INSTITUT FÜR DEUTSCHE SPRACHE

OPAL

Online publizierte Arbeiten zur Linguistik

ISSN 1860-9422

2/2014

Im Auftrag des Instituts für Deutsche Sprache
herausgegeben von Hardarik Blühdorn, Mechthild Elstermann und Annette Klosa

Andrea Abel / Lothar Lemnitzer (Hg.)

Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern

Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
opal@ids-mannheim.de

Mitglied der Leibniz-Gemeinschaft

© 2014 IDS Mannheim – Alle Rechte vorbehalten

Das Werk einschließlich seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechts ist ohne Zustimmung der Copyright-Inhaber unzulässig und strafbar. Das zulässige Zitieren kleinerer Teile in einem eigenen selbstständigen Werk (§ 51 UrhG) erfordert stets die Angabe der Quelle (§ 63 UrhG) in einer geeigneten Form (§ 13 UrhG). Eine Verletzung des Urheberrechts kann Rechtsfolgen nach sich ziehen (§ 97 UrhG). Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen. Die zugänglichen Daten dürfen von den Nutzern also nur zu rein wissenschaftlichen Zwecken genutzt werden. Eine darüber hinausgehende Nutzung, gleich welcher Art, oder die Verarbeitung und Bearbeitung dieser Daten mit dem Zweck, sie anschließend selbst oder durch Dritte kommerziell zu nutzen, bedarf einer besonderen Genehmigung des IDS (Lizenz). Es ist nicht gestattet, Kopien der Textdateien auf externen Webservern zur Verfügung zu stellen oder Dritten auf sonstigem Wege zugänglich zu machen. Bei der Veröffentlichung von Forschungsergebnissen, in denen OPAL-Publikationen zitiert werden, bitten die Autoren und Herausgeber um eine entsprechende kollegiale Information an opal@ids-mannheim.de.

Inhalt

Andrea Abel / Lothar Lemnitzer
Einleitung ... 3

Peter Meyer
Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von *elexiko* ... 9

Almuth Bedenbender
Das Deutsche Rechtswörterbuch im Netz ... 22

Frans Heyvaert
On the descriptive power of the *ANW* semagram ... 29

Jörg Didakowski / Alexander Geyken
From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions ... 39

Sabine Bartsch / Stefan Evert
Towards a Firthian Notion of Collocation ... 48

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Einleitung

Andrea Abel, EURAC Bozen / Bolzano
Lothar Lemnitzer, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

Dieser Band vereinigt Beiträge aus zwei Arbeitstreffen des von der DFG geförderten wissenschaftlichen Netzwerks „Internetlexikografie“ (www.internetlexikografie.de) und setzt damit die Reihe der Arbeitsberichte des Netzwerks (vgl. Klosa / Müller-Spitzer (Hg.) 2011 sowie Abel / Klosa 2013) fort. Das zweite Arbeitstreffen des Netzwerks fand am 5. und 6. Dezember 2011 in Berlin (DE) statt und hatte „Vernetzungs- und Zugriffsstrukturen bei Internetwörterbüchern“ zum Thema. Das folgende Arbeitstreffen wurde am 3. und 4. Mai 2012 in Bozen (IT) abgehalten und beschäftigte sich mit „Aspekten der automatischen Gewinnung von lexikografischen Angaben“. Dieser Band wird daher von uns, Andrea Abel und Lothar Lemnitzer, stellvertretend für die Organisatoren des jeweiligen Arbeitstreffens herausgegeben. Der Band enthält die Ausarbeitungen ausgewählter Beiträge der beiden Treffen.

Beim zweiten Arbeitstreffen mit dem Thema „Vernetzung und Zugriffsstrukturen“ wurden Fragen für die Internetlexikografie behandelt, die für den Bereich der Printlexikografie schon recht lange diskutiert wurden, durch die Möglichkeiten des neuen Mediums, die sich am besten mit dem Stichwort „Verlinkung“ charakterisieren lassen, aber eine ganz neue Dimension erhalten. Die Vernetzung lexikografischer Daten kann in Internetwörterbüchern über Verlinkungen zwischen Wortartikeln, zwischen Lesarten verschiedener Wortartikel oder Angaben innerhalb von Wortartikeln geschehen. Es können darüber hinaus andere, externe Ressourcen, vor allem Text- und Belegkorpora, mit den lexikografischen Daten verlinkt werden, sodass von „Hypermedia-Wörterbüchern“ (Storrer 1998) bzw. „Wortschatzinformationssystemen“ (Müller-Spitzer 2007) gesprochen werden kann. Schließlich können in sogenannten Wörterbuchportalen oder Wörterbuchnetzen mehrere lexikalische Ressourcen miteinander verbunden werden, wodurch das gleichzeitige Nachschlagen in mehreren Werken möglich wird. Solche Vernetzungen können bei entsprechender Kodierung vor allem genutzt werden, um bestimmte Relationen zwischen lexikologischen Einheiten auffindbar zu machen. Internetwörterbücher können so neben semasiologischen Zugriffsmöglichkeiten ebenso onomasiologische und ggf. auch konzeptionelle Zugriffsstrukturen anbieten. Außerdem ermöglicht es das Medium Internet, explizit kodierte Vernetzungen lexikografischer Daten den Benutzern dergestalt zu präsentieren, dass die zugrunde liegenden inhaltlichen Zusammenhänge adäquat rezipiert werden können.

Beim Berliner Arbeitstreffen wurden verschiedene Möglichkeiten der Kodierung von Vernetzungen und ihre Onlinepräsentation anhand von Beiträgen zu theoretischen und praktischen Aspekten diskutiert, um zu zeigen, wie sie für unterschiedliche Zugriffsstrukturen nutzbar gemacht werden können.

Angelika Storrer (Dortmund) präsentierte in ihrem Eröffnungsvortrag Überlegungen zum Hypertextkonzept in der Lexikografie (siehe auch Storrer 1998 und Storrer i. Dr.).

Stefan Schierholz (Erlangen) exemplifizierte seine Überlegung zu Mediostrukturen in Wörterbüchern anhand der Reihe „Wörterbücher zur Sprach- und Kommunikationswissenschaft“, die bei de Gruyter erscheint (siehe auch Schierholz / Wiegand 2004).

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Im ersten Gastvortrag stellten *Erhard Hinrichs* und *Verena Henrich* (Tübingen) Vernetzungs- und Zugriffsstrukturen im Deutschen Wortnetz *GermaNet* vor, wobei sie einen Schwerpunkt auf die Erkennung von Grund- und Bestimmungswörtern in Komposita legten (siehe auch Henrich / Hinrichs 2011).

Jörg Asmussen (København) zeigte als einer von zwei Referenten zu diesem Thema am Beispiel des dänischen Wörterbuchportals *ordnet.dk*, wie Wörterbucheinträge und Korpusbelege miteinander vernetzt sind und welche Probleme dies bei verschiedenen Gruppen von Benutzern verursachen kann (siehe auch Asmussen i. Dr.).

Peter Meyer (Mannheim, Beitrag in diesem Band) trug vor, wie ein Verweiskonzept bei informationstechnisch adäquater Modellierung der lexikografischen Daten so implementiert werden kann, dass es sowohl im lexikografischen Arbeitsprozess als auch aus Benutzersicht gute Dienste leistet.

Im Mittelpunkt des Beitrags von *Axel Herold* (Berlin) stand das Konzept einer Metalemmaliste, über die auf der Webseite des DWDS mehrere Wörterbücher so verlinkt sind, dass bei Wortabfragen die passenden Informationen aus verschiedenen Wörterbuchquellen synoptisch in mehreren Anzeigefenstern dargestellt werden. Er diskutierte die notwendige Granularität einer solchen Liste, insbesondere wenn die zugrundeliegenden Wörterbücher Homografen-Einträge enthalten, und stellte einige Probleme beim Mapping der Stichwortlisten dieser Wörterbücher vor (siehe auch Herold / Lemnitzer / Geyken 2012).

Im zweiten Gastvortrag führte *Gilles-Maurice de Schryver* (Gent) in ein Zugriffskonzept für Wörterbucheinträge im Allgemeinen und für Definitionen im Besonderen ein, das den Nachschlagebedürfnissen und Verstehensniveaus verschiedener Gruppen von Benutzern gerecht wird. De Schryver mahnte, dass Lexikografen das Verstehensniveau ihrer potenziellen Nutzer in der Regel überschätzten und deshalb verschiedene Darstellungsniveaus für die beschriebenen linguistischen Sachverhalte ein Desideratum seien (siehe auch de Schryver 2010).

Im Anschluss gab es einen Vortrag von *Frank Michaelis* (Mannheim) über die Zugriffsmöglichkeiten im Wörterbuchportal OWID (www.owid.de), das eine Reihe recht heterogener lexikalischer Ressourcen zur deutschen Sprache vereint (siehe auch Müller-Spitzer 2010).

Almuth Bedenbender (Heidelberg, Beitrag in diesem Band) stellte – als zweite Referentin zu diesem Thema – die Verlinkung von Einträgen des Deutschen Rechtswörterbuchs und vor allem von den Belegen zu den Quellen, also zu einem ebenfalls sich im Aufbau befindlichen Korpus von Volltexten und Bilddigitalisaten, vor. Darüber hinaus ging sie auf weitere Aspekte der Vernetzung vom Deutschen Rechtswörterbuch auf andere Ressourcen und auch von anderen Ressourcen auf das Deutsche Rechtswörterbuch ein.

Frans Heyvaert (Leiden, Beitrag in diesem Band) legte das Konzept der „Semagrams“ dar, eine Form der lexikalisch-semantischen Vernetzung, durch die der Zugriff auf Einträge im „Algemeen Nederlands Woordenboek“ auch über die Bedeutungsebene möglich ist.

Iryna Gurevych und *Christian Meyer* (Darmstadt) referierten über ihre Arbeiten zum Alignment von *Wiktionary* und *Wordnet* (siehe auch Meyer / Gurevych 2011).

Als letzten Beitrag dieses Arbeitstreffens stellte *Vera Hildenbrandt* (Trier) das Vernetzungskonzept im Trierer Wörterbuchnetz vor, das eine Vielzahl von historischen Wörterbüchern,

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Dialektwörterbüchern und Autorenwörterbüchern unter einem virtuellen Dach vereint und gemeinsam durchsuchbar macht (siehe auch Burch / Rapp 2007).

Auf dem Arbeitstreffen wurden somit themenspezifische Einblicke in verschiedene Wörterbuchprojekte und -produkte gegeben. Der thematische Schwerpunkt der Beiträge lag dabei sicher bei den Möglichkeiten des Internets – als gigantischer Hypertext –, Ressourcen verschiedenster Provenienz auf verschiedenen Ebenen und verschieden stark zu vernetzen. Auch Probleme der Vernetzung heterogener Ressourcen vor allem für mit dem neuen Medium weniger vertraute Nutzer wurden dabei deutlich.

Während es im zweiten Arbeitstreffen um Vernetzungs- und Zugriffstrukturen ging, drehten sich beim dritten Arbeitstreffen Vorträge und Diskussionen um das Thema „Automatische Gewinnung von lexikografischen Angaben“. Die Verfügbarkeit großer elektronischer Textkorpora hat die Lexikografie generell stark verändert. Ein- und mehrsprachige Wörterbücher werden zunehmend korpusgestützt erarbeitet, und das unabhängig vom jeweiligen Publikationsmedium. Zugleich haben Korpuslinguisten reiche Korpusrecherche- und -analysetools entwickelt, während Computerlexikografen und -linguisten z.B. zur automatischen Lesartendisambiguierung oder zur Entwicklung lexikalisch-semantischer Ressourcen (z.B. Ontologien) beigetragen haben. Welche lexikografisch relevanten Daten automatisch extrahiert werden können, welche Methoden und Werkzeuge dabei zum Einsatz kommen, wie zuverlässig die gewonnenen Daten sind und wie Qualität gewährleistet werden kann, welche Rolle den Lexikografen bei der Datengewinnung und -auswertung zukommt, welche Daten schließlich in welchem Umfang und in welcher Form in unterschiedliche Benutzerumgebungen eingebunden werden sollen, darüber tauschten sich Wissenschaftler und Praktiker während des Bozner Arbeitstreffens aus.

Marco Baroni (Trient) präsentierte im ersten Gastvortrag einen Ansatz, bei dem das Firth’sche Zitat „you can tell a word by the company it keeps“ im Hinblick auf die Eingrenzung von Wortbedeutungen nicht nur auf sprachliche, sondern auch auf bildliche Kontexte angewandt wird. Er führte neue Methoden der distributionellen Semantik vor, die durch den kombinierten Einbezug von Text- und Bildumgebungen die automatische Extraktion semantischer Informationen über Wörter aus multimodalen Korpora ermöglichen, und diskutierte, wie die Ergebnisse für lexikografische Belange nutzbar gemacht werden können (siehe auch Bruni et al. 2012).

Adam Kilgarriff (Brighton) ging im zweiten Gastvortrag, ausgehend von der provokanten Frage, ob Lexikografen in Anbetracht zunehmender Automatisierung überhaupt noch gebraucht würden, auf Perspektiven für die automatische Lexikografie ein. Anhand einer Anforderungsliste („shopping list“) für die lexikografische Praxis und der Darstellung konkreter Beispiele in Form von „word sketches“, die mithilfe der Sketch Engine automatisch gewonnen wurden, reflektierte er kritisch, in welchen Bereichen mehr (z.B. Kollokationen) oder weniger (z.B. Semantik) zuverlässige Daten zu erwarten sind. Auf die Arbeit des Lexikografen könne, so das Fazit, jedenfalls nicht verzichtet werden (siehe auch Kilgarriff et al. 2010).

Im Anschluss an die Ausführungen mit Bezug auf die „word sketches“ für das Englische und das Tschechische fokussierte der Beitrag von *Jörg Didakowski* und *Alexander Geyken* (Berlin, Beitrag in diesem Band) auf deren deutsches Pendant, nämlich die Wortprofile, die als Teil des Wortinformationssystems „Digitales Wörterbuch der deutschen Sprache“ (DWDS) entstanden sind und Ergebnisse mit unterschiedlichen syntaktischen Relationen des Deutschen liefern. Die Referenten vertieften insbesondere methodische Herausforderungen der Datenex-

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

traktion, die Besonderheiten der deutschen Sprache geschuldet sind, und Aspekte, die die Korpusgrundlage sowie die Anwendung statistischer Maße betreffen.

Stefan Evert (Erlangen) und *Sabine Bartsch* (Darmstadt, Beitrag in diesem Band) stellten eine Studie zur Gewinnung von Kollokationen in einem strikt Firth'schen Sinn vor. Sie analysierten die Rolle der Korpusgröße, -zusammensetzung und -aufbereitung sowie der Anwendung unterschiedlicher statistischer Maße bei der automatischen Kollokationsextraktion und evaluierten die Ergebnisse unter der Verwendung von Wortkombinationen des „BBI Combinatory Dictionary" als Goldstandard. Die Untersuchungen ließen unter anderem den vorsichtigen Schluss zu, dass die Zusammensetzung und „Sauberkeit" (engl. „cleanness") des Korpus eine wichtigere Rolle spielen als dessen Umfang.

Roman Schneider (Mannheim) sprach darüber, wie mit automatisch und quantitativ ermittelten Daten lexikografische Artikel von Internetwörterbüchern angereichert werden können. Er veranschaulichte Möglichkeiten und Herausforderungen am Beispiel von E-VALBU, der Onlineversion des Valenzwörterbuchs deutscher Verben (Schumacher et al. 2004), für das Daten unter der Nutzung der KoGra-Datenbank als Recherchebasis ermittelt werden (siehe auch Schneider 2012).

Magali Paquot (Louvain) nahm schließlich eine didaktische Perspektive ein. Thema ihres Vortrags waren „Dictionary-cum-corpus-query-tools" und deren Anwendungsmöglichkeiten. Ihre Ausführungen konzentrierten sich auf das pädagogische „Louvain English for Academic purposes Dictionary" (LEAD), ein webbasiertes Instrument mit einem Schwerpunkt auf formelhaften Verbindungen, das über ein Abfragesystem einen direkten Zugriff auf Korpusdaten ermöglicht, und problematisierten dabei besonders Vor- und Nachteile automatisch extrahierter Beispiele in einer solchen didaktischen Lernumgebung (siehe auch Granger / Paquot 2010).

