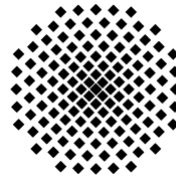


Transferbereich

**Automatische
Exzerption**



Langenscheidt



**Universität Stuttgart
IMS – Computerlinguistik**

Der Computer hilft Wörterbücher schreiben: Ein neues Projekt zu Methoden der Wörterbuch-Entwicklung

Software für Wörterbuchredakteure

Sprache lebt – und so sind Wörterbuchmacher ständig damit beschäftigt, ihr Wörterbuch auf einem aktuellen Stand zu halten und zu vervollständigen. Um neue Wörter, Bedeutungen und Wortkombinationen zu finden, exzerpieren Wörterbuchredakteure regelmäßig Zeitungen, Zeitschriften und Bücher. Das heißt: sie notieren alles, was ihnen für ihr Wörterbuch wichtig erscheint, insbesondere Beispiele.

Ziel des vorgestellten Projekts ist es, in den kommenden zwei Jahren so weitgehend wie möglich die Exzerption durch Softwarewerkzeuge zu unterstützen. Das neue Transferprojekt ist eine Kooperation zwischen der Universität Stuttgart, dem Langenscheidt Verlag (München) und dem Bibliographischen Institut & F.A. Brockhaus (Mannheim). Die Wissenschaftler des Stuttgarter Instituts für maschinelle Sprachverarbeitung haben in den letzten Jahren Computerprogramme entwickelt, die große Textmengen lesen und nach bestimmten grammatischen oder lexikalischen Phänomenen sortieren können. Verarbeitet werden dabei viele Millionen Wörter, in der Regel mehrere komplette Jahrgänge einer Tageszeitung auf einmal. Die Software ist aber nicht nur für die Wörterbucherstellung relevant; jedes System zur Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) benötigt sie als Komponente.

Die NLP-Software soll jetzt in den Dienst der Wörterbuchredaktionen gestellt werden: Die Stuttgarter Wissenschaftler lassen den Computer Beispielmateriale suchen und nach wörterbuchrelevanten Kriterien sortieren. Zur Unterstützung der Wörterbuchredakteure werden die gefundenen Beispiele mit den entsprechenden Angaben in dem zu überarbeitenden Wörterbuch verglichen, das dem Institut in elektronischer Fassung zur Verfügung steht. So kann automatisch und damit sehr schnell ermittelt werden, welche Funde die Aufnahme ins Wörterbuch lohnen. Umgekehrt können, u.a. auf-

grund von Häufigkeitsanalysen, Vorschläge zur Löschung seltener oder veralteter Wendungen gemacht werden. Die Programme werden am IMS an einigen Wörterbüchern von Langenscheidt und von Duden erprobt und dabei den Bedürfnissen der Wörterbuchredaktionen angepasst.

Projektpartner: zwei Verlage und ein Universitätsinstitut

Das Vorhaben dient dem Know-how-Transfer von der Forschung in die Wirtschaft, daher die Bezeichnung „Transferbereich“. Das Wörterbuchprojekt ist der erste DFG-Transferbereich in den Geisteswissenschaften: Techniken aus der Computerlinguistik, die seit Jahren in einem Sonderforschungsbereich der Universitäten Stuttgart und Tübingen zu den „Theoretischen Grundlagen für die Computerlinguistik“ erforscht und entwickelt worden sind, werden jetzt im konkreten Arbeitsumfeld getestet.

Allein die Personalkosten für das Projekt belaufen sich auf rund 500.000 Euro. Die Universität wird von der Deutschen Forschungsgemeinschaft, DFG, gefördert. Die Verlage kommen für ihre Kosten selbst auf.

