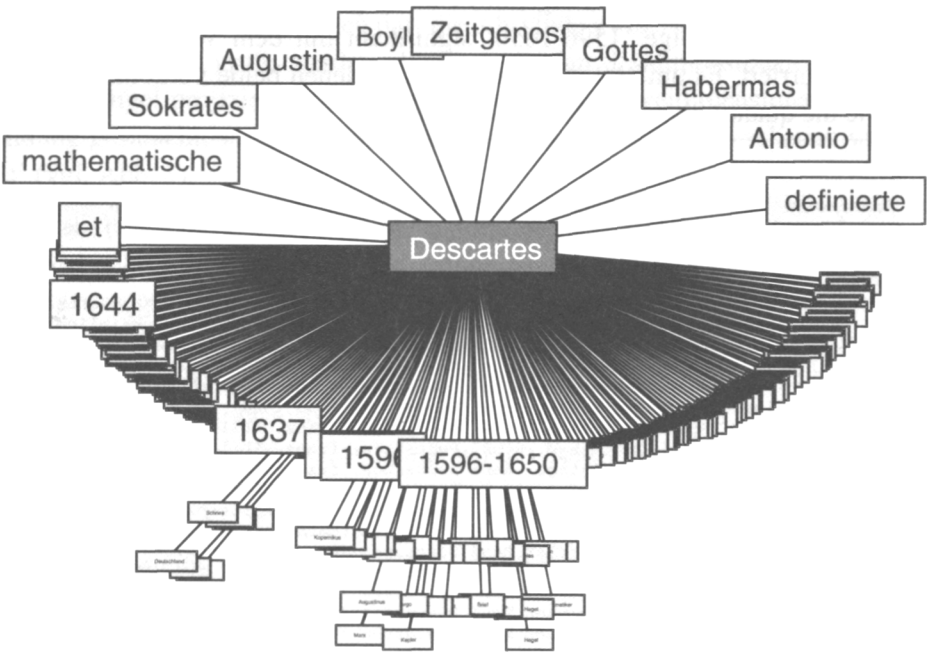

Abb. 5: Grafische Hervorhebung vermeintlich interessanter Cluster

3.2 Annotierung

Die Art der optischen Hervorhebung ließe sich über verschiedene Farbtöne und -abstufungen noch weiter verfeinern, um verschiedene Phänomene oder Grade der Zugehörigkeit auszudrücken. Der Wunsch, der sich dahinter verbirgt, entstammt dem Bedürfnis, die Elemente, d.h. die Cluster, mit Informationen anzureichern. Die direkte Art, dies umzusetzen, ist, für die verschiedenen Phänomene jeweils Attribute vorzusehen, die je nach Ausprägung mit Werten belegt werden. Typisierte Attribute erleichtern die Handhabung und ermöglichen bestimmte Operationen. So ist es z.B. möglich – alternativ zu der oben beschriebenen optischen Hervorhebung –, Attribute der Elemente mit entsprechenden Werten zu belegen.

Die in dem Beispiel erkannten, zeitlich zusammenhängenden Jahresangaben könnten z.B. als biografische Daten gekennzeichnet werden, entweder über eine Markierung dieser Eigenschaft (BIOGRAFISCHE_DATEN = 'ja') oder als Zugehörigkeit zu einer Kategorie oder Gruppe (GRUPPE = 'biografische Daten'). Die Annotierung führt dazu, dass der Clusterinhalt um diese Informationen erweitert wird.

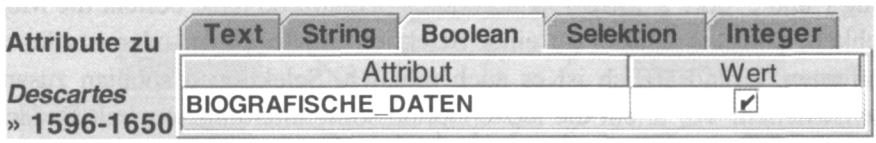

Abb. 6: Annotieren eines Clusters

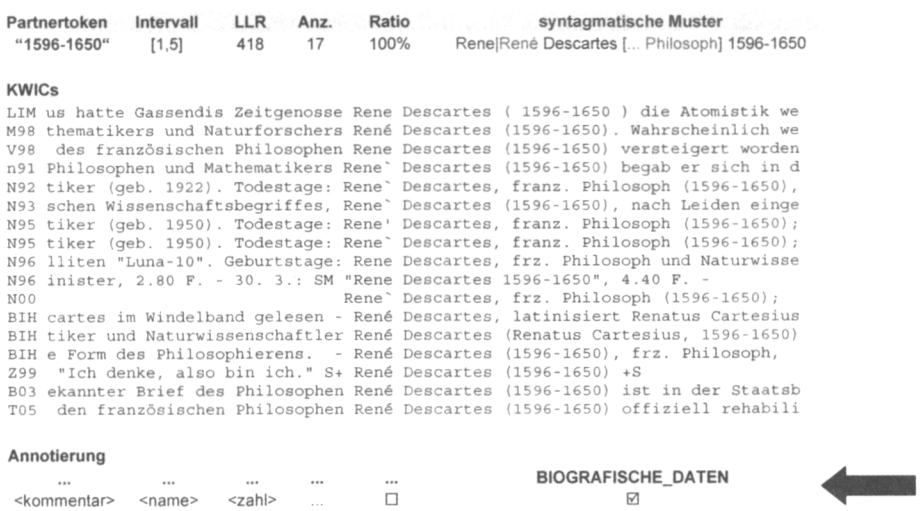

Abb. 7: Annotiertes Cluster

Geeignete Attributschemata (sog. Annotationsontologien) erleichtern die konsistente Anreicherung von Information, deren Aufbewahrung und Austausch, sowie das Sortieren annotierter Elemente gemäß bestimmter – oder die Suche nach bestimmten – Wertebelegungen. So bleibt die Information über die Sitzung hinaus erhalten und kann z.B. auch Kollegen zur Verfügung gestellt werden.