Im Rahmen des Bozner Treffens wurde deutlich, dass die zunehmende Automatisierung bei der Gewinnung lexikografischer Angaben zwar als großer Vorteil gewertet wird, dass aber bestimmte sprachliche Bereiche, wie etwa semantische, nach wie vor schwieriger zu bewältigen sind als andere. In den Diskussionen kristallisierte sich zudem heraus, dass Lexikografen automatisch generierten Ergebnissen bisweilen skeptisch gegenüberstehen und deren Zuverlässigkeit kritisch hinterfragen, aber auch, dass viele Fragen im Hinblick auf die Bereitstellung (unter anderem ob, wie, in welchem Umfang) solcher Daten für den Endnutzer nach wie vor unbeantwortet sind. Diesbezüglich seien zweifelsohne empirische Nutzerstudien nötig. Die Rolle und Relevanz des Lexikografen im Wörterbuchprozess, die sich vor dem Hintergrund der technischen Entwicklungen zwar verändern, scheinen insgesamt jedoch nicht infrage gestellt zu werden.

In den beiden Arbeitstreffen ist es gelungen, einschlägige Experten zu den jeweils zur Diskussion gestellten Themenbereichen aus ganz Europa zusammenzubringen, die in ihren Beiträgen Einsichten in laufende Forschungsarbeiten boten und die Aktualität der aufgeworfenen Fragestellungen aufzeigten. Dies und die Relevanz der aktuellen Arbeiten lässt sich nicht zuletzt auch an der unten stehenden Literaturliste sowie an den in diesen Band aufgenommenen Beiträgen erkennen.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Literatur

Abel, Andrea / Klosa, Annette (2013): Der Nutzerbeitrag im Wörterbuchprozess. 3. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. (= OPAL – Online publizierte Arbeiten zur Linguistik). Mannheim: Institut für Deutsche Sprache.

Asmussen, Jörg (i. Dr.): Combined products: Dictionary and corpus. In: Gouws, Rufus H. et al. (Hg.): Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography. Berlin / New York: de Gruyter.

Burch, Thomas / Rapp, Andrea (2007): Das Wörterbuch-Netz. Verfahren – Methoden – Perspektiven. In: Burckhardt, Daniel / Hohls, Rüdiger / Prinz, Claudia (Hg.): Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006. (= Historisches Forum 10, 1). Berlin: Clio-online, S. 607-627. Internet: http://edoc.hu-berlin.de/histfor/10_I/PDF/HistFor_2007-10-I.pdf (Stand: 10.05.2013).

Bruni, Elia et al. (2012): Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. Brave New Idea paper. Proceedings of MM 12 (20th ACM International Conference on Multimedia), New York NY: ACM, S. 1219-1228. Internet: http://clic.cimec.unitn.it/marco/publications/bruni-etal-acmmm-2012.pdf (Stand: 10.05.2013).

de Schryver, Gilles-Maurice / Prinsloo, Daniel (2010): Do Dictionaries Define on the Level of Their Target Users? A Case Study for Three Dutch Dictionaries. In: International Journal of Lexicography, 25 (1), S. 5-28.

Granger, Sylviane / Paquot, Magali (2010): Customising a general EAP dictionary to meet learner needs. In Granger, Sylviane / Paquot, Magali (Hg.): eLexicography in the 21st century: New challenges, new applications. Proceedings of ELEX2009. (= Cahiers du CENTAL). Louvain-la-Neuve: Presses universitaires de Louvain, S. 87-96.

Henrich, Verena / Hinrichs, Erhard (2011): Determining Immediate Constituents of Compounds in GermaNet. In Proceedings of Recent Advances in Natural Language Processing (RANLP 2011), Hissar, Bulgaria, September 2011, S. 420-426. Internet: http://www.aclweb.org/anthology-new_/R/R11/R11-1058.pdf (Stand: 10.05.2013).

Herold, Axel / Lemnitzer, Lothar / Geyken, Alexander (2012): Integrating lexical resources through an aligned lemma list. In: Chiarcos, Christian / Nordhoff, Sebastian / Hellmann, Sebastian (Hg.): Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata. Berlin / Heidelberg: Springer, S. 35-44.

Kilgarriff, Adam et al. (2010): A Quantitative Evaluation of Word Sketches. In: Dykstra, Anne / Schoonheim, Tanneke (Hg.): Proceedings of the XIV Euralex International Congress. Leeuwarden, 6-10 July 2010. Fryske Akademy: Leeuwarden, S. 372-379. Internet: http://www.euralex.org/proceedings-toc/euralex_2010/ (Stand: 10.05.2013).

Klosa, Annette / Müller-Spitzer, Carolin (Hg.) (2011): Datenmodellierung für Internetwörterbücher. 1. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. (= OPAL – Online publizierte Arbeiten zur Linguistik 2 / 2011). Mannheim: Institut für Deutsche Sprache.

Meyer, Christian M. / Gurevych, Iryna (2011): What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, November 2011. Chiang Mai, Thailand, S. 883-892. Internet: http://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/ijcnlp2011-meyer-wiktionary-alignment.pdf (Stand: 10.05.2013).

Müller-Spitzer, Carolin (2007): Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis. (= Studien zur Deutschen Sprache 42). Tübingen: Narr.

Müller-Spitzer, Carolin (2010): OWID – A dictionary net for corpus-based lexicography of contemporary German. In: Dykstra, Anne / Schoonheim, Tanneke (Hg.): Proceedings of the XIV Euralex International Congress. Leeuwarden, 6-10 July 2010. Fryske Akademy: Leeuwarden, S. 445-452. Internet: http://www.euralex.org/proceedings-toc/euralex_2010/ (Stand: 10.05.2013).

Schierholz, Stefan / Wiegand, Herbert Ernst (2004): Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK), herausgegeben von Stefan J. Schierholz und Herbert Ernst Wiegand. In: Forschungsnotizen. Zeitschrift für Germanistische Linguistik 32, S. 87-88.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Schneider, Roman (2012): Die Korpusdatenbank KoGra-DB. In: Grammatisches Informationssystem grammis.

Schumacher, Helmut / Kubczak, Jacqueline / Schmidt, Renate / de Ruiter, Vera (2004): VALBU – Valenzwörterbuch deutscher Verben. (= Studien zur Deutschen Sprache 31). Tübingen: Narr.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. (= Lexicographica, Series Maior 84). Tübingen: Niemeyer, S. 106-131.

Storrer, Angelika (i. Dr.): Representing dictionaries in hypertextual form**.** In: Gouws, Rufus H. et al. (Hg.): Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography. Berlin / New York: de Gruyter. Internet: http://www.studiger.tu-dortmund.de/images/Storrer-hypertext2011-preprint.pdf (Stand: 10.05.2013).

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von *elexiko*

Peter Meyer, Institut für Deutsche Sprache Mannheim

Abstract

This contribution outlines a conceptual analysis of the dictionary-internal cross-reference structure in electronic dictionaries along the lines of Wiegand's actional-theoretical text theory of print dictionaries. The discussion focuses on issues of XML-based data modeling, using the monolingual German online dictionary *elexiko* as a running example. The first part of the article demonstrates how Wiegand's formal theory of mediostructure and its intricate nomenclature can be extended in a systematic and lexicographically justified way to cover the structure of the underlying lexicographical database of online dictionaries. The second part of the article applies the concepts developed to a more technical question, examining the extent to which cross-reference information can be stored and processed separately from the dictionary entry documents, e.g., in a relational database. The results are largely negative; in most real world cases, this leads to an unwanted duplication of XML-related structural information. The concluding third part briefly describes the strategy chosen for *elexiko*: mediostructural information is not externalized at all; cross-reference consistency checks are performed by a dictionary editing tool that takes advantage of a specialized XML database index and can easily be made more efficient and scalable by using a simple caching technique.

1. Repräsentation von Verweisstrukturen in elektronischen Wörterbüchern

1.1 Fragestellungen

Konzeptuelle Überlegungen zur Verweis- oder Mediostruktur elektronischer Wörterbücher gestalten sich deutlich vielschichtiger als bei gedruckten Wörterbüchern, weil außer den – prototypisch durch Hyperlinks realisierten – Verweisen auf der Ebene der Präsentation die zugrundeliegende Ebene der Datenmodellierung betrachtet werden muss, die zu Präsentationsaspekten in einem durchaus komplexen Wechselverhältnis steht (vgl. Blumenthal/ Lemnitzer/Storrer 1988). Müller-Spitzer (2007a, 2007b) entwickelt für die konzeptionelle Datenmodellierung von XML-basierten elektronischen Wörterbüchern einen begrifflichen Rahmen analog zu Wiegands (1996, 2002) Theorie der Mediostruktur von Printwörterbüchern. Die vorliegenden Ausführungen schließen in vielerlei Hinsicht an die genannten Arbeiten an und erweitern sie um einige Aspekte mit besonderer Relevanz für die computerlexikografische Praxis. In Abschnitt 1 wird in Ansätzen eine begriffliche Analyse von Vernetzungen, d.h. Verweisstrukturen auf der Ebene der Datenmodellierung, versucht; Abschnitt 2 untersucht dann, inwieweit die Mediostruktur im XML-basierten Wörterbuch in eigenständige Datenstrukturen ausgelagert werden kann.

Die Diskussion wird am Beispiel von *elexiko*, einem korpusbasierten monolingualen Onlinewörterbuch des Gegenwartsdeutschen, geführt, das am Institut für Deutsche Sprache entwickelt wird und unter www.elexiko.de frei zugänglich ist (Haß (Hg.) 2005; Klosa (Hg.) 2011); die grundsätzlichen Überlegungen sind jedoch ohne Weiteres auf andere XML-basierte Wörterbücher übertragbar. Um die Lesbarkeit zu erhöhen, wird in diesem Text zu Beispielzwecken ein deutlich vereinfachtes XML-Schema verwendet.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

1.2 Präsentationsebene: Zur begrifflichen Analyse von Verweisen

Die Artikel des Wörterbuchs *elexiko* sind auf der Ebene der Datenbasis *inhaltsorientiert* ausgezeichnet, vergleichbar dem *lexical view* in den aktuellen TEI-Richtlinien[1]; jedem Artikel des Wörterbuchs entspricht – und diese Voraussetzung gilt für den gesamten vorliegenden Beitrag – ein separates XML-Dokument, dessen *Inhaltsstruktur* durch die DTD bzw. das XML-Schema des Wörterbuchs vorgegeben ist (zu den Begrifflichkeiten vgl. im Einzelnen Müller-Spitzer 2007b). Wir betrachten im Folgenden den typischen Fall von Verweisangaben zwischen *elexiko*-Artikeln und greifen das Beispiel paradigmatischer Relationen (Sinnbeziehungen wie Synonymie, Hyponymie etc.) zwischen Lesarten verschiedener Lemmata heraus. Auf der Präsentationsebene erscheinen die Verweise als Hyperlinks auf einer lesartspezifischen Registerkarte für sinnverwandte Wörter (vgl. Abb. 1). In der in Wiegand (2002) für das Printmedium entwickelten, begrifflich natürlich an die veränderten Bedingungen des Onlinemediums anzupassenden Terminologie ist eine solche *Verweisangabe* artikelintern an eine *Verweisausgangsangabe* (als *Bezugsadresse* der Verweisangabe, die den *Verweisausgangsbereich* identifiziert, von dem aus verwiesen wird) adressiert, die durch das betrachtete Lemma und eine die betrachtete Lesart semantisch umschreibende Kurzetikettierung oberhalb der Registerkarte gebildet wird. In Abbildung 1 ist der Verweisausgangsbereich informell durch „Artikel *Familie*, Lesart ‘Verwandte’“ beschreibbar. Mediostrukturell ist die Verweisangabe an die *Verweisadresse* adressiert, an die der Benutzer verwiesen wird und mit der er wiederum auf den zugehörigen *Verweiszielbereich* (hier ebenfalls durch die Angabe von Lemma und Lesartenkurzetikettierung identifiziert) verwiesen wird. Die Verweisangabe selber steht in einer (in der HTML-Darstellung zweidimensional zu charakterisierenden) *Verweisposition*, die ausdrücklich nicht identisch ist mit dem Verweisausgangsbereich. Die Verweisangabe[2] selber enthält eine *Verweisadressenangabe*, die offenbar nicht in herkömmlicher Weise durch ein Textsegment gegeben ist; klickt man etwa bei den Synonymen zu *Familie* / Lesart ‘Verwandte’ auf das Wort *Haushalt*, wird man per Hyperlink zum *elexiko*-Artikel *Haushalt* in der Lesart ‘Personengruppe’ weitergeleitet; relevant ist hier also das im HTML- / JavaScript-Quellcode der Seite – im typischen Fall durch eine URL-Angabe – spezifizierte Interaktionsverhalten der betreffenden Ansicht des Onlinewörterbuchs, nicht die in diesem Falle verkürzte textuelle Spezifikation, die der Benutzer im Browserfenster sieht.[3]

1.3 Datenmodellierungsebene: Zur begrifflichen Analyse von Vernetzungen

Das soeben geschilderte visuell-interaktionale Verständnis der Präsentationsebene ist begrifflich von den Verhältnissen in der *lexikografischen Datenbasis*, vereinfacht gesagt also der Gesamtheit der den Wortartikeln zugrunde liegenden XML-Dokumente, zu trennen. Diese Trennung hat zum einen naheliegende technische Gründe, denn die tatsächliche Kodierung des Verweiszieles in einem XML-Dokument wird von der Webapplikation auf komplexe und letztlich arbiträre Weise in eine URL-basierte Adressierung von dynamisch generierten Webseiten „übersetzt“, wobei Webseiten (oder, im Falle von AJAX-Anwendungen, auch HTML-

[1] Der *lexical view* wird in den aktuellen Richtlinien (TEI-P5) wie folgt erläutert: „this view includes the underlying information represented in a dictionary, without concern for its exact textual form“ (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html); weitere „views“ sind der *typographic view* und der *editorial view*.

[2] Wir unterscheiden hier nicht zwischen (rein verweisvermittelnden) Verweisangaben und (nicht ausschließlich verweisvermittelnden) Angaben mit Verweiskennzeichnung. Im Text behandeln wir nur erstere; die vorgestellten Überlegungen lassen sich jedoch auf letztere übertragen.

[3] Zu Hyperlinks aus metalexikografischer Sicht vgl. noch Kammerer (1998), dem jedoch ersichtlich das Konzept einer konzeptuellen Datenmodellierung noch nicht zur Verfügung stand.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Fragmente) in keiner simplen Entsprechungsrelation zu lexikografischen Zieladressen stehen. Zum anderen korrespondiert, wie in Müller-Spitzer (2007a, S. 142ff.) ausführlicher erläutert wird, nicht jede Verweisrelation auf der Präsentationsebene mit einer Vernetzungsbeziehung in der Datenbasis und umgekehrt.

Abb. 1: Ausschnitt aus der Onlinepräsentation des *elexiko*-Artikels *Familie* in der Lesart ‘Verwandte’; Angabebereich „Sinnverwandte Wörter“ (http://www.owid.de/artikel/3920/Verwandte; Stand Oktober 2012)

Im Folgenden betrachten wir anhand der paradigmatischen Relationen in *elexiko* den einfachsten Fall, in dem ein bestimmter Typ von Verweisangabe auf der Präsentationsebene tatsächlich systematisch – d.h. letztlich aufgrund der Programmierung der Webapplikation – einem bestimmten Typ von *Vernetzung* auf der Ebene der XML-Datenbasis entspricht. Unter Vernetzung verstehen wir hier zunächst nur den rein formal definierbaren Sachverhalt, dass ein XML-Element z.B. durch seine Attribute ein anderes XML-Element (typischerweise in einem anderen XML-Dokument derselben Datenbasis) referenziert. Die Referenzierung geschieht üblicherweise über ein Adressierungsverfahren, das – für einen bestimmten Typ von

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

XML-Elementen[4] – eine bijektive (eineindeutige) Zuordnung von Elementen eben dieses Typs zu IDs (identifizierenden Zeichenketten) oder geordneten n-Tupeln von IDs herstellt. Letzterer Fall lässt sich einfacher anhand eines deutlich vereinfachten und schematisierten Ausschnittes aus dem XML-Dokument des *elexiko*-Wortartikels *Familie* erläutern (siehe Abb. 2, der entsprechende Ausschnitt aus der Onlinepräsentation ist in Abb. 1 dargestellt).

```
<artikel id="1234">
  <lemmazeichen>Familie</lemmazeichen>
  …
  <lesart id="Verwandte">
    …
    <paradigmatik>
      …
      <synonymie>
        …
        <relpartner art-id="5678" lesart-id="Personengruppe">
          Haushalt
        </relpartner>
      </synonymie>
      …
    </paradigmatik>
    …
  </lesart>
  …
</artikel>
```

Abb. 2: Vereinfachter Ausschnitt aus dem XML-Dokument des *elexiko*-Wortartikels *Familie*: Synonymievernetzung von der Lesart ‘Verwandte’ auf die Lesart ‘Personengruppe’ des Artikels *Haushalt*. Die für Adressierungen verwendeten IDs sind fett gesetzt.