Software für alle Wörterbuchbenutzer – auf dem Weg zum innovativen elektronischen Wörterbuch

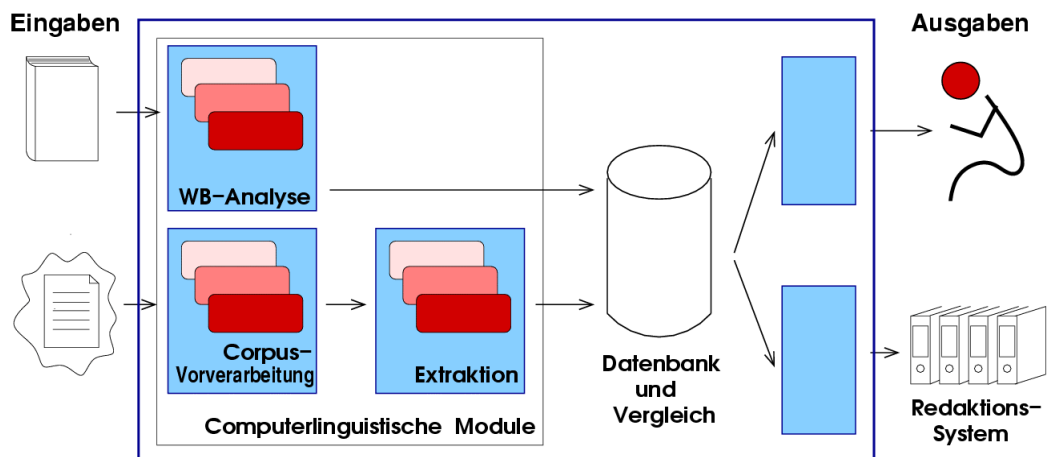
Ergebnis des Projekts soll zunächst Software sein, die dem Wörterbuchredakteur Routinearbeit abnimmt und ihm umfangreiches Material für Aktualisierungen zur Verfügung stellt. Dabei geht es nicht nur um neue Einzelwörter, wie etwa das *Kickboard*, das (im Gegensatz zu seinem in der Schweiz verbreiteten Vetter, dem *Trottinett*) Eingang in die Ausgabe 2001 des *Duden Universalwörterbuch* gefunden hat. Es geht auch um Wortverbindungen und um Details zum Wortgebrauch: Hat es sich eingebürgert, *die* E-Mail zu sagen, oder ist *das* E-Mail häufiger?

Zum anderen eröffnen sich Perspektiven für neue elektronische Wörterbuchprodukte: Wer am PC nachschlägt, kann dann nicht nur auf detailliertere, ausführlichere und möglichst anschauliche Wörterbücher hoffen, sondern auch auf Hilfsmittel zum Finden von Formulierungen und zur Überprüfung seiner eigenen Annahmen, „wie es heißen müsste“. Angenommen, ein Amerikaner will wissen, wie man im Deutschen ausdrückt, dass man „seine Steuererklärung macht“. Zu „tax return – Steuererklärung“ könnte er im künftigen elektronischen Wörterbuch *abgeben* oder *einreichen* finden, auf Wunsch mit Beispielsätzen aus Zeitungen, aus denen die genaue Verwendung des Ausdrucks hervorgeht. Das elektronische Medium bietet mehr Platz als das gedruckte Buch; dementsprechend sollen zukünftige Wörterbücher mehr sein als nur ein elektronisches Abbild des Papierwörterbuchs. Das ist unter anderem auch für das Sprachenlernen wichtig: das Wörterbuch gibt Auskunft anhand von automatisch klassifiziertem Textmaterial.

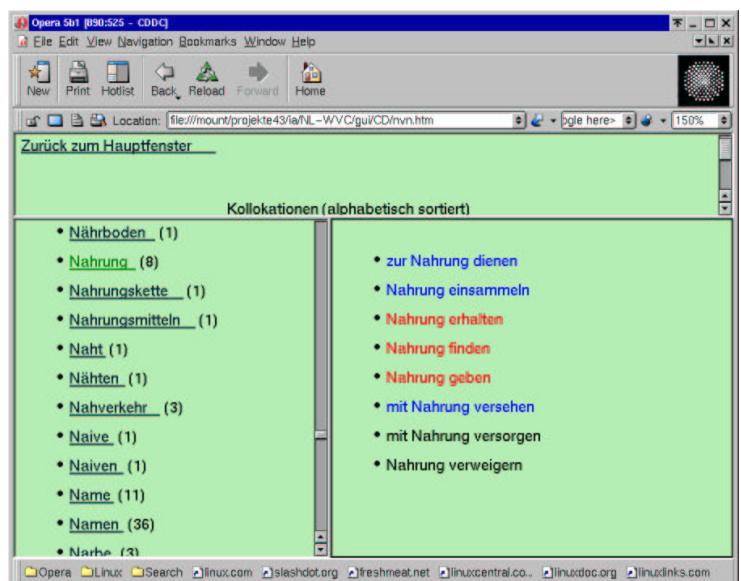
Die Rechnerunterstützung für den Wörterbuchredakteur, die in dem neuen Transferbereich im Vordergrund steht, wird somit zu besseren elektronischen Wörterbüchern führen.