Das Hauptanliegen der Annotierung ist aber, vorrangig aufgrund der Wertebelegungen Selektionen von Clustern zusammenstellen zu können. Selektionen stellen einen flüchtigen Zwischenstand für Mengen von Clustern dar, die als zusammengehörig betrachtet werden können, weil in der Gesamtheit der Cluster ein Phänomen zum Ausdruck kommt. Da die Idee der Annotierung ist, sich diesen Phänomenen zu nähern, liegt es nahe, Selektionen im Wesentlichen über Ergebnisse von Suchanfragen oder deren logischer Verknüpfung ('und', 'oder', 'nicht') zu definieren. Ergänzend dazu besteht die Möglichkeit, einzelne Cluster zu einer Selektion hinzuzufügen oder aus ihr zu entfernen. Grundsätzlich ist es auch möglich, Selektionen spontan zusammenzustellen. Vor allem die letzte Option sollte allerdings nur mit Bedacht eingesetzt werden, da hierbei der dokumentarische Wert der Annotierung vollkommen in den Hintergrund getreten ist.

In dem Descartes-Beispiel kann etwa nach 'als biografisches Datum gekennzeichneten' Elementen gesucht werden. Die Treffer stellen eine Selektion dar und werden optisch hervorgehoben. Dieselbe Selektion hätte auch spontan entstehen oder aus einer anderen Selektion hervorgehen können, z.B. hätte der Bearbeiter

- alle Cluster einzeln zur Selektion hinzufügen,
- alle vierstelligen Ziffernfolgen um entsprechende Bereichsangaben ergänzen (logische 'oder'-Verknüpfung bzw. Vereinigungsmenge) oder
- alle vierstelligen Ziffernfolgen/-bereiche auf größer als 1596 und kleiner als 1650 einschränken (logische 'und'-Verknüpfung bzw. Schnittmenge) können.

3.3 Gruppierung

Der Beweggrund hinter der hier beschriebenen Vorgehensweise ist, zu erkundschaften und zu dokumentieren, inwieweit sich in den von der Analyse vorgeschlagenen Clustern – je Cluster oder cluster-übergreifend – sprachliche Phänomene manifestieren. Die Annotierung ist dabei der erste Schritt, es kann jedoch durchaus sein, dass der Zusammenhang zwischen den sprachlichen Phänomenen und der Annotierung vielfältiger Natur ist. Die Zusammenstellungen der Selektionen bilden die nächsten Schritte, bis ein Zustand erreicht ist, der dokumentiert und festgehalten werden soll. Dazu werden die

Selektionen in Gruppen überführt, die als eigenständige Objekte für die weitere Bearbeitung zur Verfügung stehen. Eine Gruppe kann wieder zu einer Selektion werden. Dadurch kann wieder mit den obigen Operationen weitergearbeitet werden, um Zusammenstellungen zu verfeinern und aus der verfeinerten Selektion wieder eine Gruppe zu definieren.

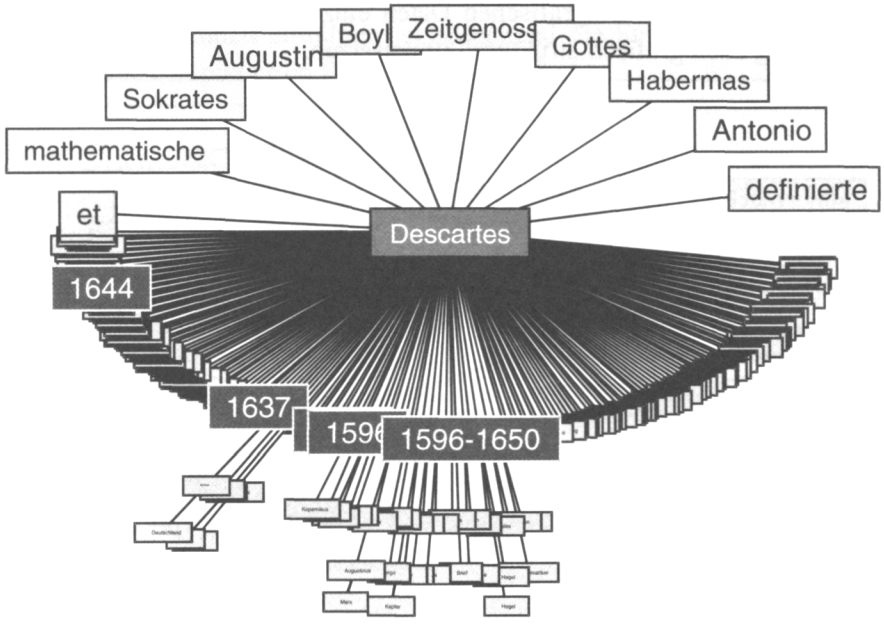

Abb. 8: Hervorhebung einer Selektion (in diesem Fall über eine Suche)

Mit dem Konzept einer Gruppe hat der Bearbeiter zum einen ein dokumentarisches Ziel erreicht, indem er aussagt, dass die in der Gruppe zusammengefassten Cluster irgendwie miteinander zusammenhängen. Zum anderen hat er damit vorbereitet, dass die gruppierten Daten Ausgangspunkt und Kern von Informationseinheiten werden können, die in entsprechenden Anwendungsszenarien bestimmte Informationsbedürfnisse stillen können. Diese Verbindung muss – in einem separaten, hier nicht erörterten Schritt – vom Bearbeiter als Abbildung ('Mapping') in Abhängigkeit von seiner Anwendung definiert werden.

Gruppen sind eigenständige Objekte, die virtuell bleiben können (und nur intern verwaltet werden), sie können aber auch grafisch umgesetzt werden.