Der Artikel als Ganzes und damit auch das den gesamten Artikel umfassende Dokumentelement ist durch die ID ‘1234’ in der zugrundeliegenden Datenbasis eineindeutig identifizierbar; die Lesart mit der Kurzetikettierung ‘Verwandte’ ist durch das geordnete Paar (‘1234’, ‘Verwandte’) eineindeutig identifizierbar; innerhalb von Lesarten werden in *elexiko* – im Beispiel nicht gezeigt – gelegentlich noch Lesartenspezifizierungen unterschieden, die dann durch ein geordnetes Tripel von IDs, z.B. (‘1234’, ‘Verwandte’, ‘Dynastie’), identifiziert werden. Die Vernetzung auf das Lemma *Haushalt* in der synonymen Lesart ‘Personengruppe’ geschieht, wie im Beispiel ersichtlich, durch Angabe des zugehörigen geordneten ID-Paares (‘5678’, ‘Personengruppe’); die Angabe der Lemmazeichengestalt *Haushalt* ist redundant und erfolgt aus pragmatischen Gründen, denn so kann die zugehörige HTML-Repräsentation ausschließlich auf der Grundlage dieses einen XML-Dokuments gewonnen werden, im Falle von *elexiko* durch XSL-Transformationen; es müssen nicht erst in der Datenbank sämtliche Lemmazeichenangaben in den adressierten Artikeln nachgeschlagen werden. – Natürlich ist die Identifikation durch geordnete Tupel von IDs, die zu ineinander verschachtelten Elementen gehören, nicht die einzige Möglichkeit eines Adressierungsschemas. Man kann auch jedes zu adressierende Element einer Datenbasis separat mit einer eineindeutigen ID versehen. Die hier demonstrierte Vorgehensweise kann aber konzeptionelle Vorteile haben, wie in Abschnitt 2.4 deutlich werden wird.

[4] Wir vereinfachen hier durchgängig, indem wir in vielen Fällen von Elementen sprechen, wo allgemeiner von Knoten im XML-Dokumentenbaum gesprochen werden müsste. Die erforderliche Verallgemeinerung ist aber trivial und trägt nichts zu den hier diskutierten Aspekten bei.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Bei der Analyse der hier exemplarisch gezeigten Vernetzung ist es durchaus möglich, die oben zur Beschreibung von Verweisen auf der Präsentationsebene verwendeten Wiegand'schen Begrifflichkeiten in ihrem Anwendungsbereich entsprechend „auszudehnen".[5] Eine Übertragung dieser handlungstheoretisch fundierten begrifflichen Analyse auf so etwas wie die Datenmodellierung ist im Fall von *elexiko* schon dadurch gerechtfertigt, dass die *lexikografischen* Handlungen direkt in den XML-Dokumenten vorgenommen werden. Die Artikelautoren arbeiten also direkt mit der XML-Repräsentation des Artikels. Natürlich ist es ihnen möglich, die Perspektive des Benutzers einzunehmen und sich beispielsweise die dem XML-Dokument korrespondierenden Onlineansichten anzuschauen. Diese können sich aber jederzeit ändern, beispielsweise um neue Darstellungsaspekte und Funktionalitäten erweitert werden. Insgesamt wird eine vollständige begriffliche Analyse – die hier nicht geleistet werden soll – durch die Existenz von zwei durch komplexe Regeln miteinander verknüpften „Textebenen" für die Produktion bzw. Nutzung / Rezeption wesentlich komplexer.

Wir zeigen nun am eingeführten Beispiel, wie die oben eingeführten Begriffe bestimmten Aspekten des betrachteten XML-Dokuments aus Abb. 2 zugeordnet werden können; dabei ersetzen wir in den verwendeten Termini jeweils den Wortbestandteil *-verweis-* durch *-vernetzung-*.

- Der zur Vernetzungsangabe gehörende *Vernetzungsausgangsbereich* ist das artikelintern adressierte übergeordnete XML-Element, von dem aus auf ein anderes XML-Element vernetzt wird; da hier eine Synonymievernetzung zwischen Lesarten von zwei Artikeln vorliegt, wäre es vielleicht naheliegend, das XML-Element <lesart>[6] mit diesem Bereich zu identifizieren. Korrekterweise ist jedoch <synonymie> der Vernetzungsausgangsbereich, da ja auf *Haushalt* in der Lesart 'Personengruppe' *als Synonym* und nicht z.B. als Partonym vernetzt (und auf der Präsentationsebene verwiesen) werden soll. Die zum genannten Bereich gehörende *Vernetzungsausgangsangabe* besteht aus der für eine eindeutige Identifikation erforderlichen Eigenschaft (z.B. ID-Attributwert oder Elementname) des betreffenden Elements selber sowie der nach dem verwendeten Bijektionsschema *relevanten*, mit IDs versehenen Vorfahren des Elements. Genauer lässt sich die Vernetzungsausgangsangabe mit der standardisierten XML-Abfragesprache XPath beschreiben; im Beispiel handelt es sich um den XPath-Ausdruck **/artikel[@id='1234']//lesart [@id='Verwandte']//synonymie**, der die gewünschte Einer-Knotenmenge ausdrückt. Die Vernetzungsausgangsangabe spezifiziert die *Bezugsadresse* der Vernetzung, die hier aus dem geordneten ID-Tripel ('1234', 'Verwandte', 'synonymie') besteht.[7]
- Die Vernetzungsangabe selber befindet sich innerhalb des Vernetzungsausgangsbereichs[8] (in XML-Terminologie: als Nachkomme an einer nach dem verwendeten XML-Schema dafür vorgesehenen *Vernetzungsposition*). In diesem Fall handelt es sich einfach um einen von mehreren Kindelementen des Ausgangsbereichs. Die

[5] Nur am Rande sei vermerkt, dass ein solcher Umgang mit Begriffen (nicht bloß Termini) nicht dem grundsätzlichen wissenschaftstheoretischen Verständnis von Begriffsbildungen zuwiderläuft, sondern sogar grundlegende Eigenschaft des „Funktionierens" von Begriffen ist; so argumentiert umfassend und mit zahlreichen Beispielen v.a. aus den Natur- und Ingenieurswissenschaften Wilson (2006).

[6] Die hier verwendete verkürzende Bezeichnungsweise sollte hinreichend klar sein; gemeint ist ein konkretes XML-Element des Beispieldokuments, nämlich hier das einzige im Beispiel gezeigte Element, dessen Tagname gleich *lesart* ist.

[7] Im Allgemeinen müssen IDs also keine XML-Attributwerte sein, es kann sich auch z.B. um Elementnamen handeln. – Hier soll ohne Beschränkung der Allgemeinheit davon ausgegangen werden, dass sich mit den Eigenschaften der IDs (z.B. Zahlenbereich) und der Stelligkeit des ID-Tupels eindeutig das zugehörige XML-Element identifizieren lässt.

[8] Dies ist natürlich keine logisch zwingende Entscheidung; die Angabe könnte z.B. auch ein Geschwisterknoten des Ausgangsbereichsknotens sein.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Vernetzungsangabe selber ist hier das <relpartner>-Element, das die *Vernetzungsadressenangabe* enthält, in diesem Falle in Gestalt von zwei Attributwerten, die die Artikel- und Lesart-IDs des synonymen Relationspartners angeben. Als die von der Angabe spezifizierte *Vernetzungsadresse* fassen wir wiederum das zugehörige geordnete Paar (‘5678’, ‘Personengruppe’) von IDs auf. – Weiter unten besprechen wir den Fall, dass die Vernetzungsangabe keine Vernetzungsadressenangabe enthält, sondern über eine eindeutige ID verfügt, aus der in einer externen Tabelle die Vernetzungsadresse ermittelt werden kann.

- Die genannte Vernetzungsadresse ist die Adresse des *Vernetzungszielbereichs*, der sich im hier nicht gezeigten XML-„Zieldokument“ des Wortartikels *Haushalt* befindet; dabei handelt es sich um das <lesart>-Element dieses Artikels mit dem ID-Attribut ‘Personengruppe’.

2. Wie ‘selbstständig’ ist die Mediostruktur in XML-basierten Wörterbüchern?

2.1 Vernetzungsspezifikationen

Das geordnete Paar aus Bezugs- und Vernetzungsadresse, im Beispiel: ([‘1234’, ‘Verwandte’, ‘synonymie’], [‘5678’, ‘Personengruppe’]), stellt gewissermaßen den informationellen, mediostrukturellen Kern der Vernetzungsbeziehung dar und soll hier als (*lexikografische*) *Vernetzungsspezifikation* bezeichnet werden. Das damit adressierte Paar aus Vernetzungsausgangs- und -zielbereich möge nunmehr auch kurz als (*lexikografische*) *Vernetzung* bezeichnet werden; dies entspricht im Wesentlichen, wenn auch mit abweichender Terminologie, der Definition unidirektionaler Vernetzungen als geordneten Paaren aus adressierten Quell- und Zielressourcen bei Müller-Spitzer (2007a: 167).

Aus computerlexikografischer Sicht von besonderer Relevanz ist die Menge der Vernetzungsspezifikationen eines elektronischen Wörterbuchs, die mengentheoretisch-extensional eine Relation ist. Sie kann als Grundlage für wesentliche Konsistenzprüfungen dienen, die für ein automatisiertes Vernetzungsmanagement von Bedeutung sind und nicht sinnvoll in manueller lexikografischer Arbeit erledigt werden können. Hierbei geht es zum einen darum, ob die Vernetzungsadresse in der Datenbasis überhaupt existiert, zum anderen aber auch, ob die genannte Relation oder Teilmengen davon (z.B. die Menge der Synonymierelationen) bestimmte Bedingungen wie Symmetrie[9] oder Transitivität erfüllen.

Da Vernetzungen in der Regel artikelübergreifend sind, liegt es nahe, bei einer XML-basierten Repräsentation der Artikel die Vernetzungsspezifikationen in einer gesonderten Datenstruktur zu speichern, beispielsweise in einer relationalen Datenbanktabelle, die performante Suchen gestattet. Ganz allgemein ist die Frage, inwieweit sich die Mediostruktur eines XML-basierten elektronischen Wörterbuchs unabhängig von der hierarchischen Struktur der Einzelartikel repräsentieren lässt, von computerlexikografischem Interesse, und soll im Folgenden ausführlicher betrachtet werden.

[9] Die Darstellung ist hier wieder etwas vereinfacht: Wenn man die Vernetzungsspezifikationen z.B. auf Synonymie einschränkt, kann man sie verkürzend notieren, ohne den Elementnamen <synonymie> anzugeben; erst dann kann man bezüglich geordneter Paare der Art ([‘1234’,‘Verwandte’], [‘5678’,‘Personengruppe’]) von Symmetrie der Relation sprechen.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

2.2 Vernetzungspositionen und Separabilität der Mediostruktur

Konkret lässt sich zunächst fragen, ob es im Allgemeinen durch eine Reorganisation der XML-Datenbasis eines Wörterbuchs möglich ist, Vernetzungsspezifikationen dergestalt in einer separaten Datenstruktur zu verwalten, dass das Anlegen oder Löschen einer Vernetzung *keine Änderungen am XML-Dokument erforderlich macht*. Diese mögliche Eigenschaft der XML-Datenbasis eines Wörterbuchs sei hier kurz als *Separabilität der Mediostruktur* bezeichnet. Die Antwort auf die gestellte Frage ist negativ; für eine nähere Begründung kommen wir auf den in Wiegands Arbeiten eher unscharf verwendeten Begriff der Verweis- bzw. hier Vernetzungs*position* zurück. In unserem Beispiel sind die Vernetzungsangaben zu den Synonymen zu *Familie* in der Lesart 'Verwandte' eine Menge von Geschwisterknoten, die laut XML-Schema sämtlich Kindelemente des Elements <synonymie> sind und in diesem Sinne eine bestimmte *schemainduzierte generische Vernetzungsposition* gemein haben. Die Reihenfolge der Vernetzungsangaben ist jedoch nicht mehr durch ein Schema vorgebbar. Die *konkrete Vernetzungsposition* der Vernetzungsangaben kann daher im Allgemeinen nicht mechanisch aus der Vernetzungsspezifikation abgeleitet werden. In diesem speziellen Fall ist sie in *elexiko* einfach durch eine alphabetische Sortierung gegeben, so dass es allein aufgrund der Vernetzungsspezifikationen immerhin noch möglich wäre, die konkreten Positionen der Angaben mechanisch zu bestimmen – allerdings nur dann, wenn sämtliche Spezifikationen, die zu einer und derselben schemainduzierten generischen Vernetzungsposition gehören, bekannt sind. Bestünden die Angabebereiche zu den verschiedenen paradigmatischen Relationen (Synonymie, Hyperonymie, …) sämtlich nur aus Verweisen auf alphabetisch geordnete Relationspartner, wäre die Mediostruktur – bzw. zumindest der hier betrachtete Ausschnitt der Mediostruktur – separabel, es wäre sogar möglich, das gesamte Element <paradigmatik> aus dem XML-Schema herauszunehmen und statt dessen die zugehörigen Informationen in einer simplen relationalen Tabelle von Vernetzungsspezifikationen zu speichern.

Der eben eingeführte Begriff der konkreten Vernetzungsposition lässt sich formal genauer fassen: Er wird durch einen XPath-Ausdruck repräsentiert, der auf dem Vernetzungsausgangsbereich operiert und das zur Vernetzungsangabe gehörende XML-Element eindeutig spezifiziert. Diese Präzisierung geht davon aus, dass die Vernetzungsposition nicht durch eine eigene ID erschließbar ist, so dass der XPath-Ausdruck im Allgemeinen mit *Zugriffsindizes* arbeiten muss.[10] In unserem Beispiel aus Abb. 1 / 2 wäre, da *Haushalt* das an dritter Stelle genannte Synonym und der Vernetzungsausgangsbereich das Element <synonymie> ist, die konkrete Vernetzungsposition einfach durch den auf das <synonym>-Element anzuwendenden XPath-Ausdruck **relpartner[3]**, oder, in der ausführlichen XPath-Notation, **child::relpartner[position() = 3]** gegeben.

2.3 Nichtseparable Mediostrukturen I: Die Rolle positionaler Informationen

Wäre die Abfolge der Relationspartner Gegenstand inhaltlich-lexikografischer Entscheidungen und könnte deswegen nicht mechanisch aus den Vernetzungsspezifikationen erschlossen werden, wären die oben gegebenen Bedingungen für Separabilität nicht mehr gegeben; dadurch würde eine getrennte Datenhaltung für die Mediostruktur, ähnlich wie bei den nachfolgenden Beispielen, allerdings nicht unmöglich, aber bereits deutlich komplexer, da man nun beispielsweise den XPath-Zugriffsindex als zusätzliche Information in die Tabelle der Vernetzungsspezifikationen aufnehmen müsste und damit einen Aspekt der Hierarchisierung

[10] Abstrakter formuliert muss der XPath-Ausdruck im Allgemeinen auch entlang der Achsen **preceding** und **following** navigieren.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

und Abfolge von Elementen der XML-Struktur zusätzlich in die separate Datenstruktur auslagern müsste. Ein Redaktionssystem müsste dafür sorgen, dass diese Aufteilung der Datenhaltung in der Benutzeroberfläche für den Lexikografen transparent ist, und müsste zudem die Konsistenz der Datenhaltung sicherstellen, etwa dadurch, dass geprüft wird, ob die Zugriffsindizes für zusammengehörige konkrete Vernetzungspositionenen wirklich bei 1 beginnen und fortlaufend sind. Dieses neue Konsistenzproblem wäre ein Artefakt der getrennten Repräsentation von Mikro- und Mediostrukturen.