Systemarchitektur

Das System ist modular aufgebaut und besteht aus mehreren Komponenten, die mittelfristig auch für andere Sprachen als Deutsch bereitgestellt werden könnten. Diese sind schematisch in der untenstehenden Abbildung angegeben. Zunächst werden die Daten des bestehenden Wörterbuchs analysiert und in eine Datenbank geschrieben. Parallel dazu werden sehr große Mengen Text mit computerlinguistischer Software durchsucht; wieder werden die Ergebnisse in die Datenbank eingetragen. Wörterbuch und Textsammlung werden jetzt im Hinblick darauf verglichen, welche wörterbuchrelevante Information sie liefern.



Die Ergebnisse des Vergleichs können zum Beispiel in einer Benutzeroberfläche inspiziert werden (siehe den nebenstehenden Bildschirmabdruck). Dort sind Verbindungen mit dem Wort *Nahrung* in einem zweisprachigen Wörterbuch mit dem verglichen worden, was zu *Nahrung* in der Textsammlung des Instituts gefunden wurde. Dabei ist in roter Schrift angegeben, was im Text gefunden wurde, so aber nicht im Wörterbuch steht; umgekehrt wird in blau markiert, was nur im Wörterbuch vorkommt, während Wortkombinationen, die in beiden Quellen zu finden sind, in schwarz am Bildschirm erscheinen.



Computerlinguistische Module

Warum Vorverarbeitung des Korpus? Warum nicht Einwort-Suche?

Ohne eine linguistische Analyse lassen sich nur einfache Suchanfragen, wie man sie aus dem Internet kennt, durchführen: „Geben Sie ein oder mehrere Wörter ein, die in dem zu suchenden Dokument auftreten sollen“, oder „Geben Sie eine genaue Wortgruppe ein“. Der Computer sucht dabei nach Zeichenfolgen, die mit denen der Anfrage identisch sind. Für viele lexikographische Fragestellungen ist diese Methode jedoch nicht ausreichend. Die Suche nach Belegen für den Ausdruck *einen Vortrag halten* (vgl. Englisch *give a talk*) verdeutlicht dies.

- Wörter treten in verschiedenen Formen auf:

Professor Maier hält einen Vortrag.
Professor Maier hielt zwei Vorträge.

- Die Wortabfolge im Deutschen ist relativ frei. Anders als z.B. im Englischen kann von der Anordnung nicht unmittelbar auf die Funktion geschlossen werden, z.B. auf Subjekt oder Objekt:

*Der Vorstand hält heute einen Vortrag.
Den Vortrag hält heute der Vorstand.
Heute wird der Vorstand den Vortrag halten.
Heute wird den Vortrag der Vorstand halten.*

- Einen ganz anderen Sinn ergibt die Sequenz im folgenden Beispiel. Die Form *hält* gehört hier zum Verb *abhalten*.

Der Vortrag hält den Vorstand von einer Teilnahme an der Besprechung ab.

Um die Gemeinsamkeiten bzw. Unterschiede der genannten Beispiele zu erfassen, muss die Software auf die Ergebnisse einer linguistischen Analyse zurückgreifen können: zu jeder Wortform muss die Grundform angegeben sein (die als Schlüssel beim Zugriff auf das Lexikon des Programms dient). Außerdem müssen Satzbausteine und deren Funktionen erkannt werden.