Sind Cluster Elemente verschiedener Gruppen, kann die Umsetzung allerdings sehr unübersichtlich werden. Darüber hinaus kann die Motivation für Gruppierungen z.T. sehr anwendungsbezogen sein. Nur bei anwendungsneutralen und eindeutigen Gruppierungen bietet es sich tatsächlich an, diese auch als strukturelle Elemente darzustellen.

In dem Beispiel entscheidet sich der Bearbeiter dafür, eine neue Einheit für die Cluster der biografischen Daten einzuführen und diesen Zusammenhang auch grafisch-strukturell umzusetzen.

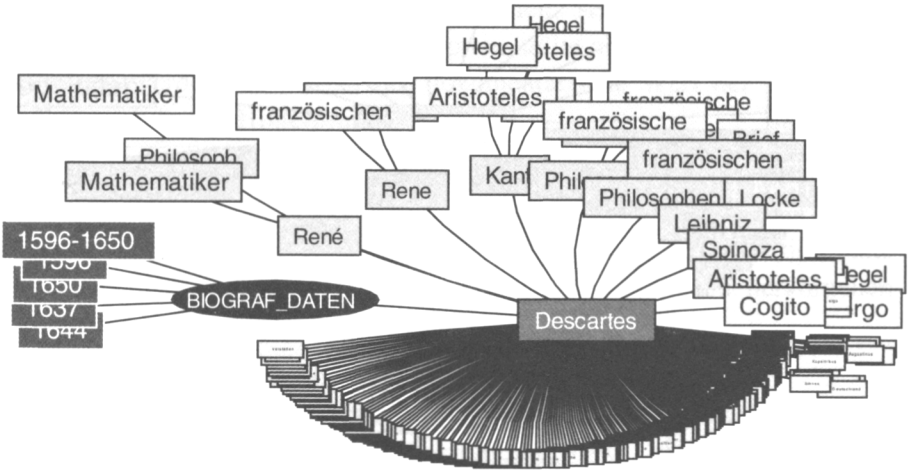

Abb. 9: Grafische Umsetzung einer Gruppierung

Die neu eingeführten Elemente sind im Gegensatz zu den berechneten Clustern epistemische Strukturierungselemente. Sie sollen helfen, das Modell zum Ausdruck zu bringen, das während der Analyse aufgebaut wird. Diese Elemente können auch spontan angelegt werden und Cluster bzw. künstliche Gruppen können mit ihnen verbunden werden, ohne dass Attribute vergeben, belegt und abgefragt werden müssten. Dies erspart, dass der Interpretierende etwas in Worte fassen muss, was er vielleicht noch gar nicht in Worte fassen kann. Der fehlende suggestiv-dokumentarische Charakter der Attribute kann annähernd durch eine adäquate Bezeichnung des künstlichen Elements ausgeglichen werden. Diese Bezeichnungen können aber im weiteren Verlauf durchaus revidiert werden.

Für das Beispiel sei angenommen, dass der Bearbeiter mehrere Cluster, in denen Zeitgenossen von Descartes genannt werden, spontan unter einem neuen künstlichen Element 'ZEITGENOSSEN' zusammengefasst hat. Dabei kann dem Interpretierenden durchaus aufgefallen sein, dass diese Einheit genau wie die vorherige im Grunde besondere Aspekte all dessen sind, was man beim Zusammentragen an Informationen zu einer Biografie sammeln würde. Daraus könnte er ein Revision seines Modells ableiten, eine neue Einheit einführen, entsprechend die Struktur umgestalten und den sehr komplizierten Bezeichner vereinfachen.

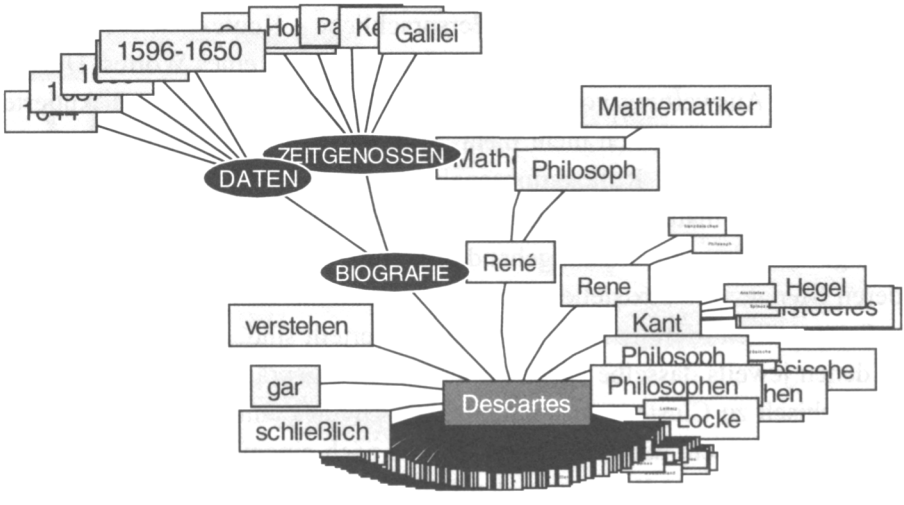

Abb. 10: Verfeinerung des Modells

Alle hier beschriebenen Möglichkeiten unterstützen eine Vorgehensweise, um aus vorstrukturierten Daten Schritt für Schritt über eine Interpretation angereicherte und (leicht) modifizierte Strukturen entstehen zu lassen, die für sich genommen bereits in gewisser Weise neue Erkenntnisse zum Ausdruck bringen bzw. in verschiedenen Szenarien diese Erkenntnisse vermitteln (helfen) können. Grundlage dieser Strukturen ist wohlgemerkt nach wie vor der gesamte Ausgangsdatenbestand, dem sich der Ansatz stets verpflichtet fühlt. Die Strukturen stellen lediglich eine Zwischenschicht dar, die die Wahrnehmung der in den Daten vorhandenen Muster unterstützen sollen.