Noch deutlicher wird die Problematik anhand des in Abbildung 3 gezeigten weiteren Ausschnittes aus der Paradigmatik des Artikels *Familie*, zu dem in Abbildung 4 ein wiederum deutlich vereinfachter Ausschnitt des XML-Dokuments gezeigt wird. In *elexiko* können mehrere Relationspartner, für die ein gemeinsamer Korpusbeleg vorliegt, zu einer <beleg-gruppe> zusammengefasst sein. Überdies können sämtliche Relationspartner bzw. Beleggruppen für eine gegebene Sinnrelation zu mehreren „Relationspartnergruppen“ zusammengefasst sein; so wird hier zwischen Partonymen zu ‘Familie im engeren Sinne’ und zu ‘Familie im weiteren Sinne’ unterschieden. Um diese mehrfache hierarchische Untergliederung nach Gruppen von Vernetzungsangaben in eine separate Datenstruktur auslagern zu können, würde man bereits ein recht ausgefeiltes System von Positions- und/oder Gruppenindizes benötigen, dessen einziger Zweck zudem darin bestünde, Lagebeziehungen zwischen Knoten in XML-Bäumen zu reproduzieren. Man kann in diesem Zusammenhang ein Konzept von *schwacher Separabilität* der Mediostruktur definieren, das genau dann zutrifft, wenn es möglich ist, Vernetzungsspezifikationen *zusammen mit positionalen Indizes für die zugehörigen Vernetzungsangaben* in einer separaten Tabelle so zu speichern, dass das Anlegen oder Löschen einer Vernetzung ohne Veränderung des XML-Dokuments möglich ist.

Abb. 3: Weiterer Ausschnitt aus der Onlinepräsentation des *elexiko*-Artikels *Familie* in der Lesart ‘Verwandte’ (http://www.owid.de/artikel/3920/Verwandte; Stand Oktober 2012), Angabebereich „Sinn-verwandte Wörter“

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

```
<artikel id="1234">
  <lemmazeichen>Familie</lemmazeichen>
    …
    <lesart id="Verwandte">
      …
      <paradigmatik>
        …
        <partonymie>
          <relpartner-gruppe titel="Familie im engeren Sinne">
            …
            <relpartner art-id=… lesart-id=…>Elternteil</relpartner>
              …
            <beleg-gruppe>
              <relpartner art-id=… lesart-id=…>Mutter</relpartner>
              <relpartner art-id=… lesart-id=…>Tochter</relpartner>
              <beleg> … </beleg>
            </beleg-gruppe>
             …
          </relpartner-gruppe>
          <relpartner-gruppe titel="Familie im weiteren Sinne">
            …
          </relpartner-gruppe>
        </partonymie>
        …
    </paradigmatik>
    …
  </lesart>
  …
</artikel>
```

Abb. 4: Vereinfachter Ausschnitt aus dem XML-Dokument des *elexiko*-Wortartikels ‘Familie’: Vernetzung von der Lesart ‘Verwandte’ auf eine Gruppe von partonymen Relationspartnern. Die gegenüber Abb. 2 neu hinzugekommenen XML-Elemente sind fett gesetzt

2.4 Nichtseparable Mediostrukturen II: Angabezusätze

Eine weitere Komplikation ergibt sich offenbar durch die bislang ignorierten *Angabezusätze* zu Vernetzungsangaben, wie etwa Korpusbelege und weitere Hinweise (vgl. erneut Abb. 1), die sich inhaltlich ausschließlich auf die Vernetzung beziehen. In der hier verwendeten, technischen Definition von Vernetzungsspezifikation sind solche Zusätze nicht berücksichtigt: Wenn man eine lexikografische Datenbasis ohne vernetzungsbezogene Angabezusätze, deren Mediostruktur separabel oder schwach separabel ist, um solche Zusätze anreichert, bleibt die Mediostruktur separabel bzw. schwach separabel. Dies lässt sich an den obigen Beispielen leicht nachvollziehen: Es ist beim Vorhandensein von vernetzungsbezogenen Angabezusätzen zwar nicht mehr möglich, durch Auslagern der Vernetzungsspezifikationen das gesamte <paradigmatik>-Element zu streichen, aber man kann – trivialerweise – immer noch die Vernetzungsadressenangaben (Attribute) aus den <relpartner>-Elementen entfernen. Aus lexikografischer wie informatischer Sicht ist ein solches Vorgehen nicht überzeugend, weil nunmehr vernetzungsbezogene Informationen in wenig einleuchtender Weise auf XML-Dokument und externe Tabelle der Vernetzungsspezifikationen verteilt sind. Inhaltlich gesehen lassen sich Vernetzungen auch als ternäre Beziehungen auffassen – es wird a) von etwas b) mit etwas c) auf etwas vernetzt; die Angabezusätze sind ein Aspekt des „mit etwas“. Man kann daher ein Konzept *erweiterter Vernetzungsspezifikationen* einführen, die geordnete Paare aus einer

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Vernetzungsspezifikation sowie einer Spezifikation der ausschließlich zur jeweiligen Vernetzung gehörigen Angabezusätze sind. Letztere müssten in komplexeren Fällen, so wie sie in *elexiko* vorliegen, als XML-Fragment oder auch, falls überhaupt möglich, als komplexes relationales Äquivalent eines solchen Fragments repräsentiert werden, wenn man versuchen möchte, diese Spezifikationen als separate Datenstruktur abzuspeichern. Solange solche Zusätze strikt jeweils genau einer Vernetzungsangabe zugeordnet sind, ist ein solches Auslagern aller vernetzungsbezogenen Informationen in erweiterte Vernetzungsspezifikationen dergestalt möglich, dass wiederum das Anlegen oder Löschen von Vernetzungen einschließlich der Angabezusätze ohne Veränderungen am XML-Dokument möglich ist. Dies lässt sich im hier eingeführten terminologischen Stil als *erweiterte Separabilität* bzw. *erweiterte schwache Separabilität der Mediostruktur* bezeichnen, je nachdem, ob auch positionale Indizes mit den Vernetzungsspezifikationen abgespeichert werden müssen. Die Mediostruktur von *elexiko* ist jedoch nach der eben absichtlich so restriktiv eingeführten Definition nicht einmal erweitert schwach separabel, da Angabezusätze sich häufig auf mehr als eine Vernetzung beziehen, wie sofort aus Abb. 1 und Abb. 3 ersichtlich ist. So ist es gerade das charakteristische Merkmal von Beleggruppen, dass sie einen gemeinsamen Korpusbeleg haben; Vernetzungen in Relationspartnergruppen haben eine gemeinsame erläuternde Überschrift; usw. Natürlich kann man auch solche übergreifenden Angabezusätze in eine separate Struktur auslagern; der Preis dafür wäre dann jedoch, dass man die ausgelagerten Vernetzungsspezifikationen analog den Strukturen des ursprünglichen XML-Dokuments hierarchisch gliedern und die auszulagernden Zusätze dann an Gruppen von Vernetzungsspezifikationen zuweisen müsste – womit die grundsätzliche Idee, die hierarchische interne Gliederungsstruktur eines Wörterbuchartikels von den mediostrukturellen Beziehungen zwischen Artikeln zu trennen, endgültig ad absurdum geführt wäre.

Dort, wo keine hinreichend einfache Form von Separabilität der Mediostruktur gegeben ist, steht nun noch ein anderer Weg offen, um die u.U. gewünschte informationelle Trennung zu erreichen: Man kann zunächst die Vernetzungsangaben selber in das allgemeine Adressierungsschema mit einbeziehen, also durch IDs oder ID-Tupel identifizierbar machen. In einem zweiten Schritt entfernt man dann lediglich die Vernetzungsadressenangaben, in unseren Beispielen die Attribute @art-id und @lesart-id, und lagert die reine Vernetzungsinformation in eine *strukturelle Vernetzungsspezifikation* aus, die einfach ein geordnetes Paar aus der Adresse der Vernetzungsangabe und der Vernetzungsadresse ist. Im Beispiel aus Abb. 1 hätten wir dann statt eines Vernetzungsangabe-Elements **<relpartner art-id="5678" lesart-id="Personengruppe">…</relpartner>** ein Element **<relpartner id="rel_haushalt">…</relpartner>**; in einer separaten Tabelle würde dann die strukturelle Vernetzungsspezifikation (['1234', 'Verwandte', 'synonymie', 'rel_haushalt'], ['5678', 'Personengruppe']) abgespeichert. Die soeben erneut illustrierte Adressierungstechnik mittels geordneter ID-Tupel hat in der hier gezeigten Form den Vorteil, dass eine solche strukturelle Vernetzungsspezifikation immer zugleich auch eine gewöhnliche Vernetzungsspezifikation ist und mithin Auskunft über den Vernetzungsausgangsbereich liefert, was für Konsistenzprüfungen nützlich ist.

Im Ergebnis ist die Auslagerung nur der strukturellen Vernetzungsspezifikationen eine in allen Fällen verfügbare, computerlexikografisch recht saubere Lösung, denn sie lässt die hierarchische, XML-basierte Datenmodellierung innerhalb der Wortartikel intakt und separiert nur dasjenige Minimum an artikelübergreifender mediostruktureller Information, das für Korrektheits- und Konsistenzprüfungen benötigt wird, nämlich die Vernetzungsadresse und ihre Zuordnung zu einer konkreten Vernetzungsposition im Ausgangsartikel. Solange man keine Redundanzen in Kauf nehmen möchte, bedeutet aber auch hier die Auslagerung, dass aus den einzelnen Dokumenten die zugehörige HTML-Präsentation nicht vollständig ermittelt werden kann.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

2.5 Zusammenfassung: Typen von Separabilität der Mediostruktur

Es hat sich gezeigt, dass die Möglichkeit, auf sinnvolle Weise mediostrukturelle Informationen aus den XML-Dokumenten der Wortartikel auszulagern, nur unter recht speziellen Bedingungen gegeben ist. Wenn man aus der Kenntnis von Bezugs- und Zieladresse einer Vernetzung mechanisch die konkrete Position der Vernetzungsangabe im XML-Dokument bestimmen kann, kann man dieses Adressenpaar, oben Vernetzungsspezifikation genannt, in einer eigenen Datenstruktur verwalten (Separabilität der Mediostruktur); ist die konkrete Position hingegen selber Gegenstand genuin lexikografischer Entscheidungen, kann allenfalls eine Kombination aus Vernetzungsspezifikation und positionellen Indizes ausgelagert werden (schwache Separabilität), was bereits zu einer wenig wünschenswerten doppelten Repräsentation von hierarchisch-positionalen Informationen führt. – Gibt es auf die Vernetzungen bezogene Angabezusätze, sind beide genannten Formen von Separabilität wenig relevant, da man sinnvollerweise die (allerdings möglicherweise selber komplexe XML-Teilbäume bildenden!) Angabezusätze mit auslagern können sollte; ist dies der Fall, sprechen wir hier von erweiterter bzw. erweiterter schwacher Separabilität – aber nur dann, wenn es keine auf mehrere Vernetzungen gleichzeitig bezogenen Zusätze gibt. Anderenfalls kommt es zwangsläufig zu einer doppelten Repräsentation der gesamten wortartikelinternen, durch die XML-Struktur vorgegebenen hierarchischen Gruppierung der Vernetzungsangaben. – In allen Fällen steht jedoch die Option zur Verfügung, nur die strukturellen Vernetzungsspezifikationen eines Artikels auszugliedern, indem man eineindeutige Adressen an die Vernetzungsangaben selbst vergibt.

3. Anwendungen und Schlussfolgerungen

3.1 Computerlexikografische Behandlung von Vernetzungen in *elexiko*

Die Vernetzungen in *elexiko*-Wortartikeln sind, wie gesehen, ein Beispiel für den Fall, dass die Mediostruktur eines Wörterbuchs aufgrund ihrer Komplexität allenfalls schwach separabel ist und die computerlexikografischen Kosten einer Ausgliederung ersichtlich höher als ihr Nutzen wären. Da die lexikografische Bearbeitung der XML-Instanzen von *elexiko* ursprünglich ausschließlich mit einem handelsüblichen XML-Editor durchgeführt wurde, war es sinnvoll zu fordern, dass die Vernetzungsadressenangaben lokal in den Instanzen erscheinen, so dass auch die Speicherung struktureller Vernetzungsspezifikationen außerhalb der XML-Instanzen keine Option war. Mittlerweile arbeitet das *elexiko*-Projekt mit einem im Haus entwickelten Editor-Plugin für das Vernetzungsmanagement (Meyer 2011).[11] Im Rahmen eines konservativen Herangehens wurde entschieden, die grundsätzlichen Datenstrukturen und die mit ihnen verbundenen Arbeitsabläufe nicht umzugestalten. Um die ein- und ausgehenden Vernetzungen eines im XML-Editor bearbeiteten Artikels aufzufinden sowie Integrität und Konsistenz seiner Vernetzungen zu prüfen, führt das Plugin folgende Prozesse durch:

- Das Dokument im Editor wird geparst und alle Vernetzungsangaben sowie die allgemeine Artikelstruktur (Lesarten, Adressen) werden ermittelt;

[11] Auch wenn grundsätzlich schon aus Wirtschaftlichkeitsgründen die Verwendung eines handelsüblichen Wörterbuch-Redaktionssystems wünschenswert ist, kann es doch – *pace* de Schryver (2011) – gute Gründe für eine gegenteilige Entscheidung geben. Im Fall von *elexiko* sind bereits die komplexe, schon vorhandene Hardware- und Software-Infrastruktur und die in lexikografischen Forschungsprojekten zu erwartenden häufigen Umstrukturierungen des XML-Schemas sowie die Existenz von Vernetzungen zwischen verschiedenen Wörterbüchern mit unterschiedlichen Schemata solche Gründe.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

- alle Artikel, auf die sich die gefundenen Vernetzungsadressenangaben beziehen, werden, sofern wirklich vorhanden, aus der Datenbank geholt und geparst, um die Vernetzungsadresse zu prüfen;
- mit einer XPath-basierten Suche auf einem für XML optimierten Volltextindex werden alle Artikel in der Datenbank gesucht und ausgelesen, die eine Vernetzung auf das Dokument im Editor haben;
- diese Artikel werden ebenfalls geparst, um die zugehörigen Vernetzungsspezifikationen zu ermitteln;
- abschließend werden alle gefundenen Informationen miteinander abgeglichen und die Resultate in zwei Tabellen angezeigt, die für ein- bzw. ausgehende Vernetzungen die Vernetzungsspezifikationen und Konsistenzstatusinformationen liefern.

Aufgrund der relativ kleinen Menge an bearbeiteten Wortartikeln (< 2000) in *elexiko* ist dieses Vorgehen trotz der bemerkenswert komplexen Vernetzungsstruktur der Artikel und der hohen Kosten einer Volltextsuche für den lexikografischen Alltag hinreichend performant; eine Analyse der beschriebenen Art nimmt wenige Sekunden in Anspruch. Würde man ein skalierbareres Verfahren benötigen, wäre dennoch kein Auslagern von mediostruktureller Information erforderlich, da grundsätzlich noch ein anderes Verfahren zur Verfügung steht: Man kann die für Integritäts- und Konsistenzprüfungen benötigten Vernetzungsspezifikationen einfach in eine separate Tabelle *kopieren*, die über einen Datenbanktrigger bei jeder Änderung eines Artikels geprüft und ggf. aktualisiert wird. Die separate Tabelle fungiert dann als schneller relationaler Cache der benötigten Informationen (vgl. Joffe/Schryver/Prinsloo 2003; Meyer / Müller-Spitzer 2010).