Die Verarbeitung von Textmaterial für die automatische Exzerption verläuft in zwei Schritten: Ein Programm zur Vorverarbeitung reichert den Text mit linguistischer Information an, z.B. mit den Grundformen der Wörter sowie mit Satzbausteinen und deren Funktionen. Das Auswertungsprogramm extrahiert in einem zweiten Schritt lexikographisch relevante Information aus dem analysierten Text. Dieses Material wird automatisch mit Angaben aus dem Wörterbuch verglichen und dann dem Lexikographen präsentiert.

Werkzeuge für die Vorverarbeitung

Chunking mit TLIPP (Three Level Incremental Partial Parser)

Das Chunking hat zum Ziel, einzelne Wörter zu größeren Einheiten („Chunks“) zusammenzufassen. Solche Einheiten sind Satzbausteine wie Subjekt und Objekt oder sinnvolle Teilstrukturen davon. Man nennt diese Einheiten auch „Phrasen“. Die Linguisten unterscheiden verschiedene Typen von Phrasen, je nach deren „Kopf“ (Kern oder Hauptelement). Subjekt und Objekt sind i.d.R. Nominalphrasen (NP), d.h. Phrasen mit einem Substantiv (Nomen) als Kopf. Wenn man sich einen Satz wie den folgenden näher betrachtet, wird klar, dass eine Analyse von Phrasen sinnvoll, wenn nicht sogar notwendig ist.

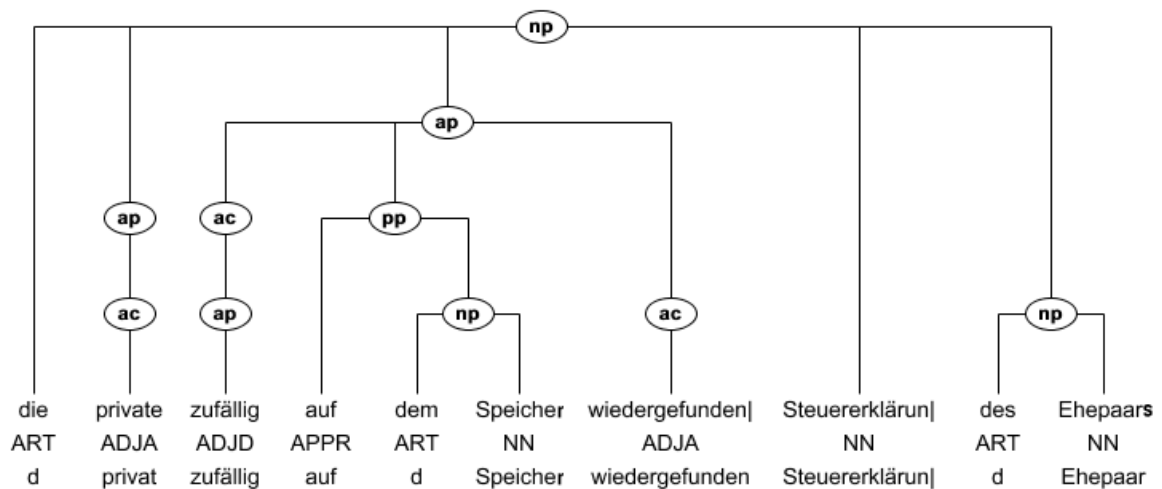
Auch die private zufällig auf dem Speicher wiedergefundene Steuerklärung des Ehepaars stand am zweiten Verhandlungstag erneut zur Debatte.

Um herauszufinden, welches Adjektiv mit welchem Substantiv kombiniert werden kann, reicht es nicht aus, nur Elemente zu betrachten, die besonders nahe beieinander stehen. Im oben genannten Beispiel würden fälschlicherweise *privat* und *Speicher* extrahiert und nicht *privat* und *Steuerklärung*. Das korrekte Paar kann nur extrahiert werden, wenn festgestellt werden kann, auf welches Substantiv sich ein Adjektiv bezieht.