4. Das Vorgehensmodell angewandt auf die Kookkurrenzanalyse

You shall know a word by the company it keeps.
(Firth 1968, S. 179)

Ausgangspunkt für die Anwendung unseres Vorgehensmodells sind die Ergebnisse der Kookkurrenzanalyse (vgl. Belica 1995). Die Kookkurrenzanalyse fasst Originaltextausschnitte zu hierarchischen Gruppen, den bereits vorher erwähnten Clustern, zusammen, z.B. aufgrund eines gleichen bzw. ähnlichen Kontextverhaltens eines zentralen Trefferobjekts. Die Cluster sind die zentralen Elemente der weiteren Bearbeitungsschritte.

Der sprachliche – aber auch der außersprachliche – Kontext ist das wichtigste Charakteristikum zur Beschreibung des Gebrauchs eines Wortes. Die systematische Auswertung von Gebrauchskontexten ist in der hier diskutierten Größenordnung händisch nicht mehr zu bewerkstelligen. Die Kookkurrenzanalyse (oder auch Kollokationsanalyse, vgl. engl. „collocation(al) analysis") ist ein mathematisch-statistisches Verfahren, das eine vorgegebene Menge von Textstellen dahingehend überprüft, ob im Vergleich zu einer anderen Menge von Textstellen die darin vorkommenden textuellen und/oder außertextuellen Merkmale auffällig anders verteilt sind. Alle die Textstellen, in denen jeweils dasselbe Merkmal auffällig ausgeprägt ist, lassen sich evtl. hierarchisch zu Clustern zusammenfassen. Die bekannteste Anwendung des Verfahrens besteht darin, in einem Korpus nach einem Wort zu suchen und einen (sprachlichen) Kontext zu den Fundstellen zu definieren. Diese vorgegebene Menge von Textstellen wird dann daraufhin überprüft, ob die darin vorkommenden Wörter auffällig oft vorkommen, d.h. im Vergleich zum Vorkommen im Gesamtkorpus statistisch überrepräsentiert sind.

Falls dem so ist, liegt nahe zu vermuten, dass das Zusammenstellungskriterium der vorgegebenen Menge mit der Auffälligkeit zusammenhängt; in dem Fall, dass nach einem Wort gesucht wurde (d.h., dass in allen Textstellen dieses Wort vorkommt) und ein anderes Wort auffällig oft in dessen Nähe beobachtet werden kann, besteht der Verdacht, dass dieser Umstand in einer 'Anziehungskraft' zwischen den Wörtern begründet ist. Welcher Natur diese Affinität ist, darüber sagt die Analyse nichts, sie erschließt sich aber

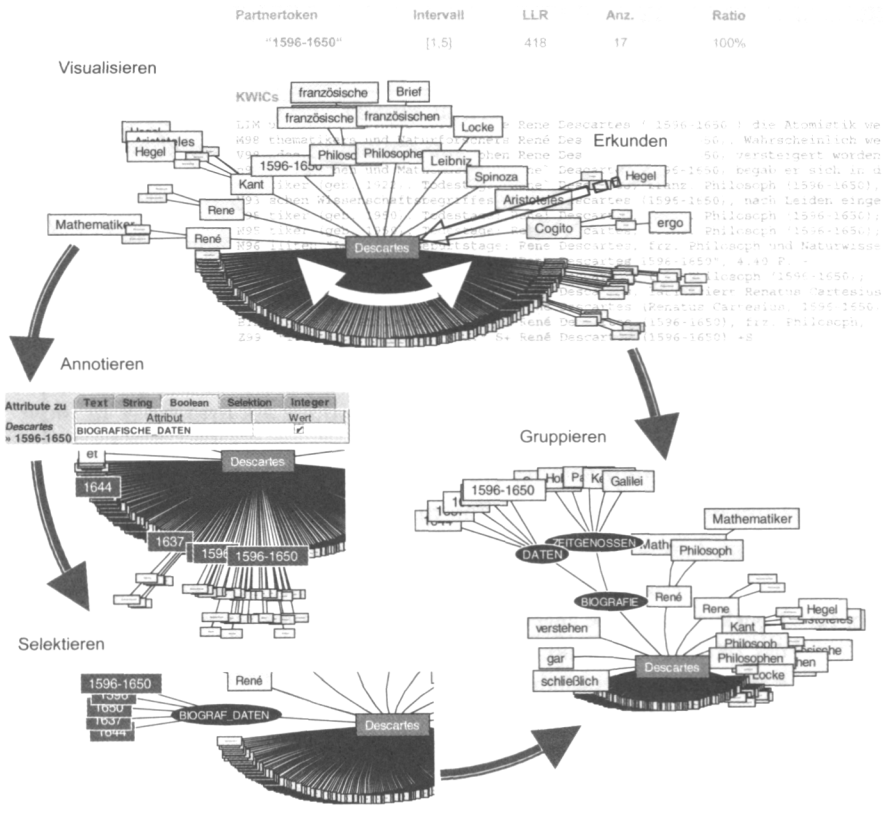

Abb. 11: Das Modell im Überblick

normalerweise bei der Betrachtung der Cluster, in denen jeweils die Textstellen[6] zusammengefasst sind, die dieselben 'auffälligen' Wörter enthalten. Als zusätzliche Interpretationshilfe wird zu jedem Cluster das 'auffällige Wort' ('Partnerwort') vermerkt und ein syntagmatisches Muster angegeben, das innerhalb eines Clusters die prägnanteste Wortstellung und -form der Partnerwörter sowie weiterer prägnanter Füllwörter ermittelt hat. Kookkurrenzcluster können hierarchisch strukturiert sein, wenn zu auffälligen Wort-

[6] Aus urheberrechtlichen Gründen müssen die Textausschnitte auf eine kleine Anzahl von Absätzen beschränkt sein, aus technischen Gründen beschränken wir uns zunächst auf KWIC-Zeilen (= **K**ey**w**ord **i**n **C**ontext; eine Zeile links- und rechtsseitiger Kontext, symmetrisch um das zentrale Trefferobjekt, auf 80 Zeichen insgesamt begrenzt).