3.2 Zusammenfassung

Vor dem Hintergrund einer relationalen lexikografischen Datenmodellierung schrieben Blumenthal et al. im Jahre 1988:

> Auf der Ebene der konzeptionellen Datenmodellierung gibt es Beziehungen. Beziehungen sind ungerichtet, d.h. in beiden Richtungen in gleicher Weise zugreifbar. Dies liegt daran, daß auf der konzeptionellen Ebene die Asymmetrie von Verweisursprung und Verweisziel im Hinblick auf ihre ‚Repräsentationsbedürftigkeit' verschwindet, da man dort direkt die Beziehung zwischen Verweisursprung und Verweisziel modelliert. (Blumenthal/Lemnitzer/Storrer 1988: 356)

Der vorliegende Beitrag ist ein Beleg dafür, dass durch die Verwendung von XML-basierten Repräsentationen die genannte Asymmetrie wieder in die Computerlexikografie Einzug gehalten hat; der „Verweisursprung" ist in den XML-Instanzen, wie auch auf der Präsentationsebene, „nicht symbolisch repräsentiert, sondern qua Lokalität … faktisch gegeben" (ebd.: 145). Der Versuch, Vernetzungen doch relational zu modellieren, führt, wie gezeigt, in den meisten Fällen zu unbrauchbaren Ergebnissen. Dies ist zunächst eine sehr unbefriedigende Diagnose, denn durch den Verzicht auf eine solche Modellierung kehren viele Probleme zurück, die man mit der Verwendung eines relationalen Datenbanksystems lösen konnte. Der einzige sinnvolle Ausweg ist eine Doppelstrategie, die es bei einer ausschließlich XML-basierten Datenhaltung und lokal in den Instanzen repräsentierten Vernetzungsadressenangaben belässt und mediostrukturelle Informationen in einem stets aktuell gehaltenen relationalen Repräsentationsformat vorhält, das performante Prüfungen und Suchvorgänge ermöglicht.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

4. Literatur

Blumenthal, Andreas / Lemnitzer, Lothar / Storrer, Angelika (1988): Was ist eigentlich ein Verweis? Konzeptionelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung. In: Harras, Gisela (Hg.): Das Wörterbuch. Artikel und Verweisstrukturen. (= Jahrbuch 1987 des Instituts für deutsche Sprache). Düsseldorf/Bielefeld, S. 351-373.

Haß, Ulrike (Hg.) (2005): Grundfragen der Elektronischen Lexikographie: *elexiko* – Das Online-Informationssystem zum deutschen Wortschatz. Berlin / New York.

Joffe, David / de Schryver, Gilles-Maurice / Prinsloo, Daniel Jacobus (2003): Computational features of the dictionary application "TshwaneLex". In: Southern African Linguistics and Applied Language Studies 21(4), S. 239-250.

Kammerer, Matthias (1998): Hypertextualisierung gedruckter Wörterbuchtexte: Verweisstrukturen und Hyperlinks. Eine Analyse anhand des FRÜHNEUHOCHDEUTSCHEN WÖRTERBUCHES. In: Storrer, Angelika / Harriehausen, Bettina (Hg.): Hypermedia für Lexikon und Grammatik. Tübingen, S. 145-171.

Klosa, Annette (Hg.) (2011): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur Deutschen Sprache 55). Tübingen.

Meyer, Peter (2011): vernetziko: a cross-reference management tool for the lexicographer's workbench. In: Kosem, Iztok / Kosem, Karmen (Hg.): Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex2011, Bled, Slowenien, 10-12 November 2011. Ljubljana, S. 191-198. Internet: http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-25.pdf. (Stand: Oktober 2012).

Meyer, Peter / Müller-Spitzer, Carolin (2010): Consistency of sense relations in a lexicographic context. In: Mititelu, Verginica Barbu / Pekar, Viktor / Barbu, Eduard (Hg.): Proceedings of the workshop "Semantic Relations. Theory and Applications", 18 May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta. Internet: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf. (Stand: Oktober 2012).

Müller-Spitzer, Carolin (2007a): Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. In: Hermes 38, S. 137-171.

Müller-Spitzer, Carolin (2007b): Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis. (= Studien zur Deutschen Sprache 42). Tübingen.

Schryver, Gilles-Maurice de (2011): Why opting for a dedicated, professional, off-the-shelf dictionary writing system matters. In: Akasu, Kaoru / Uchida, Satoru (Hg.): ASIALEX 2011 Proceedings. LEXICOGRAPHY: Theoretical and practical perspectives. Papers submitted to the Seventh ASIALEX Biennial International Conference. Kyoto Terrsa, Kyoto, Japan, August 22-24, 2011. Kyoto, S. 647-656.

Wiegand, Herbert Ernst (1996): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Zettersten, Arne / Pedersen, Viggo Hjørnager (Hg.): Symposium on Lexicography VII. Proceedings of the Seventh International Symposium on Lexicography, May 5-6, 1994, at the University of Copenhagen. (= Lexicographica, Series Maior 76). Tübingen, S. 11-43.

Wiegand, Herbert Ernst (2002): Altes und Neues zur Mediostruktur in Printwörterbüchern. In: Lexicographica 18, S. 168-252.

Wilson, Mark (2006): Wandering significance. An essay on conceptual behaviour. Oxford / New York.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Das Deutsche Rechtswörterbuch im Netz

Almuth Bedenbender, Heidelberger Akademie der Wissenschaften

Abstract

The "Deutsches Rechtswörterbuch" (DRW) is a dictionary of historical German and older West Germanic legal terms, spanning the time from the first word occurrences in written sources until the beginning of the 19th century. The project started at the end of the 19th century. So far twelve volumes have been published, containing more than 90,000 main entries (*Aachenfahrt – schwedisch*). While the work is still primarily based on an archive of approx. 2.5 million handwritten index cards, the DRW team has used a system of several interlinked databases (including a collection of digitised source texts) as a tool for the lexicographical process since 1993.

"DRW Online" (www.deutsches-rechtswoerterbuch.de) reflects these internal databases in some ways. There are several indices providing easy access not only to the main entries (by either headwords or by word forms in their original spelling), but also to the full text of the source citations and the explanations. Additionally, DRW Online comprises a bibliographical database of the DRW sources, digitised source texts (partly as searchable full text, partly as images) and information about online resources.

The dictionary text is enriched by several types of links: Entries are connected to other entries and to bibliographical data, reflecting the cross referencing of the print version. In addition to this, there are links to (possibly) corresponding entries in other historical dictionaries. An added functionality for the user who wants to evaluate the context of a source citation is the direct linking of citations with their (publicly available) online source texts – up to now, more than 200,000 citations (i.e. more than 40 %) have been linked to their respective pages. Furthermore, searching for additional source material within the database is facilitated by means of a button which combines the look-up of all spelling forms occurring in an entry. Finally, there are "implicit links": By double clicking on a word, the reader starts a search for this word form in several indices of DRW Online.

Das Deutsche Rechtswörterbuch (DRW), eines der großen diachronen Wörterbuchprojekte, wurde in der Zeit der Zettelkästen begründet, und auch heute noch ist das Wörterbucharchiv mit ca. 2,5 Millionen Belegzetteln die Hauptarbeitsgrundlage bei seiner Erstellung. Schon seit über zwanzig Jahren wird dieses Material aber durch elektronisch vorliegende Quellen ergänzt, und aufgrund der großen Retrodigitalisierungsprojekte steht ein immer größerer Teil des DRW-Quellenkorpus auch im WWW zur Verfügung. Es ist ein besonderes Anliegen der Forschungsstelle, auch in der Onlineversion des Deutschen Rechtswörterbuchs (www.deutsches-rechtswoerterbuch.de) die Verlinkung von den Belegen zu ihrem jeweiligen Quellenkontext herzustellen, um so nicht nur eine bessere Überprüfbarkeit zu gewährleisten, sondern auch für weiterführende Fragen eine Hilfestellung zu bieten. Das ist auch die Intention bei verschiedenen weiteren Zugriffsstrukturen und Vernetzungen, in denen DRW Online über das gedruckte Wörterbuch hinausgeht. Im Folgenden sollen zunächst das Projekt und die intern genutzte lexikographische Datenbank kurz vorgestellt und dann die verschiedenen Möglichkeiten, die die Onlineversion bietet, näher beschrieben werden.

Das Deutsche Rechtswörterbuch hat einen wesentlich umfangreicheren Gegenstand, als sein Titel vermuten lässt. Es beschreibt, wie es im Untertitel heißt, die „ältere deutsche Rechtssprache". Behandelt wird der Zeitraum vom Beginn der schriftlichen Überlieferung bis zum

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

frühen 19. Jahrhundert.[1] „Deutsch“ ist im Sinne Jacob Grimms als Westgermanisch zu verstehen, wobei von den anderen Sprachen nur die älteren Sprachstufen berücksichtigt werden. Neben dem Hoch- und Niederdeutschen der verschiedenen Epochen werden also insbesondere auch Altenglisch, Altfriesisch und Mittelniederländisch behandelt. „Rechtssprache“ geht über die juristische Fachsprache hinaus; gerade für die ältere Zeit werden z.B. auch literarische Werke ausgewertet. Das Projekt wurde Ende des 19. Jahrhunderts initiiert und spiegelt insbesondere in seiner gemeinsamen Behandlung der verschiedenen westgermanischen Sprachen damalige Vorstellungen wider;[2] aus heutiger Perspektive ergibt sich aber gerade dadurch die Möglichkeit einer Zusammenschau der Rechts- und Kulturgeschichte verschiedener europäischer Länder. Bisher sind über 90.000 Wortartikel von *Aachenfahrt* bis *schwedisch* (Band I bis XII) erschienen; der geplante Gesamtumfang beträgt 16 Bände mit ca. 120.000 Artikeln.

Seit 1993 werden die DRW-Artikel mit dem Datenbanksystem FAUST erstellt, einem Programm, das eigentlich für die Objektdokumentation in Archiven, Museen und Bibliotheken gedacht ist, sich aber insbesondere durch seine flexiblen Datenstrukturen und die vielfältigen Verknüpfungs- und Recherchemöglichkeiten auch für die Arbeit in der Forschungsstelle sehr bewährt hat.[3] So sind bei Verweisen zwischen Artikeln die Verweisziele per Mausklick sofort zu erreichen, ebenso auch die Informationen im Quellenverzeichnis. Die Belegzitate und Worterklärungen sind nicht nur durchsuchbar, sondern stehen auch bei der Arbeit an den Artikeln als eine Art Kontexthilfe zur Verfügung. Es lässt sich z.B. bei der Formulierung einer Erklärung per Doppelklick auf ein Wort feststellen, inwieweit es der üblichen Beschreibungssprache im DRW entspricht, und bei einem unverständlichen Wort in einem Belegzitat kann ein Doppelklick oft zu weiteren Stellen und teilweise auch zum passenden Wortartikel führen.

Neben dem Wörterbuchtext gibt es als weitere Datenbestände das Quellenverzeichnis mit bibliographischen Angaben und sonstigen Informationen über die Texte, ein Textarchiv mit Quellen im Volltext sowie eine Bilddatenbank. Letztere wird im DRW dafür genutzt, die Belegzitate mit Faksimiles der entsprechenden Buchseiten zu verknüpfen, so dass bei Bedarf ein Zitat schnell im originalen Zusammenhang überprüft werden kann. Die Arbeit am Deutschen Rechtswörterbuch kann keineswegs nur auf der Basis der Belegzettel erfolgen, auf denen in vielen Fällen nur eine Fundstelle mit Datierung, aber kein Belegtext angeführt wird. Selbst wenn die Stelle auf dem Zettel tatsächlich zitiert wird, ist nicht nur unsicher, wie genau der Text übernommen wurde, sondern zum genauen inhaltlichen Verständnis ist oft ein größeres Textumfeld erforderlich. Deshalb ist ein Rückgriff auf das Buch bzw. das Faksimile nicht nur bei der Belegaufnahme in die Datenbank, sondern auch bei der weiteren Überarbeitung der Artikelstrecken immer wieder erforderlich.

Durch eine zugehörige Programmiersprache ist es möglich, in FAUST eigenentwickelte Zusatzfunktionen zu integrieren, die insbesondere der Prüfung auf Fehler sowie der Entlastung der Mitarbeiter von automatisierbaren Aufgaben dienen. Dadurch kann man z.B. ohne weiteren Aufwand eine Verknüpfung zu schon vorhandenen Faksimiles herstellen oder auch die URLs zahlreicher online verfügbarer Digitalisate seitengenau einfügen.

---

[1] Die obere Zeitgrenze war zunächst nicht ganz fest und wurde im Laufe des Projekts verschiedentlich geändert. Für die aktuelle Arbeit werden nur Quellen bis zum Jahr 1815 berücksichtigt; Kompositen werden bis auf Ausnahmen nur noch behandelt, wenn sie schon vor dem 18. Jahrhundert belegt sind – vgl. dazu Dickel/Speer (1979: 30f.). Die nach diesen Kriterien ausgeschlossenen Wörter werden inzwischen allerdings in der DRW-Datenbank mit Jahr und Fundstelle des Erstbelegs sowie der Zahl der Belegzettel dokumentiert; diese Informationen stehen auch online zur Verfügung.

[2] Vgl. über die Vorstellungen der Gründungsväter des Deutschen Rechtswörterbuchs Deutsch (2010: 23-25).

[3] Vgl. zur Entwicklung des EDV-Einsatzes und zur Arbeit mit FAUST in der Forschungsstelle Speer (1994) sowie Lill (1998) und Lemberg (2001).

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

1997 wurde mit der Entwicklung einer ersten Internetversion des Deutschen Rechtswörterbuchs begonnen.[4] Im Rahmen eines DFG-Projekts konnten dann die alten Bände des Wörterbuchs in die Datenbank übertragen werden, so dass inzwischen fast der gesamte bisher gedruckte Artikelbestand – unter Ausschluss des jeweils zuletzt erschienenen Doppelhefts – kostenlos online zugänglich ist. Bei der Eingabe der alten Bände in die Datenbank waren an verschiedenen Punkten Anpassungen erforderlich, z.B. eine Vereinheitlichung der Siglen und eine Wiedergabe aller Lemmata in voll ausgeschriebener Form. Darüber hinaus wurden dort, wo Fehler zutage traten, diese stillschweigend verbessert, so dass die Volltextfassung von DRW Online den Druck nicht ganz exakt wiedergibt. Sie ist aber mit einem Faksimile des Drucks verknüpft, so dass bei Bedarf von jedem Artikel aus auf die entsprechende Seite zugegriffen werden kann.

Seit 2004 beruht auch DRW Online auf einem Datenbanksystem. Dadurch können insbesondere verschiedene Zugriffsmöglichkeiten angeboten werden, die über die alphabetische Lemmasortierung hinausgehen. Wie in der internen FAUST-Datenbank lässt sich z.B. in den Belegzitaten recherchieren. Diese sind freilich nicht lemmatisiert. Man muss also alle Schreibformen angeben (oder durch Trunkierung zusammenfassen), die berücksichtigt werden sollen. Da die Wörterbuchbelege in vielen Fällen recht prägnante Textausschnitte sind, ist die Wahrscheinlichkeit nicht gering, dass sie auch für Recherchen zu anderen Wörtern von Interesse sein können. Auf diese Weise lassen sich z.B. auch Materialsammlungen durchführen zu Wörtern, zu denen es bisher noch keinen Artikel im Deutschen Rechtswörterbuch gibt (vgl. Abb. 1).