TLIPP ist ein vollautomatisches Tool, das zum Chunking eingesetzt wird. Das Ziel ist, eine Basis für die Extraktion von lexikographischem und linguistischem Wissen für verschiedene Anwendungen zu schaffen. Dazu wird versucht, möglichst große Phrasen zu bilden, im Idealfall Satzbausteine wie Subjekt oder Objekt.

Die Software enthält relativ einfache Regeln, die mehrfach auf denselben Text angewendet werden. Nach jedem Regeldurchlauf werden die Ergebnisse im Korpus festgehalten. In den weiteren Analy-

seschritten kann auf die bereits erkannten Strukturen zugegriffen werden. Diese können in den Aufbau größerer Strukturen mit einbezogen oder selbst erweitert werden. Der mehrstufige Aufbau erlaubt die Verschachtelung von Phrasen derselben Kategorie, wie folgendes Beispiel verdeutlicht.



Hier ist z.B. eine NP (Nominalphrase) in eine AP (Adjektivphrase) und diese in eine größere NP eingebettet. Die Regel für die eingebettete NP und für die größere NP bleibt dieselbe.

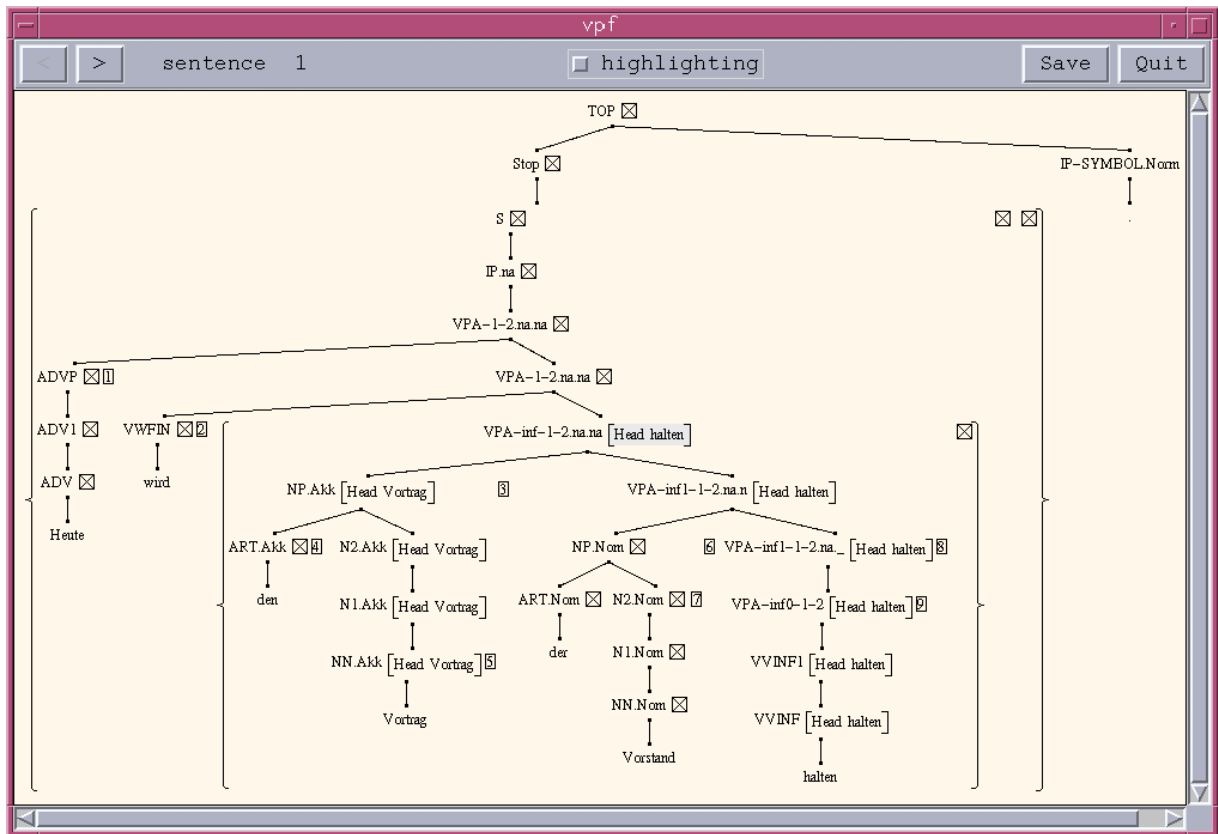
GRAMOTRON: Analyse mit einer statistischen Grammatik