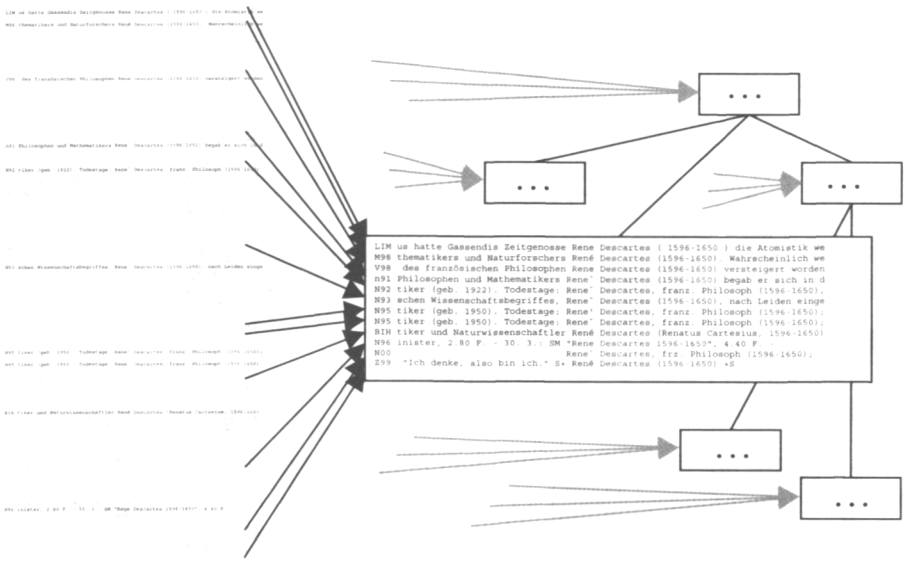

Abb. 12: Kookkurrenzanalyse angewandt auf das Ergebnis der Suchanfrage „Descartes“ an das DEUTSCHE REFERENZKORPUS des IDS und ausgewertet in Bezug auf das Korpus als Grundgesamtheit

kombinationen noch weitere auffällige Wörter hinzutreten, so dass auch auffällige Kombinationen von mehr als zwei Wörtern erfasst werden. Der erste ermittelte Partner, der ‘primäre Partner’ ist der statistisch Auffälligere – unabhängig von der tatsächlichen Wortfolge in den ausgewerteten Textstellen.

Die Kookkurrenzanalyse wird z.Zt. interaktiv über das System COSMAS II (COSMAS II 2005) zur Erschließung der IDS-Korpora angeboten, in einer Kookkurrenzdatenbank (CCDB) stehen die Ergebnisse einer Menge von Kookkurrenzanalysen zur Verfügung.[7] Eine Schnittstelle unseres Ansatzes zu der in COSMAS II implementierten Kookkurrenzanalyse und zur CCDB ist spezifiziert. Sofern nichts anderes angemerkt wird, gelten im Folgenden die Parametereinstellungen, wie sie für die Analysen der CCDB gewählt wurden.

[7] `http://corpora.ids-mannheim.de/ccdb/`, ca. 150.000 lexikografisch interessante Lemmata; Parametereinstellungen: Kontext: -5 ... +5, Satzgrenzen: beachten, Zuverlässigkeit: mittel, Granularität: mittel, Funktionswörter: ignorieren, Autofocus: ja, Zuordnung: mehrfach.

Der eigentliche Inhalt eines Clusters, eine Menge von authentischen Textstellen, kann bezüglich verschiedener Facetten oberflächlich betrachtet sehr inhomogen sein. Dies betrifft zum einen die bibliografischen bzw. Metainformationen (Region, Zeit, Thema usw.) – was durchaus erwünscht sein kann.[8] Zum anderen sind die Quellen der Variabilität

- die konkrete Wortform des Bezugswortes (sofern lemmatisiert oder über Platzhalter gesucht),
- die konkrete Wortform des Partnerwortes (sofern lemmatisiert),
- die Wortstellung des Bezugswortes und der Partnerwörter und
- der restliche nicht als Bezugs- oder Partnerwort erfasste Kontext – sowohl in Wortform als auch -stellung.

Unsere These ist, dass ein Cluster über die Berechnung gerade deshalb zustande gekommen ist, weil alle in ihm enthaltenen Textstellen ein spezielles sprachliches Phänomen zum Ausdruck bringen. In diesem Sinne gehen wir von einer sprachsystemischen Homogenität der Cluster aus.

Die ersten Arbeitsschritte sind dergestalt konzipiert, dass das Informationsnetz ausgehend von den berechneten Clustern als Grundlage aufgebaut werden kann. Die Cluster oder ggf. Mengen von Clustern stellen dabei (Vorschläge für) die sprachlichen Daten der Informationseinheiten dar. In einer ersten Phase soll – vor dem Hintergrund des Ergebnisses einer Kookkurrenzanalyse als Gesamtheit – die Möglichkeit geboten werden, die berechneten Cluster bzw. -strukturen zu verstehen. Insbesondere soll bei der Erkundung der Clusterstrukturen reflektiert werden, welche sprachlichen Phänomene sich in einzelnen Clustern oder clusterübergreifend manifestieren.[9] Dazu werden ständig alle Cluster überblicksartig dargestellt. Ein interaktiv wählbarer Ausschnitt wird detaillierter angezeigt.

