Die *Sammlung unikaler Wörter des Deutschen* Aufbauprinzipien und erste Auswertungsergebnisse

Manfred Sailer und Beata Trawiński
Sonderforschungsbereich 441
Universität Tübingen
Nauklerstr. 35
72074 Tübingen, Deutschland

Allgemeiner Hintergrund Die *Sammlung unikaler Wörter des Deutschen* (SuWD) ist der erste abgeschlossene Teil des Forschungsportals *CoDII* (*Collection of Distributionally Idiosyncratic Items*), das im Rahmen des Projekts A5 des Sonderforschungsbereichs 441 (*Linguistische Datenstrukturen*) an der Universität Tübingen erstellt wird. Die Grundidee von CoDII ist es, eine Ausgangsbasis zu sein für die linguistische Untersuchung von lexikalischen Elementen mit Distributionsbeschränkungen. Dies beinhaltet, dass CoDII die entsprechenden Elemente auflistet, ihren linguistischen Dokumentationsstand angibt und Möglichkeiten der Datenerhebung zu diesen Elementen aufzeigt. In der ersten Projektphase (2002–2004) wurden die allgemeinen Aufbauprinzipien von CoDII definiert und das hier zu präsentierende Teilportal SuWD realisiert. Für die kommende Projektphase (2005–2008) ist eine Erweiterung um ein Portal zu Polaritätselementen des Deutschen (Kürschner, 1983; Zwarts, 1997) vorgesehen. Ergänzungen um distributionell beschränkte Elemente in anderen Sprachen sind jederzeit möglich.

Unikalia Der Bestand an unikalen Wörtern des Deutschen ist in der phraseologischen Literatur seit Dobrovol'skij (1988) sehr gut dokumentiert. Dies sind Wörter wie *Tacheles*, die jeweils nur in einer bestimmten Wendung auftreten können (beispielsweise nur in: *Tacheles reden*). In Dobrovol'skij (1988), Dobrovol'skij (1989), Dobrovol'skij und Piirainen (1994b) und Dobrovol'skij und Piirainen (1994a) finden sich verschiedene Klassifikationskriterien für unikale Wörter und die Wendungen, in denen sie auftreten.[1] Ein wichtiges Augenmerk liegt bei diesen phraseologisch orientierten Publikationen auf der Frage der klaren Abgrenzung von unikalen Wörtern zu nicht-unikalen. So sprechen Dobrovol'skij und Piirainen (1994b) von etwa 600 potenziellen Unikalia des Deutschen, von denen sie dann 180 als zweifelsfrei unikal und die dazugehörigen Wendungen als Teil des Gemeinsprachschatzes klassifizieren.

Einträge in SuWD In SuWD werden zunächst unselektiv alle potenziellen Kandidaten für unikale Wörter aufgenommen, wobei wir von der Sammlung in Dobrovol'skij (1988) ausgegangen sind. Zu jedem Element werden Informationen in fünf Gruppen gesammelt. Zunächst wird jedes Element einzeln zusammen mit der/den Wendungen aufgeführt, in denen es auftreten kann. Als zweites sind die Klassifikationen gemäß Dobrovol'skij (1988), Dobrovol'skij (1989), Dobrovol'skij und Piirainen (1994b) sowie eine an Nunberg et al. (1994) orientierte semantische Klassifikation angegeben. Soweit möglich wird eine Klassifikation auch auf andere in der urspünglichen Publikation nicht genannte Elemente übertragen. Die Datenstruktur in SuWD ist so gewählt, dass problemlos weitere Klassifikationen hinzugefügt werden können (wie beispielsweise die von Feyaerts (1994)). Auch weitere Publikationen, in denen das Element diskutiert wird, sind hier aufgelistet.

Ein dritter Informationsblock beschreibt die syntaktische Struktur, innerhalb derer das unikale Element auftritt. Hierbei wird auf mögliche Variationen (Passivierung, Pronominalisierung, usw.) geachtet. Für alle Kontexte wird auf Belegsätze aus verschiedenen Korpora, dem Internet oder der Fachliteratur verwiesen.

Der linguistischen Erfassung des unikalen Elements folgen Hinweise zur weiteren Datenrecherche durch den Benutzer. Hierzu stellen wir für verschiedene öffentlich zugängliche Korpora des Deutschen optimierte Suchanfragen bereit. Diese Korpora umfassen zum Zeitpunkt der Tagung: das Internet mittels Google, die öffentlich zugänglichen Korpora des Instituts für Deutsche Sprache mittels Cosmas II, sowie die deutschen Korpora der Korpussammlung TUSNELDA des SFB441. Auch hier ist die Aufnahme weiterer Korpora und Suchwerkzeuge problemlos möglich.

Als letzte Rubrik der SuWD-Einträge stehen Angaben für die Überprüfung von empirischen Thesen zu Unikalia, wie sie im Projekt A5 aufgestellt wurden. So formulieren Soehn und Sailer (2003) Hypothesen über die möglichen Dependenzbeziehungen zwischen Unikalia und den Elementen, die obligatorisch zusammen mit dem unikalen Element auftreten. Zur Überprüfung dieser Hypothesen wird teilweise die Dependenzstruktur angegeben.

[1] Für Verweise auf eine Vielzahl weiterer Publikationen zu unikalen Wörtern siehe die *Bound Words Bibliography*, `http://www.sfb441.uni-tuebingen.de/a5/bwb/`.