Im Rahmen des Meta-Projekts GRAMOTRON¹ des Lehrstuhls für Theoretische Computerlinguistik wurden verschiedene Verfahren entwickelt, um mit statistischen Methoden linguistische Information aus Korpora zu gewinnen. Die Grundannahme des statistischen Ansatzes ist, dass sich Sprachkonventionen (d.h. Regeln und Formulierungen) durch ihre Häufigkeit bemerkbar machen. Nicht nur bei der linguistischen Vorverarbeitung sondern auch bei der Extraktion gilt: eine Analyse, die häufiger ist als andere, wird als plausibler eingestuft.

Die statistische Grammatik besteht aus Regeln, die zunächst von Hand erstellt, dann aber automatisch bewertet werden. Die Grammatik analysiert ein großes Korpus. Dabei wird für jede Regel mitgezählt, wie oft sie anwendbar war; aus dieser Frequenzangabe wird ein Wahrscheinlichkeitswert für die Regel berechnet. Mittels eines Algorithmus aus dem maschinellen Lernen (des EM-Algorithmus) kann das System die Gültigkeit der einzelnen Regeln lernen.

Für die lexikographischen Fragestellungen ist besonders wichtig, dass jede einzelne Grammatikregel für jeden möglichen lexikalischen Kopf gesondert bewertet wird. Das erlaubt es, dass auch Beziehungen zwischen nicht benachbarten Einheiten ausgewertet werden können. In der Abbildung unten ist erkennbar, dass der Kopf der Verbalphrase, *halten*, über die syntaktische Struktur in direkter Beziehung zum Kopf des Objekts, *Vortrag*, steht, unabhängig von der linearen Abfolge: Die Kopfinformation wird in der Struktur weitergegeben.

¹ siehe: <http://www.ims.uni-stuttgart.de/tcl/>



Extraktionsverfahren und Ergebnisse

Die Regeln von TLIPP sind in der CQP-Anfragesprache (Corpus Query Processor) der IMS Corpus Workbench (CWB)² geschrieben. Mit derselben Anfragesprache wird auf das von TLIPP mit Analysen angereicherte („annotierte“) Korpus zugegriffen. So kann die Extraktionssoftware nicht nur die Phrasen, sondern auch deren Köpfe und lexikalische Eigenschaften abfragen und gleichzeitig Belege dafür sammeln. Die Ergebnisse können in HTML formatiert und durchgesehen werden. Die nebenstehende Abbildung zeigt Verben, die mit dem Nomen *Steuererklärung* auftreten. Sie sind nach Häufigkeiten sortiert und mit Beispielsätzen aus dem Korpus versehen.

Opera 5b1 [985:636 - Index]
 Location: [oun/lexicon/colllocation/Steuererklarung/TFB-STEUER/index.html] [search with Google here-] [150%]

abgeben --- freq: 58
 angeben --- freq: 20
 ausfüllen --- freq: 9
 einreichen --- freq: 8
 machen --- freq: 7
 auftauchen --- freq: 5
 absetzen --- freq: 4
 bearbeiten --- freq: 4
 tun --- freq: 4
 akzeptieren --- freq: 3
 anrechnen --- freq: 3
 erledigen --- freq: 3
 vorlegen --- freq: 3
 ablaufen --- freq: 2
 abliefern --- freq: 2
 abschicken --- freq: 2
 berücksichtigen --- freq: 2
 bewegen --- freq: 2
 erleichtern --- freq: 2
 finden --- freq: 2
 fordern --- freq: 2
 gehen --- freq: 2