---

[8] Auffälligkeiten bei der statistischen Verteilung innerhalb dieser Facetten sollen in einer zukünftigen Version der Kookkurrenzanalyse automatisch überprüft werden.

[9] Sollte in einem Cluster kein sprachliches (sondern vielleicht ein kulturelles) Phänomen dokumentiert sein, ist diese Erkenntnis ggf. auch relevant dafür, ob bzw. dass dieses Cluster in bestimmten Zusammenhängen ignoriert werden soll. Dass ein Cluster in disjunkten Teilmengen mehrere sprachliche Phänomene dokumentiert, ohne dass die Analyse Subcluster ermittelt hat, wird für das Vorgehensmodell zunächst zurückgestellt. Derartige Fälle lassen sich gemeinhin durch eine erneute Analyse mit einer feineren Granularität handhaben. Rückkopplungen dieser Art werden in diesem Beitrag nicht betrachtet.

Zu allen Clustern lassen sich die Originaltextausschnitte einblenden, sowie die quantitativen Angaben, die bei der Kookkurrenzanalyse dazu geführt haben, diese Menge von Originaltextausschnitten als ein Cluster zu betrachten, und weitere Angaben (Partnerwort, Kohäsionsstärke (LLR) des primären Partners,[10] Anzahl, Intervall, syntagmatische Muster). Die überblicksartige Darstellung ist deshalb sinnvoll, da ein Cluster ohne Ko-Cluster genauso wenig aussagekräftig ist wie ein Wort ohne Kontext: nachdem ein Cluster interpretiert wurde, kann dies – bei Verdacht – durch Hervorhebung bewusst gehalten werden, um Beziehungen zu anderen Clustern zu hinterfragen, z.B. auf morphosyntaktische oder semantische Verwandtschaft. Drückt dieses Cluster mit Clustern anderer Wortbildungsformen oder von Synonymen/ Kohyponymen dasselbe sprachliche Phänomen aus und sind die Cluster Kandidaten für eine spätere 'Zusammenfassung'? Bei Verdacht ist es auch möglich, in der Darstellung gezielt nach Varianten zu suchen, was jedoch die allumfassende Abhandlung aller Cluster nicht ersetzen sollte. Es kann auch über Attribute nach Clustern gesucht werden, die sich aus der Berechnung (oder später aus der Interpretation) ergeben haben. Die Darstellung kann nach unterschiedlichen Kriterien sortiert werden und es können mehrere Kookkurrenzanalyse-Ergebnisse parallel 'erkundschaftet' und verglichen werden.

In der zweiten Phase geht es vorrangig darum, die Cluster mit Annotationen anzureichern, die den Erkenntnisgewinn widerspiegeln sollen, die der Betrachter bei der Interpretation der Clusterstrukturen gewonnen hat, d.h. insbesondere welche sprachlichen Phänomene sich in welchen Clustern ausdrücken. Die Annotationen bestehen nicht aus nur einem Informations- oder Kommentarfeld, sondern sind selbst wiederum – gemäß einer Annotationsontologie – strukturierte Objekte. D.h., dass bei jedem Cluster mehrere Attribute zur Verfügung stehen, die jeweils mit Werten belegt werden können. Da unterschiedliche Angaben verschiedenen Anforderungen genügen sollen (z.B. um sie später für unterschiedliche Zwecke nutzen zu können), sind die Attribute typisiert. Neben Zeichenketten und Texten können Attribute Zahlen, Wahrheitswerte und Aufzählungen ausdrücken. Die Typen der Attribute ermöglichen eine kontrollierte Werteingabe und sie eröffnen besondere Möglichkeiten sowohl bei der Suche als auch bei der Sortierung.

---

[10] LLR = *Log-Likelihood Ratio* (vgl. Dunning 1993), ein bewährtes Maß für die Berechnung der Nicht-Zufälligkeit des Auftretens eines Wortes, z.B. als primäres Partnerwort.