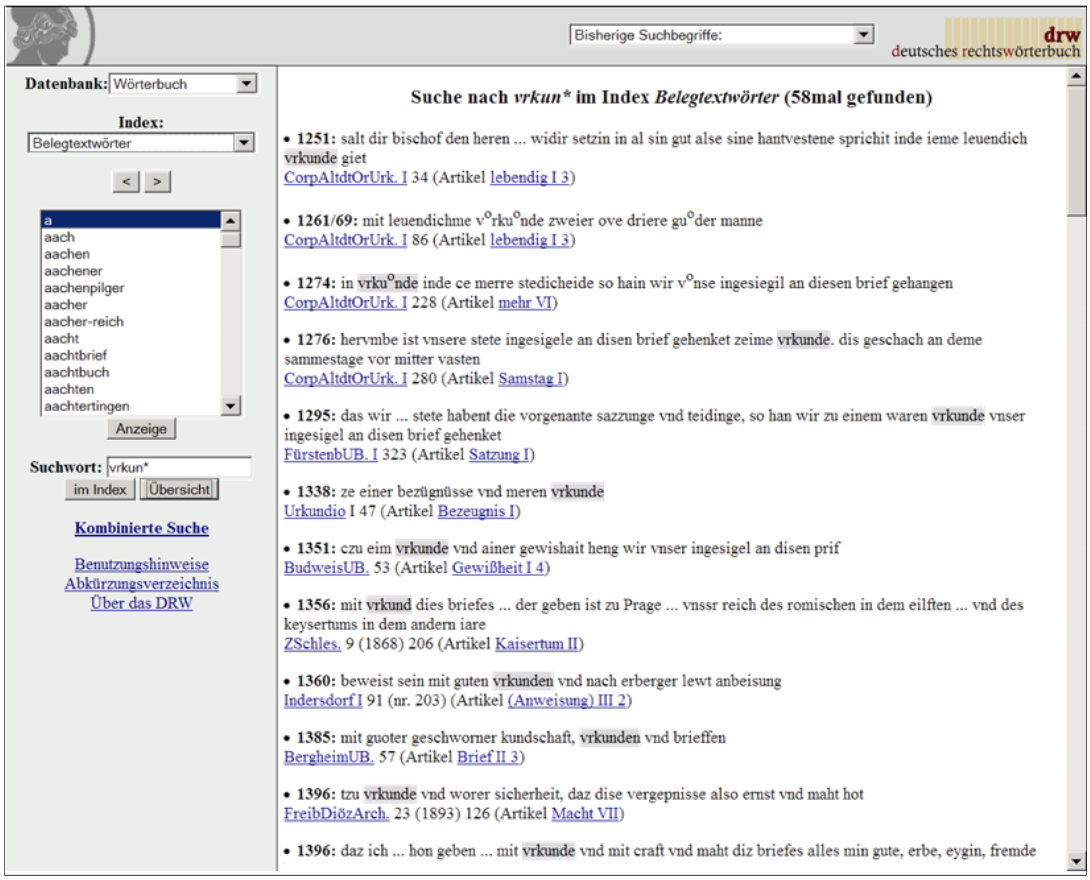

Abb. 1: Suche nach „vrkund*“ in den Belegzitaten

[4] Vgl. Lemberg/Petzold/Speer (1998).

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Etwas einfacher kann es sein, wenn ein DRW-Artikel existiert und man an weiteren Belegen zu einem Wort interessiert ist. Im Wörterbuch kann schon aus Platzgründen, aber auch im Interesse der Prägnanz vor allem bei häufig vorkommenden Wörtern nur ein kleiner Teil der vorhandenen Belegstellen wiedergegeben werden. Die Onlineversion kann zwar bei Belegen, die im Druck nur als Fundstelle ohne Zitat oder nur als Verweis auf ein Zitat in einem anderen Artikel erscheinen, das in der Datenbank gespeicherte Zitat bringen, aber ansonsten folgt die Belegauswahl der gedruckten Version. Es gibt allerdings sicherlich viele Fragestellungen, für die das nicht ausreicht. So lässt sich daraus z.B. nicht erkennen, wie typisch bestimmte Wortkombinationen sind, da der Artikel formelhafte Wendungen zwar berücksichtigt, zugleich aber auch ein breites Spektrum abbilden soll, so dass gerade die häufigeren Fälle nur in starker Auswahl angeführt werden. Hier bietet sich also wiederum die Recherche im Gesamtbestand der Belegzitate des DRW an. Trotz der bisher nicht möglichen automatischen Lemmatisierung der Texte bietet die Onlineversion Unterstützung für eine formenübergreifende Suche: Da die Schreibformen des jeweiligen Stichworts in den Belegzitaten gekennzeichnet sind, wird automatisch eine übergreifende Recherche konstruiert, die all diese Formen berücksichtigt. Falls sich ihnen weitere Belegzitate zuordnen lassen, wird am Ende eines Artikels ein Button gezeigt, über den auf diese Stellen zugegriffen werden kann. Das bedeutet freilich nicht, dass die gefundenen Stellen wirklich zu dem entsprechenden Wort gehören. Es gibt nicht nur Homonyme auf der Ebene der Lemmata – gerade flektierte Wortformen lassen sich ohne Prüfung des syntaktischen Zusammenhangs zunächst einmal oft verschiedenen Wortartikeln zuordnen. Zudem bedingt das sprachenübergreifende Korpus des Deutschen Rechtswörterbuchs, dass manche Schreibformen in ganz unterschiedlichen Wortartikeln auftauchen. Deshalb gibt es neben dem Button für die „Zusatzrecherche“ (siehe Abb. 2) noch ein Eingabefeld, in dem man bei Bedarf bestimmte Formen ausschließen kann.

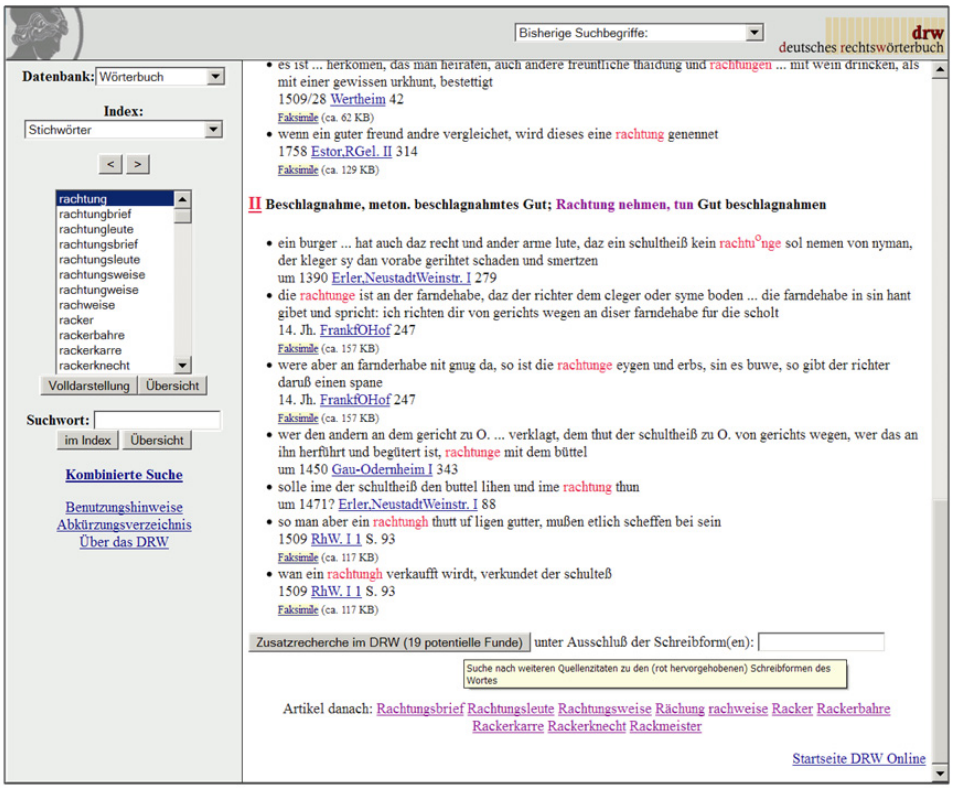

Abb. 2: Button für die Zusatzrecherche über die Schreibformen im Artikel

Die Schreibformenliste steht darüber hinaus auch als eigener Index zur Verfügung, der eine Hilfe bieten kann, wenn die Lemmatisierung eines Wortes unklar ist.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Auch in den Worterklärungen kann eine Volltextrecherche durchgeführt werden. In den Erklärungstexten wird zwar kein normiertes Vokabular verwendet oder gar eine Thesauruserschließung o.Ä. geboten, gleichwohl besteht bei Wörtern, deren Verwendung in einem bestimmten Sachzusammenhang naheliegt, eine recht hohe Wahrscheinlichkeit, dass sie auch in den diesbezüglichen Worterklärungen vorkommen. Als Beispiel sei eine Suche nach „Handschlag“ angeführt (vgl. Abb. 3).

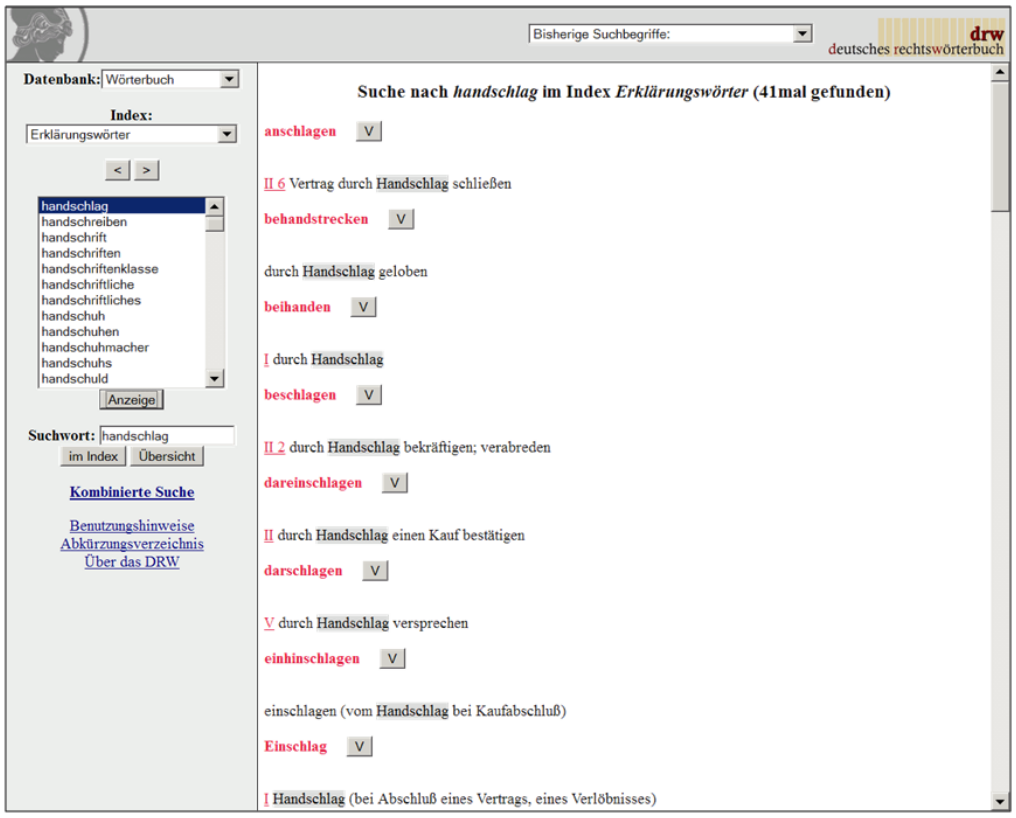

Abb. 3: Suche nach „Handschlag“ in den Worterklärungen

DRW Online ist in einem starken Maße hypertextualisiert. Dies betrifft zunächst einmal natürlich die Verweise auf andere Artikel und die Quellensiglen. Und um die Wortartikel in den Kontext der jeweiligen Artikelstrecke einzubetten, gibt es nicht nur die neben dem Hauptbereich stehende Indexliste, sondern auch vor und nach dem jeweiligen Artikel Links zu den zehn Artikeln davor und danach. Wie bei den Verweiswörtern sind diese Links mit einem Tooltip versehen, der bei entsprechender Positionierung der Maus die Worterklärung (oder bei längeren Erklärungen deren Anfang) bietet, so dass sich oft auch ohne Klick ergibt, ob ein Wort für die eigene Fragestellung von Interesse ist.

Daneben gibt es aber auch zahlreiche Links, die aus dem Wörterbuch herausführen. Sie sind in der derzeitigen Darstellung durch eine leichte farbliche Unterlegung in beige gekennzeichnet und erscheinen nur in der Volldarstellung der Artikel.[5] Zum einen konnte eine Verknüpfung mit den Artikeln verschiedener historischer Wörterbücher eingebaut werden. Sie beruht auf einem automatisch durchgeführten Lemmaabgleich. Da für eine Überprüfung der Zuordnungen keine Personalkapazitäten zur Verfügung stehen, ist diese Verlinkung bewusst „großzügig“ gestaltet, insbesondere bei der Zuordnung von mittelhochdeutschen Lemmata. Wir

[5] Neben der Volldarstellung steht auch eine Anzeige im Übersichtsmodus zur Verfügung, die bei längeren Artikeln den Überblick über verschiedene Gliederungspunkte erleichtern soll und nur die Worterklärungen enthält. Volldarstellung, Übersicht und Faksimile des gedruckten DRW sind jeweils miteinander verlinkt, so dass man stets zwischen den verschiedenen Anzeigemodi wechseln kann.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

wissen, dass darin auch viele Verweise enthalten sind, die in die Irre führen, gehen aber davon aus, dass dem verständigen Benutzer damit besser gedient ist als mit einer stärkeren Filterung. Obwohl das Deutsche Rechtswörterbuch primär auf der Basis der Belegzettel erarbeitet wird, ist die Verlinkung zu Digitalisaten der Quellen, wie schon beschrieben, ein wichtiges Anliegen der Forschungsstelle. DRW Online umfasst deshalb neben dem Wörterbuch auch eine Reihe von Quellen im Volltext (im „Textarchiv“) bzw. als Faksimile. Zudem bietet das an der Forschungsstelle betriebene DFG-Projekt DRQEdit (http://drqedit.de) eine virtuelle Bibliothek der deutschsprachigen juristischen Literatur des 15. und 16. Jahrhunderts – nicht nur als Faksimile, sondern bei inzwischen über 100 Drucken auch im Volltext. Daneben kann mittlerweile aber vor allem auch zu Digitalisaten aus anderen Projekten verlinkt werden. Durch den Aufbau von Zuordnungstabellen konnten mittlerweile über 200.000 Belegstellen mit einem Digitalisat der betreffenden Textstelle verknüpft werden, das sind über 40% der DRW-Belege. Diese Verlinkung soll – im Rahmen der knappen personellen Ressourcen – auch in Zukunft weiter ausgebaut werden.

Bisher ging es um Links in DRW Online. Daneben ist es natürlich auch wichtig, dass die Artikel von außen adressiert werden können. Vorerst sind ihre URLs zwar aufgrund der derzeit noch verwendeten Framestruktur des Onlineangebots nicht unmittelbar ersichtlich, sie sind in ihren wesentlichen Bestandteilen aber leicht zu konstruieren und sollen auch längerfristig funktionieren.[6] So lässt sich das Deutsche Rechtswörterbuch z.B. als Hilfsmittel in die Präsentation von Texten einbeziehen. Die ins Textarchiv von DRW Online eingebundenen Quellen sind in dieser Weise hypertextualisiert, und zwar entsprechend den Stellen, die im Wörterbuch als Belege zitiert werden (vgl. Abb. 4).