abgeben --- freq: 58

- 2124: Zudem soll er für die Einnahmen seiner Tochter vier Jahre lang **keine Steuererklärung abgeben** und anschließend für geschätzte Einnahmen von 35 Millionen Mark nur rund sieben Millionen Mark nachgezahlt haben .
- 19348: Die betroffenen Unternehmen sind verpflichtet , zwei Wochen nach Abschluß eines jeden Quartals **eine Steuerklärung abzugeben** , in der sie die Anzahl der nicht wieder verwerteten Becher , des Geschirrs und damit gleichsam auch ihre Steuerschuld errechnen .
- 15796: Auch Hoffmann selbst soll für die eigene Person **unrichtige Steuererklärungen abgegeben haben** , wodurch nach Ermittlungen der Staatsanwaltschaft rund 1,2 Millionen DM Einkommensteuer hinterzogen worden seien .
- 1655: Die Steuerberater wandten sich gegen die von der Bundesregierung geplante Möglichkeit , **Steuererklärungen für zwei Jahre abzugeben** .
- 13377: Wird die Immobilie nach Ablauf von mindestens zwei Jahren verkauft , lohnt es sich , **eine Steuererklärung abzugeben** , denn bei einem Verkauf danach mindert sich die

² siehe: <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>

Außer dem CQP-Format kann das Korpus auch in XML ausgegeben werden, ein Zugriff mit XSL-Stylesheets ist somit möglich. Außerdem kann das Korpus in andere Formate konvertiert werden, wie z.B. das TIGERSearch Format. TIGERSearch³ ist ein Korpus-Anfragetool, das am IMS entwickelt wurde und das erlaubt, strukturell annotierte Korpora als Baumstruktur darzustellen. Für TIGERSearch gibt es auch eine graphische Benutzerschnittstelle (GUI).

Beim statistischen Ansatz kann ein Großteil der lexikographischen Information direkt aus den statistisch bewerteten Grammatikregeln abgelesen werden. Die Abbildung rechts zeigt die Verben, die mit *Vortrag* als Objekt aufgetreten sind. Die Frequenzen zeigen deutlich, dass die Verbindung *Vortrag halten* überdurchschnittlich oft aufgetreten ist. Für den Lexikographen ist das ein wichtiger Hinweis.

Anwendungen für diese Software sind u.a. der Lexikonaufbau, z.B. typische Wortverbindungen (vgl. die Beispiele oben) oder die Identifikation typischer grammatischer und semantischer (bedeutungsmäßiger) Einschränkungen im Wortgebrauch. Daneben aber auch andere, in der Forschung diskutierte Ansätze, wie z.B. semantische Clustering-Verfahren (die zunächst unsortierte Wortmengen in Bedeutungsgruppen ordnen), vollständige oder partielle Satzanalyse sowie maschinelle Übersetzung.

Freq.	Verb
125.37	halten
14.38	geben
6.44	planen
6.13	anbieten
6.00	schließen
4.67	hören
4.13	stehen
3.92	vorsehen
3.90	organisieren
3.65	bieten
3.61	angebotien
3.00	verbinden
3.00	lesen
3.00	vorbereiten
3.00	ergänzen
2.95	ansetzen
2.93	zeigen
2.91	veranstalten
2.68	dokumentieren
2.00	absagen
2.00	widmen
2.00	gestalten
2.00	aufnehmen
2.00	liefern
2.00	übersetzen
2.00	betreffen
2.00	würzen
2.00	einplanen
2.00	aufteilen
2.00	übernehmen
2.00	ankündigen

Ansprechpartner		
<p>Dr. Vincent J. Docherty Leiter Redaktion Wörterbücher Langenscheidt KG Mies-van-der-Rohe-Straße 1 80807 München Tel.: 089/36096 – 400 Fax.: 089/36096 – 383</p>	<p>Dr. Matthias Wermke Leiter der Dudenredaktion Bibliographisches Institut & F.A. Brockhaus AG Postfach 100 311 68003 Mannheim Tel. 0621/3901 – 420 Fax. 0621/3901 – 430</p>	<p>Prof. Dr. Christian Rohrer Institut für Maschinelle Sprach- verarbeitung Computerlinguistik Azenbergstraße 12 70174 Stuttgart Tel. 0711/121 – 1365 oder 1373 Fax: 0711/121 – 1366</p>

³ siehe: <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/>