Vergleich zu anderen Sammlungen Die SuWD versteht sich als ein Hilfsmittel zur linguistischen Forschung. Das bedeutet, dass vorhandene linguistische Analysen zusammengetragen und ergänzt werden, eigene Analyseansätze werden als solche gekennzeichnet. Die SuWD unterscheidet sich prinzipiell von anderen Sammlungen zu Wortkombinationen, die im Moment im Aufbau sind. Zu erwähnen sind dabei besonders: das Projekt *Usuelle Wortverbindungen* (UWV) im Rahmen von *elexiko* am IDS-Mannheim und das Projekt *Kollokationen im Wörterbuch* an der Berlin-Brandenburgischen Akademie der Wissenschaften.

Im Rahmen von UWV (Steyer, 2004) wird von statistisch hochfrequenten Wörtern ausgegangen. Diese Wörter werden einer Kookkurrenzanalyse unterworfen, deren Ergebnisse dann die Basis für eine linguistische und lexikographische Beschreibung des typischen Gebrauchs eines Wortes bilden. Diese Auswertung wird in ein elektronisches Informationssystem integriert und kann dort zum Beispiel von Deutschlernern genutzt werden. Als Datenbasis dienen dabei die am IDS gesammelten Korpora. Das methodologisch Besondere an UMV ist, dass die gesamte Datenauswahl rein statistischen Kriterien unterliegt und die linguistische Interpretation im Vergleich zu anderen Projekten relativ spät ansetzt. Im Gegensatz dazu beruht die Zusammenstellung der SuWD-Einträge auf einer primär kompetenzbasierten Klassifizierung von Elementen als potenziell unikal. Interessante Überschneidungen ergeben sich beispielsweise im Bereich der gebundenen/unikalen Lesarten von hochfrequenten Wörtern.

Auch im Rahmen des Projekts *Kollokationen im Wörterbuch* ist eine Nutzung für ein breiteres Publikum anvisiert. Die Datenbasis bietet hier der Sprachgebrauch des Deutschen im 20. Jahrhundert, wie er in den Korpora des Projekts *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts* erfasst ist. Bei der Untersuchung von Mehrwortkombinationen werden wie bei der SuWD syntaktische Flexibilität der Wendungen und andere linguistische Kriterien mit berücksichtigt. Es geht im Projekt *Kollokationen im Wörterbuch* jedoch um eine lexikographische Erfassung eines etablierten Wortkombinationenbestands des Deutschen unter Beschränkung auf verbale Phraseologismen. Im Gegensatz dazu will die SuWD (und die übrigen Teile von CoDII) eine Basis liefern, um den Möglichkeitenraum für Distributionsbeschränkungen lexikalischer Elemente zu erforschen. Dabei ist das Prinzip der "Offenheit" des Korpus relevant, d.h., dass Linguisten das Portal jederzeit verwenden können, um eigene Datenrecherche auf ständig aktualisierten Korpora (wie dem Internet) zu betreiben, und so zu aktuelleren Erkenntnissen kommen können, als die Klassifikationen und Variationsbeobachtungen, die in SuWD eingang gefunden haben.

Technische Realisierung Die SuWD wird zunächst intern in XML kodiert und ist als XHTML-Datei über das Internet öffentlich zugänglich. Die Informationen zu Unikalia werden dabei anfangs statisch angezeigt. In der nächsten Projektphase ist geplant, die SuWD in eine Datenbank zu überführen, auf die dann über eine Internet-Schnittstelle zugegriffen werden kann.

Literatur

Dobrovol'skij, Dmitrij (1988). *Phraseologie als Objekt der Universallinguistik.* Verlag Enzyklopädie, Leipzig.

Dobrovol'skij, Dmitrij (1989). Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse. In W. Fleischer, R. Große, und G. Lerchner (Hg.), *Beiträge zur Erforschung der deutschen Sprache*, Band 9, S. 57–78. Leipzig, Bibliographisches Institut.

Dobrovol'skij, Dmitrij und Piirainen, Elisabeth (1994a). PGF: Auf dem Präsentierteller oder auf dem Abstellgleis? *Zeitschrift für Germanistik* (NF 4), 65–77.

Dobrovol'skij, Dmitrij und Piirainen, Elisabeth (1994b). Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative. *Folia Linguistica 27*(3–4), 449–473.

Feyaerts, Kurt (1994). Zur lexikalisch-semantischen Komplexität der Phraseologismen mit phraseologisch gebundenen Formativen. In *Sprachbilder zwischen Theorie und Praxis*, S. 133–162. Bochum.

Kürschner, Wilfried (1983). *Studien zur Negation im Deutschen.* Gunter Narr, Tübingen.

Nunberg, Geoffrey, Sag, Ivan A., und Wasow, Thomas (1994). Idioms. *Language 70*, 491–538.

Soehn, Jan-Philipp und Sailer, Manfred (2003). At First Blush on Tenterhooks. About Selectional Restrictions Imposed by Nonheads. In G. Jäger, P. Monachesi, G. Penn, und S. Wintner (Hg.), *Proceedings of Formal Grammar 2003*, S. 149–161.

Steyer, Kathrin (2004). Kookkurrenz. Korpusmethdik, linguistisches Modell, lexikographische Perspektiven. In K. Steyer (Hg.), *Tagungsband der 39. Jahrestagung des IDS.* IDS Mannheim, Mannheim. in Vorbereitung.

Zwarts, Frans (1997). Three Types of Polarity. In F. Hamm und E. W. Hinrichs (Hg.), *Plurality and Quantification*, S. 177–237. Kluwer Academic Publishers.