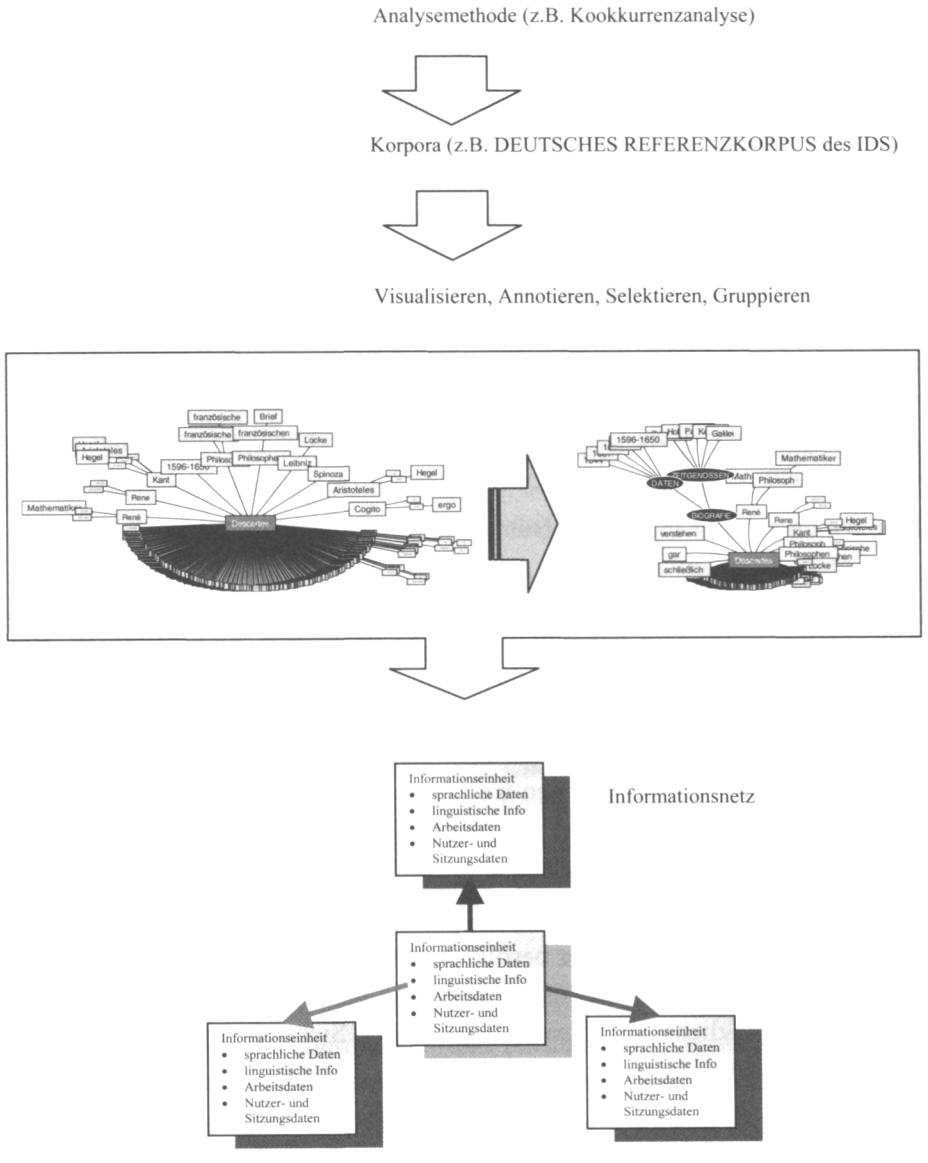

Abb. 13: Überblick über das Vorgehensmodell

Ziel dieser Arbeitsschritte ist es, Teile der Analyse (gedanklich und grafisch) hervorheben bzw. ausblenden zu können. So ist z.B. häufig die erste Erkenntnis, die man beim Betrachten der Cluster gewinnt, dass in einzelnen Clustern keine sprachlichen Phänomene, sondern Phänomene des Tages- und Zeitgeschehens ausgedrückt werden. Beispiele dafür sind Cluster aufgrund von Filmtiteln oder der Sportberichterstattung. In diesen, aber auch in anderen Fällen bewirken Eigennamen häufig überlagernde Effekte. Sollte sich herausstellen, dass die genaue Beschreibung für eine spätere Betrachtung von besonderem Wert ist, bleibt es dem Interpretierenden überlassen, entsprechende Attribute für diese Cluster zu vergeben. Aus unserer Sicht würde es angemessen sein, sie zunächst als 'Noise' ('Verunreinigung') zu deklarieren, die für weitere Schritte ausgeblendet werden kann. Zu einem späteren Zeitpunkt können diese Cluster ggf. wieder eingeblendet und weiter untersucht werden.

Des Weiteren stellt man beim Betrachten der Cluster fest, dass die Analyse den Sprachgebrauch zu filigran aufgespaltet und eventuell zu viele Cluster erstellt hat. In manchen Clustern kommt – nicht nur aus Sicht bestimmter Anwendungen, sondern generell – dasselbe Phänomen zum Ausdruck. Bei der Analyse zu „stattfinden" ergeben sich Cluster zu verschiedenen Zeitangaben (Monatsnamen, Tagesnamen). Die Analyse zu „Grund" liefert viele Cluster mit Farbbezeichnern, aber auch mit Flächenangaben, zwischen denen genauso wenig unterschieden werden müsste. In diesen Fällen böte sich an, zu allen diesen Clustern das generische Konzept zu annotieren, so dass es möglich ist, sie entsprechend zusammenzufassen.

Darüber hinaus treten aber auch Phänomene hervor, deren Annotierung sich nur vor dem Hintergrund einer besonderen Anwendung rechtfertigt. So können die Grenzen zwischen Clustern durchaus von Bedeutung sein, wenn es darum geht, mögliche Interferenzprobleme aus Sicht einer Kontrastsprache zu kennzeichnen.

Unser Ansatz ist konzipiert für zwei verschiedene Nutzergruppen, die jeweils in einem eigenen Modus arbeiten. Eine Gruppe sind aus unserer Sicht Experten, die Annotationsontologien aufbauen bzw. erweitern. Dieser Prozess erfordert z.T. sehr unterschiedliche Kompetenzen. Zum einen müssen die Experten beurteilen können, welche Relevanz welches Phänomen vor dem Hintergrund einer bestimmten Anwendung hat bzw. welche Phänomene einen anwendungsunabhängigen Stellenwert haben. Zum anderen sollten die

Experten dazu in der Lage sein einzuordnen, inwieweit aus den Daten gewonnene Erkenntnisse mit den Erwartungen bestehender Theorie harmonieren, Erwartungen der Theorien unerfüllt bleiben oder nicht-erwartbare Phänomene hervorgetreten sind. Experten können Attribute mit verschiedenen 'Typen' ('Wertebereichen') deklarieren. Die andere Gruppe sind aus unserer Sicht Anwender, die evtl. aus vorgegebenen Annotationsontologien auswählen und diese benutzen können. Anwender und Experten können für die Attribute jedes Clusters Werte vergeben.

Die Attributbelegung kann u.a. auch die Funktion der Ablaufsteuerung und der Koordinierung der weiteren evtl. kooperativen Arbeit übernehmen. Ein wesentliches Anliegen dieser Phase besteht aber darin, ein- oder mehrelementige Mengen von Clustern definieren zu können, die als sprachliche Daten einer Informationseinheit fungieren sollen. Die Zusammenstellung dieser Mengen kann über Attribute (bzw. Suche nach Attributwerten) erfolgen oder durch explizites Zu- oder Abwählen einzelner Cluster.