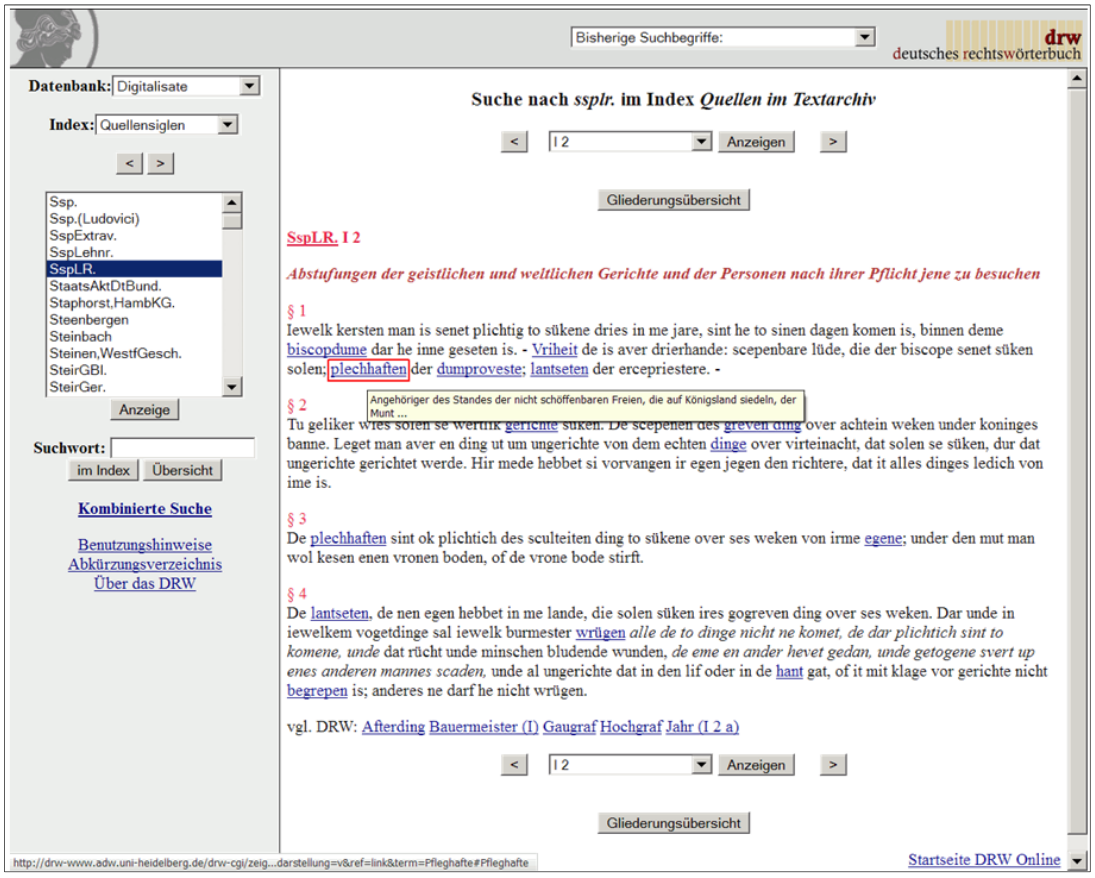

Abb. 4: Abschnitt aus dem Sachsenspiegel mit Links zu DRW-Artikeln

[6] Bei Interesse an einer Verlinkung kann gerne Rücksprache genommen werden.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Eine Verlinkung wird in aller Regel nur an bestimmten Stellen vorgenommen – sei es einzeln durch einen Bearbeiter, sei es nach bestimmten Kriterien automatisch. Gerade wenn es um die Erschließung von Sprache geht, kann aber potentiell jedes Wort für einen Link in Frage kommen, der zu entsprechenden Stellen führt. Deshalb gibt es in DRW Online zusätzlich zu den bisher vorgestellten Verknüpfungen auch noch etwas, was man als „implizite Links" bezeichnen kann: Sobald man auf ein Wort doppelklickt (und der Browser mitspielt), wird eine übergreifende Suche in den verschiedenen Komponenten von DRW Online durchgeführt. Auf diese Weise kann man also ggf. (nämlich wenn es sich um das Lemma oder um eine im Artikel belegte Schreibform handelt) schnell zum entsprechenden Wortartikel gelangen, man kann aber auch weitere Belege, Stellen im Textarchiv oder Worterklärungen finden, die diese Wortform enthalten. Unmittelbar einleuchten dürfte der Nutzen als Verständnishilfe für die Belegzitate und die Quellen im Textarchiv. Daneben soll diese Funktion aber auch eine Einladung sein, die verschiedenen Suchmöglichkeiten in DRW Online für Fragestellungen zu nutzen, die über die reine Worterklärung hinausgehen.

Literatur

Deutsch, Andreas (2010): Von „tausend Wundern" und einem „gewaltigen Zettelschatz". Aus der Geschichte des Deutschen Rechtswörterbuchs. In: Deutsch, Andreas (Hg.): Das Deutsche Rechtswörterbuch – Perspektiven. (= Akademiekonferenzen 8). Heidelberg: Universitätsverlag Winter, S. 21-45.

Dickel, Günther / Speer, Heino (1979): Deutsches Rechtswörterbuch. Konzeption und lexikographische Praxis während acht Jahrzehnten (1897-1977). In: Henne, Helmut (Hg.): Praxis der Lexikographie. Berichte aus der Werkstatt. Tübingen: Niemeyer, S. 20-37.

DRW (1914ff.): Deutsches Rechtswörterbuch. Wörterbuch der älteren deutschen Rechtssprache. Hg. von der Preußischen Akademie der Wissenschaften [Bände 1-3]/der Deutschen Akademie der Wissenschaften zu Berlin [Band 4]/der Heidelberger Akademie der Wissenschaften [Band 5ff.], bearb. von Richard Schröder u.a. Weimar: Hermann Böhlaus Nachfolger.

Lemberg, Ingrid (2001): Die Belegbearbeitung in der lexikographischen Datenbank des Deutschen Rechtswörterbuchs. In: Moser, Stephan et al. (Hg.): Maschinelle Verarbeitung altdeutscher Texte V. Beiträge zum Fünften Internationalen Symposion Würzburg 4.-6. März 1997. Tübingen: Niemeyer, S. 129-147.

Lemberg, Ingrid / Petzold, Sybille / Speer, Heino (1998): Der Weg des Deutschen Rechtswörterbuchs in das Internet. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion III. (= Lexicographica. Series Maior 84). Tübingen: Niemeyer, S. 262-284.

Lill, Eva-Maria (1998): Die EDV – das Ende aller Verzettelung? Der Einsatz der elektronischen Datenverarbeitung am Deutschen Rechtswörterbuch. In: Große, Rudolf (Hg.): Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern. Beiträge zu einer Arbeitstagung der deutschsprachigen Wörterbücher, Projekte an Akademien und Universitäten vom 7. bis 9. März 1996 anläßlich des 150jährigen Jubiläums der Sächsischen Akademie der Wissenschaften zu Leipzig. (= Abhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig, Philologisch-historische Klasse 75, 1). Leipzig: Hirzel, S. 237-247.

Speer, Heino (1994): DRW to FAUST. Ein Wörterbuch zwischen Tradition und Fortschritt. In: Lexicographica. Internationales Jahrbuch für Lexikographie 10, S. 171-213.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

On the descriptive power of the *ANW* semagram

Frans Heyvaert, Instituut voor Nederlandse Lexicologie, Leiden, The Netherlands

1. Introducing the *ANW*

The *ANW* (*Algemeen Nederlands Woordenboek* = Dictionary of Common Dutch) is an online lexicographic project conceived around the turn of the century and emerging from the Dutch *Instituut voor Nederlandse Lexicologie* (INL). It has introduced some noteworthy methodological and descriptive innovations in Dutch lexicography that might also be of interest for dictionary projects outside the frontiers of The Netherlands. Most of these novelties make part of a methodological framework designed to make dictionary content as consistent and uniform as possible. The uniformity pursued is meant to serve two aims at the same time: (1) to produce a dictionary that can not only be used to look up individual entries but that at the same time works as a useful and directly operational tool for the study of the internal semantic and grammatical organisation of the lexicon and (2) to offer to the user maximum retrieval opportunities in both directions (from word to meaning and from meaning to word) on the basis of, among other means, sameness of the terminology used in descriptions of related words. The major novelty in the *ANW*, playing a crucial role in the realisation of both goals, is the introduction of the so-called *semagram* (a term coined by the first editor-in-chief F. Moerdijk; see Moerdijk 2007, 2008).

2. What does a semagram look like?

The origin of the semagram is to be traced back to a project of lexical definition analysis that took place from 2002 to 2005 as part of the preparatory studies for the conception of the *ANW* format. In this analysis the definitions of all monomorphematic words of the major lexical categories (nouns, verbs, adjectives) in present day Dutch as they are available in the major monolingual dictionaries were involved[1]. The analysis aimed at two goals: listing and clustering the different syntactic headwords in the dictionary definitions and investigating into the types of modifications that go with those headwords.

The clustering of headwords has resulted in an organisation of the word stock of the major parts of speech into a fixed set of semantic categories intended to become the basic ordering principle in the description of the entire vocabulary. However, the fact that this set of categories is based on only a rather small part of the vocabulary reduces its status to that of a working hypothesis that will have to be improved continuously in the course of the editing. A definitive set of categories will only be defined at the end of the editing process. The genesis of this kind of categorisation – syntactic headwords of definitions! – might suggest a basic assumption that category names should be identical with hypernyms. To a large degree this is true for nouns but even there the hypernym stock in Dutch proves not to be sufficient to cover all categories that can empirically be discerned. This is e.g. illustrated by dictionary definitions having as their head *deel van* (“part of”), as is frequently the case with definitions of parts of furniture, buildings, tools etc. It does not make sense to put forward a category “part”, for its distribution would be too incoherent to make it useful for retrieval operations. So the criterion for something being a category cannot be that there is a single word for it, but that it

[1] The total amount of word meanings involved in the project was about 10,000.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

defines a coherent set of semantic properties inherited from the meaning of the category name by all the category members. Apart from that also common elements in grammatical behavior may play a role. This is for example the case with the distinction to be made between related categories of countable and non-countable nouns: the next higher category for *scallop* cannot be *meat*, the latter being a mass noun; it should be something like *piece of meat*.

The editing procedure of the *ANW* is organised per category: all entries of the same category are produced simultaneously. By this procedure the risk of deviation from uniformity is seriously reduced. Moreover the consistency of the category is checked once more when its editing is completed. Polysemy, of course, is a serious obstacle to this procedure, different senses normally belonging to different categories. This can be remedied in two ways. Either one can describe only the one sense that fits in the category under treatment. But that is quite unpractical for it would require an enormous accountancy of unfinished entries. Or one can describe provisionally senses belonging to categories that are still in preparation and revise them when that category is being taken in hands.

On each semantic category a specific internal organization is imposed (provisionally, just as for the categorization) in the form of a template consisting of a set of feature slots. These sets are the result of the analysis of types of modifications in existing definitions as mentioned above. In principle each category is provided with a specific set of feature slots. The result of this can be observed in the two examples of *ANW* semagrams below, originally for the Dutch words *hond* and *school*, but here, for ease's sake, adapted to be useable for English *dog* and *school*.

The semagram for *dog*:

A dog
[CATEGORY] is a mammal
[SIZE] can, according to the race, be from very small, like a chihuahua to quite large, like a Great Dane
[SOUND] barks; yelps; howls; growls
[LOOKS] has a coat that can have various colours from white over grey and all types of yellow and brown to black, either plain or spotted; can be long-haired or short-haired; can be wire-haired or silky; has sharp teeth, mostly with very pronounced canine teeth; has a tail, in some cases long, in some cases short, sometimes hairy, sometimes not, sometimes hanging down, sometimes erect
[ATTRIBUTES] sometimes wears a necklace; sometimes wears a muzzle to prevent it from biting
[FUNCTION] is kept for company; is used to guard properties; is, if it belongs to a race of sheep-dogs, used to herd sheep; is, if it belongs to a race of sporting-dogs, used to trace and hunt game; is sometimes used by blind people to guide them outside on the streets; is sometimes used by the police for detective work because of its sense of smell
[DESCENT] descends from the wolves
[AGE] is called a pup or puppy before its adult age
[SITUATION] is domesticated
[CHARACTER] is normally loyal and obedient to his human master; is generally very alert; can be agressive, threatening and unreliable towards strangers
[ABILITIES] has a highly developed sense of hearing; has a very strong sense of smell; can in most cases run very fast
[TREATMENT] is educated and trained by its owner, sometimes with the help of a dog training school; is when it is kept for company or as a show dog, very well groomed, often brushed, trimmed, combed and washed
[BEHAVIOUR] is carnivorous, eats meat; eats as a domestic animal often kitchen leftovers; often eats special dog food preserves; is very playful when it is young; has a typical body language, like wagging its tail when it is gay, growling when it is afraid of something and lying down to show submissiveness; slavers sometimes; sometimes is panting with its tongue out of its mouth; sometimes eats at bones and buries them; has inimical reactions to cats; sometimes attacks people; sometimes bites people; sometimes relieves nature on the pavement, to the annoyance of many people

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

[APPRECIATION] is quite generally considered man's best friend; sometimes frightens people because it can be aggressive and unreliable; is in some cultures an impure animal that has to live on the street

For *school*, the semagram would look like:

A school
[CATEGORY] is an institution
[ACTIVITY] provides education and instruction, which can be basic education, different kinds of secondary education or higher education
[PRODUCT] provides the pupils or students with certificates; turns pupils into people with a certificate
[PLACE] is typically housed in a big building or a complex consisting mainly of classrooms and of a playground for recreation
[PEOPLE CONCERNED] is meant for pupils or students
[AGENS] makes use of the services of teachers
[ORGANISATION] is either a public or a private organisation; is governed by a headmaster etc.
[TIME] performs its activity every day on working days; works according to a regular time schedule, usually from about nine in the morning till four in the afternoon; is active during the whole year with the exception of some school holidays of which the summer holiday is the most important; divides its working time in school years which run from the end of one summer holiday to the beginning of the next summer holiday
[MEANS] uses an established educational method, either classic or alternative
[GROUP] often brings together people who share the same religion or philosophy
[APPRECIATION] is considered by many people as an institution where one can only acquire artificial knowledge that is far away from "real life"

The semagram offers opportunities to increase the descriptive power of the dictionary entry considerably, not only at the service of the language user and the language student but also for the benefit of the professional linguist. These improvements do not only pertain to the internal bulk of information that becomes available in each separate entry, but also to the revelation of the relationships between words and between different senses of a word. Because of their systematicity semagrams need not be just isolated semantic descriptions of separate words. Simultaneous description of the whole stock of members of a category strongly promotes comparative and contrastive investigation and creates optimal facilities for that. Insight into the semantics of one word benefits from direct accessibility of knowledge of cognate words. In fact in some cases, like descriptive adjectives, the semantic structure of the category makes part of the meaning of the individual word and semantic features of one word only get their full meaning in relation to comparable features of cognate words (like size and colour adjectives).

The descriptive power of semagrams with regard to such paradigmatic relations will be discussed further on in this paper. First some enrichments will be presented which the semagram can bring in the internal structure of an entry.

3. Unification of conceptual and referential information

Prototype semantics in the last few decades has led indirectly to the insight that a conceptual semantic description does not always lead to a correct prediction of reference (e.g. the introductory chapter to Margolis and Laurence (1999) pays a lot of attention to that problem). This is an idea that seriously complicates the lexicographer's task, for an adequate meaning description requires that both aspects should be done justice. Both are semantic realities. Words are used to refer to things in the world but also – like e.g. in generic sentences – to talk about the "ideas" that they express. Adequate meaning rendering in a dictionary requires information about the one as well as the other, if possible in such a way that the dictionary users can learn from it whether their own use of any word corresponds to the standards or not. But

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

one immediately understands that joining together a sufficient conceptual description of a word with a complete account for its referential valencies would lead lexical entries to proportions unseen as yet. Due to limitations of available space in paper dictionaries and consequently the demand for conciseness, the lexicographer is often forced into making a choice between on the one hand focusing on (prototypical) concepts and thus neglecting borderline referential aspects and, on the other hand, showing the referential potential of a word and as a consequence, offering a rather diffuse account of the concept. In fact there should be no need to make that choice. Extensive conceptual analysis in lexicography, as introduced by Wierzbicka (1985), was originally conceived in order to solve some puzzling referential problems such as the *cup – mug* distinction discussed in Labov (1978). Moerdijk originally introduced it as a descriptive device to render the prototypical features of concepts, i.e. to extend the classical "minimal" definition in terms of putative necessary and sufficient condition with information of a more encyclopaedic nature that is also reflected in language use. But the semagram as a descriptive tool is strong enough to fulfill both functions together: to point at the most typical semantic features associated with the concept denoted by a word and at the same time specify the conditions on reference as they are shown in the corpus used and are present in the lexicographer's own language intuition. To distinguish between both types of information, one can make use of "hedges" (for the use of this term see Lakoff 1972) in the feature fillers, such as *typically*, *in some cases*, *generally*, *always* and so on, to indicate the status of the feature given. Or in the case of words with a diffuse reference, one can use enumerations or such expressions as *either … or …* . As such both generic and referential use of the word can be explained within the same framework and yet be clearly distinguished one from the other. The different types of feature descriptions can for instance be observed in the above *dog* semagram. If one compares the feature slots [SIZE] and [LOOKS] with [CHARACTER] one can remark that the fillers in the first two slots consist of enumerations of possible physical appearances of dogs and as such pertain in the first place to reference, while the fillers of [CHARACTER] contain such warnings as *normally*, *generally*, *can* and so on to indicate that these are properties of typical dogs, not of all dogs.