Während bei der ersten Variante der Schritt quasi durch die Attributwerte dokumentiert wird, fehlt dieser Aspekt bei der zweiten Variante. Die Clustermengen können benannt und 'editiert' werden (hervorgehoben und verändert). Aus diesen Clustermengen werden die Informationsnetze aufgebaut. Den Anspruch, der aus Sicht des Informationsnetzes an eine Clustermenge gestellt wird, lässt sich so umschreiben, dass diese Menge ein Informationsbedürfnis besser stillen kann als seine Elemente (da diese zu differenziert sind). Die Attribute dienen dazu, in der Anwendung die Relevanz der Menge zu beschreiben, d.h., dass und wie aus ihnen abgeleitet werden können soll, warum in einer bestimmten Anwendungssituation für einen bestimmten Nutzer gerade diese Einheit präsentiert wird.

Im letzten Schritt ist mit Hilfe eines zu definierenden Mappings die gesammelte Information (Clustermengen und Attributbelegungen) auf das (Ziel-) Informationsnetz abzubilden. Das Mapping hängt einerseits von den vereinbarten Attributen (der 'Annotationsontologie'), andererseits von der Zielstruktur (der 'Anwendungsontologie') ab. Wir gehen davon aus, dass sich auf beiden Seiten eine überschaubare Anzahl an Ontologien herauskristallisieren wird, die weitestgehend stabil sind. Im Idealfall müssen zukünftige Endnutzer nur zwischen verschiedenen vorgegebenen Ontologien und Mappings auswählen, um ihr System aufzubauen.

5. Fazit

In diesem Beitrag haben wir gezeigt, dass und wie es möglich ist, sich bei linguistischen Untersuchungen den Evidenzen, die sich aus der Analyse sehr großer Datenmengen ergeben, zu verpflichten. Mit der entsprechenden Unterstützung ist es sogar möglich, gleichzeitig eng an den Daten zu arbeiten und dabei den Überblick über die Gesamtheit zu behalten. An vielen Stellen ist bereits dokumentiert, wie die Kookkurrenzanalyse sinnvoll eingesetzt (Belica/Steyer i.Dr.) und dass sie erfolgreich angewendet wird (Hanks 2004, Steyer 2004, Steyer/Vachková i.Dr.). Die Einbettung in das hier vorgestellte Vorgehensmodell eröffnet bereits für diese Anwendungen neue Perspektiven. Wir erwarten aber gerade für den Bereich der Fremdsprachendidaktik eine große Resonanz, da dort die Schüler „in authentischen Texten zunehmend selbstständig die Bedeutung unbekannter Wörter aus dem Kontext erschließen" sollen (Mukherjee 2002, S. 67) – und gerade diese Kompetenz lässt sich durch die komprimierten und adäquat aufbereiteten Sammlungen von Kontexten besonders gut trainieren. Und wir spekulieren darauf, dass dies nicht nur für die Bedeutung eines Wortes gilt, sondern für alle Verwendungsaspekte von Wörtern oder abstrakteren Kategorien.

Die in diesem Beitrag – der Illustration halber – bewusst einfach gehaltenen Beispiele sollen nicht darüber hinwegtäuschen, welche Bandbreite neuer Erkenntnisse diese Vorgehensweise ermöglicht. Dies wird im folgenden Beitrag von Kathrin Steyer und Meike Lauer („Corpus-Driven": Linguistische Interpretation von Kookkurrenzbeziehungen) angedeutet, in dem einzelne Schritte dieser abstrakten, generischen Vorgehensweise bei der Untersuchung sprachlicher Phänomene konkret umgesetzt werden.

6. Literatur

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. COSMAS-Korpusanalysemodul. Mannheim 1995. Internet: `http://corpora.ids-mannheim.de` (Stand: November 2006).

Belica, Cyril/Steyer, Kathrin (i.Dr.): Korpusanalytische Zugänge zu sprachlichem Usus. In: AUC (Acta Universitatis Carolinae), Germanistica Pragensia XX, Praha.

Church, Kenneth W./Mercer, Robert L. (1993): Introduction to the Special Issue on Computational Linguistics Using Large Corpora. In: Computational Linguistics 19, 1, S. 1-24.

COSMAS II (2005): Corpus Search, Management and Analysis System. Internet: http://www.ids-mannheim.de/cosmas2/ (Stand: November 2006).

DEREKO (2005): Das Deutsche Referenzkorpus – DEREKO. Internet: http://www.ids-mannheim.de/kl/projekte/dereko/ (Stand: November 2006).

Dunning, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational Linguistics, 19, 1, S. 61-74.

Firth, John R. (1968): A Synopsis of Linguistic Theory 1930-1955. In: Palmer, Frank (Hg.): Selected Papers of J. R. Firth. Harlow. S. 168-205. [Zuerst ersch. in: Studies in Linguistic Analysis. Philological Society. Oxford: 1957.].

Hanks, Patrick (2004): The Syntagmatics of Metaphor and Idiom. In: International Journal of Lexicography 17, 3, S. 245-274.

Köhler, Reinhard (2005): Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In: LDV Forum 20, H. 2, S. 1-16.

Lamping, John/Rao, Ramana/Pirolli, Peter (1995): A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Denver, May 1995, S. 401-408.

Mukherjee, Joybrato (2002): Korpuslinguistik und Englischunterricht: Eine Einführung. Frankfurt a.M.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hg.): Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York. S. 87-116.

Steyer, Kathrin/Vachková, Marie (i.Dr.): Kookkurrenzanalyse kontrastiv. Zum Nutzen von Korpusanalysemethoden für die bilinguale lexikographische Praxis am Beispiel des GDTAW. In: AUC (Acta Universitatis Carolinae), Germanistica Pragensia XX, Praha.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam/Philadelphia. (= Studies in Corpus Linguistics 6).