4. Distinguishing word senses

Descriptions in the form of semagrams also hold an implicit definition of what a single meaning or a word sense is. One can only say that a word in a set of using instances has one and the same sense if in all these instances the word belongs to the same category and shares the same set of semantic features. A word has two senses when it can be related to two different categories for different referents. So *wheat*, for one set of referents, denotes plants and as such belongs to the category *plant*. But the word can also refer to the grains produced by this plant. When used like that, it belongs to the category *seed*. The ambiguity of the word goes together with a clearly defined class difference. The same holds for example for *cocoa* in the senses "powder" and "drink". Also words belonging to only one category may show different senses. So in Dutch the word *marmot* can be used for two quite distinct animals: either one of the species *Marmota*, a kind of squirrel living in the Alps or a guinea-pig (species *Cavia*). In both uses the word belongs to the same category (*Mammal*, just like *dog* does). But within that category they have different sets of features to a degree that there are features that can be attributed to some or to all animals of the one kind but can never be attributed to any specimen of the other kind. To name only one: the shrieking noise that guinea-pigs can make.

Of course there are other well-known empirical tests for distinguishing word senses, in most cases based on referential identity and applicability of a variety of anaphoric expressions. A white dog and a black dog make two dogs. A marmot and a guinea-pig never make two mar-

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

mots, not even in Dutch. One would expect that both methods for detecting ambiguity would yield the same results. Generally spoken they do but one can also observe some frictions between them. So as an illustration the compound *car dealer* can refer to either a person or a company. According to the categorisation criterion it should thus be ambiguous for belonging to two different categories: *person* and *organisation*. But according to the referential criterion it is not, as is shown in *All car dealers will be visited by a tax official*, in which sentence the word refers to both persons and companies. Examples like this, however, do not falsify the categorisation principle. The set of categories used at present in the ANW is, as has been said, at this stage only a working hypothesis, which needs to be refined in the course of the editing. Now here is such a case. What the example proves is that, apart from the categories *person* and *organisation* one must assume a third one: something like *person or entity represented as such.*[2]

5. Compounds and combinatorics

Not only does the semagram offer a considerable enrichment of the descriptive apparatus for the semantics of the individual word, it is also involved in the description and explanation of its paradigmatic as well as syntagmatic relations. For one thing, it can offer some of the information about arguments and modifications that also the lexical functions Mel'čuk style do provide (See, for instance, Mel'čuk 1996). One may have noticed that in the semagram for *dog* above the semantic relations between this word and a series of terms which are idiomatically related to it, like *bark*, *pup*, and *wag* are made explicit. This type of information can in principle be incorporated in every semagram where it is relevant. Other phenomena in which the semagram has an explanatory role are compound formation, combinatorics and polysemy relations. What follows is a brief survey of how this works.

First, descriptions in terms of semagrams have some predicting qualities with regard to the formation and interpretation of compounds. When a word is used as the right part (the head) of a compound of the specificans-specificatum type, the word used in the left part of the compound very often provides information about one of the meaning dimensions included in the feature slots of the semagram. As such the semagram also functions as an identification tool for the relation between the components of compound words, especially nouns. This role of the semagram has been elaborated extensively in Moerdijk (1987, 1988) among others for the compound with *mes* ("knife") and in Heyvaert (2009) for the compounds formed with Dutch *school*.

Also the degree of idiomaticity of syntactic combinations in which a word occurs is influenced by the semagram content. As an illustration this can be observed in the combinations of a noun with different adjectives, where one combination "feels" more idiomatic than the other. (For clarity's sake: *idiomatic* here is not to be taken in the sense of non-compositional but as more or less fixed by frequent use or "commonness".). If we compare the combinations *a*(*n*) [*aggressive*, *black*, *dangerous*, *lazy*, *wild*] *dog* with *a*(*n*) [*Australian*, *old*, *short-sighted, sleepy*] *dog*, we notice that we judge the former set of adjectives more typical of dogs and the latter rather accidental. The relation of the former set of adjectives to the noun feels less "free" than in the latter case, their meaning being related to some part of the *dog* semagram, either as a specification or as a denial. The latter set of adjectives has no meaning relation

2 As L. Lemnitzer (personal communication) remarks, alternatives could be to interpret this example as a case of underspecification (so no amendment to the definition needed) or as a case of type coercion VISIT – (PERSON ← ORGANISATION)

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

with the noun semagram and as a consequence forms a completely free combination with the noun and has no idiomatic feel to it.

6. Defining sense relations

Finally the semagram also offers an excellent tool to define and explain sense relations within a polysemous complex in a much more detailed and explicit way than is common practice in most dictionaries. Usually the whole explanation consists of a label like (*meton*.) or (*fig*.) or something like that, leaving the user uninformed about what the source sense of the metaphor or the metonymy is as well as about the way in which the derived sense is metonymic or metaphoric. By means of semagrams the latter two information categories can be expressed explicitly in a natural and understandable way. To show how this works, let us first take the simplest forms of polysemy: generalisation and specialisation. Generalisation is an extension of referential scope, paired to some loss of semantic structure. Specialisation is the opposite. Use of semagrams allows us to show exactly what this loss our gain consists of. To demonstrate this with a simple example: the noun *drink*, in its general sense, means "something to drink"; next to that, there is a specialised sense "beverage containing alcohol". The semagram of the latter sense contains, apart from all the features present in the original sense, an extra feature in the slot [INGREDIENT]. So the specialisation relation can be defined, in terms of the formalism of the *ANW* editing form: $drink_2$ has a specialization relation to $drink_1$ made explicit by addition of a feature in the slot [INGREDIENT]. In order for this information to be presented to the user in a uniform and easily accessible way, an extra information category is attached to the derived semagram, defining its semantic relation to the source semagram. This information category takes the form

Meaning relation General: [values: metaphor, metonymy, specialization, generalization, other, unclear][3] Specific: […….][4] Senses involved: [source sense number: goal sense number]

This formalism not only describes the relation between both senses more explicitly than usual, but, by its uniform application to all similar cases, it also allows semanticists to look up and gather similar instances of polysemy for purposes of study.

Metonymy is a more complicated type of semantic transition, appearing in such a wealth of different types (for a provisional but not at all exhaustive list see Apresjan 1992) that it seems an impossible task to impose some descriptive systematicity on it. Yet also here the semagram can offer substantial help to discover patterns. As a demonstration let us take *school* and its semagram as shown above. Apart from the "literal" use as in *My children go to a good school* we have other – metonymical – kinds of uses, such as in *The school is being painted*, *The school has decided to postpone the exams*, *School is out* and *The whole school was on the ice*. *School* in these sentences can be paraphrased roughly as – in the same order as the example sentences – *school building*, *school government*, *school time* and *school population*. Now if we look back to the semagram of the source meaning for these metonyms, we notice that

[3] The options between square brackets have the form of a choice menu window in the editing form.

[4] This information category has not been realised yet. It is to be incorporated in a later phase of the editing.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

these paraphrases are each related to the content of a specific feature slot: [PLACE], [ORGANISATION], [TIME] and [PEOPLE CONCERNED]. So metonymy – at least in the case of these types – can be defined as a category shift in which the original category has been replaced by the content of one of the feature slots in the semagram (for more details see Heyvaert 2009). So the use of semagrams also allows cases of metonymy to be defined in terms of a general formula, something like "X_2 bears a relation of metonymy to X_1 on the basis of feature slot A". For linguists such a uniform characterisation once again offers opportunities to extract large pre-ordered collections of study materials from a dictionary without having to do great efforts for that. For more elaborate treatments of metonymy along the same line of thought, see among others Moerdijk (1989, 1990a, 1990b).

Finally also for metaphorical transitions the use of semagrams creates opportunities for making generalisations and discovering patterns. One complication with lexicographical description of metaphor however is that it has been subject to dispute for ages in the hands of semanticists, philosophers and psychologists (see Johnson 1980; Ortony 1980). The task of the lexicographer in these matters is not to intervene in the discussion and take outspoken theoretical stances himself but rather to gather and order the phenomena and to present them in a well-organised but theory-neutral way. So we take as our starting point a working definition of metaphor that is general enough to please the whole audience and to grant everybody the opportunity of completing it according to his own insights. A lexical metaphor is a derived word sense in which the word is used for something belonging to a different semantic category than the thing named in the source sense and which is motivated by an amount of semantic content in the source sense that is applied in some way or another to the derived sense. The word *applied* is used deliberately here instead of the commonly used *transferred*, because the latter term already suggests some theoretical option.

The ANW account for metaphor is presented in just the same form as the ones for generalisation, specialisation and metonymy. It contains a specification of a category for the source meaning, a category for the derived meaning, and one or more semantic features which form the motivation of the sense derivation. Let us take *mouse* "animal" and *mouse* "computer utensil" as an example to demonstrate this. The semagram of the former surely must contain the features

> [LOOKS] has a grey coat; is small; has a pointed little head with two striking dark eyes looking like pearls; [MOVEMENT] moves quickly in all directions without making a sound.

The computer mouse is usually grey, is small, has two buttons in front which may remind one of a mouse's eyes or ears, and moves in all directions without making a sound.

These correspondences can be stated formally and explicitly in a formula of the type mentioned above, linking the semagrams of both senses:

Meaning relation General: [metaphor] Specific: [from category animal (or mammal) to category ustensil with features [LOOKS] and [MOVEMENT] involved] Senses involved: [1 : 2]

This descriptive device makes it possible to cross the borders of the individual word in the dictionary and to create easily retrievable typological categories of metaphors. This can be

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

demonstrated with some other metaphors using animal names to name persons or things. We take *aap* "monkey", *beer* "bear", *ezel* "donkey", *koe* "cow" , *melkkoe* "milk cow", *geit* "goat", *bok* "billy-goat", *schaap* "sheep", *paard* "horse", *hond* "dog" and *egel* "hedgehog":

based on the feature [BEHAVIOUR]:
aap "person who does nothing of his own but only imitates other people" – [FEATURE] imitates the behaviour of humans; "young boy who has to be tamed and civilised yet" [FEATURE] is sometimes very vivid;
bok "man with an unreasonably big sexual appetite [FEATURE] has a very developed sexual instinct[5]
egel "somebody who secludes himself from the others as a reaction to criticism instead of fighting back" [FEATURE] rolls itself into a ball of spines when it feels menaced

based on the feature [LOOKS]:
beer "tall, strong and somewhat ponderous man" [FEATURE] is tall, strong and somewhat ponderous
bok "vaulting-buck" [FEATURE] has four legs and a short body

based on the feature [SOUND]:
geit "woman who is taken for stupid and contempted" [FEATURE] makes a high, irritating and frequently sound; bleats

based on the feature [CHARACTER]:
beer, "big and heavy man whom one likes to cuddle" [FEATURE] is in childrens' stories presented as a kind-hearted but greedy and somewhat silly character
schaap "person in unfortunate circumstances, whom one must pity" [FEATURE] counts in Christian mythology as the symbol of innocence, of the innocent and submissive victim that raises pity

based on the feature [FUNCTION]:
melkkoe "person who is only used to raise maximum profit and who for the rest is not taken seriously or respected" [FEATURE] is mainly kept for milk production

based on the feature [APPRECIATION]:
hond "somebody not worthy of being treated as a human being" [FEATURE] is in some cultures considered impure and forced to live on the streets
Just as in *mouse* some metaphorical uses may be based on more than one feature:

based on the features [BEHAVIOUR] and [APPRECIATION]:
aap "man who makes a very primitive impression, either by his (hairy?) looks or by his uncivilised behaviour" [FEATURE 1] behaves more or less like man, but in a primitive, uncivilised way – [FEATURE 2] is judged comic or ridiculous because of its behaviour

based on the features [CHARACTER] and [APPRECIATION]
ezel "stupid and / or headstrong person" [FEATURE 1, FEATURE 2[6]] has the undeserved reputation of being, stupid and stubborn

based on the features [SOUND] and [BEHAVIOUR]
beer "big and brutal man" [FEATURE 1] roars – [FEATURE 2] sometimes attacks people

based on the features [LOOKS] and [BEHAVIOUR]
koe "woman who is fat, ponderous, bad-mannered or all these together" [FEATURE 1] has an unwieldy, ponderous appearance – [FEATURE 2] runs through and over things without paying attention

It must be admitted that there is something odd with the idea of transfer of feature content. It would be a ridiculous assumption that e.g. for *egel* the feature "rolls itself into a ball of spines when it feels menaced" would be literally incorporated in the metaphorical sense of the word. If we really pretend to give an explicit account of lexical metaphors, we should also make ex-

[5] Since *bok* is a very depreciatory term, having a connotation of filthiness, one may guess that also other less nice elements of the source meaning play a role here, like [SMELL] has a disgusting smell; stinks.

[6] This is a feature content that clearly fits under two different feature slots. Here it is the slot name which makes the difference. Things like that happen.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

plicit how this feature is translated into "secludes himself from the others as a reaction to criticism instead of fighting back" or on what grounds it can refer to a person who secludes himself that way. What we need here is a good definition of what should be understood by *transfer*. But that again is something that should be found out by theoretical semanticists, philosophers and psychologists. The dictionary can better have a little faith in the imagination of its users. It must at least leave something to that.

Another phenomenon it is difficult to account for in this approach is the degree of conventionalisation of metaphors. A metaphor that is fully conventionalized is also completely dead. Most of the time one does not even notice that it is a metaphor anymore, since any association with the source meaning has become unnecessary to understand it. This is, for instance, the case in *foot of a mountain, of a page*. Metaphors that have not yet reached that stage show different degrees of activity for their use often keeps showing interaction with the use of the source sense. The *egel* example above can be used to illustrate this. If someone is called *egel*, one can also say about him that he is using his spines again, that you can seriously hurt yourself in contacts with him and so on. In fact this is another kind of typology than the one proposed above and the phenomenon only pertains to the word in its derived sense. Maybe to cope with this one should incorporate some kind of slot for this in the semagram of the derived meaning. But that is something that has to be found out in some later stage of the project.

Something that has also not yet been taken in account in the intended project are the so-called conceptual metaphors, i.e. coherent metaphorical organisation underlying the manner of speaking about a whole cognitive field (see for this Lakoff/Johnson 1980). This does not mean that they will remain excluded from treatment in the long run. One may guess that the categorical treatment of words, as is practiced in the *ANW*, could provide a solid basis for making statements about this type of metaphor within the framework of a dictionary database.

7. Summary

A semantic description of a word that really does justice to it should not limit itself to giving just enough information to identify the thing denoted. Words are burdened with all kinds of associations that we can find back in the typical contexts in which it is used, in a variety of "secondary" properties that are ascribed to their denotatum, which are responsible for the derivation of new senses from the original meaning and so on. Words are linked to some "idea" that they evoke when considered outside of a specific context and they have a referential domain that often cannot be explained by that idea alone or may even be contradictory to some aspects of that idea. These are all phenomena that semagrams as used in the *ANW* can cope with. A semagram is not just an elaborate semi-encyclopaedic description of the thing named by a word that exceeds by far the information necessary to identify that thing. Quite some aspects of the linguistic behaviour of the word can be explained at least in part from the semagram information: the formation and interpretation of most compounds, the occurrence of some more or less lexicalised syntactic combinations, regular polysemy are all phenomena which seem to find their motivation in the conceptual associations that constitute the word meaning in a broader sense and that are made explicit in semagrams.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

8. References

Apresjan, Yuri D. (1992): Lexical semantics. User's guide to contemporary Russian vocabulary. Ann Arbor: Karoma Publishers.

Heyvaert, Frans (2009): Terug naar *school*! In: Beijk, Egbert et al. (eds.): Fons Verborum. Feestbundel voor prof. dr. A. M. F. J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie. Amsterdam: Gopher, p. 143-159.

Johnson, Mark (1980): A philosophical perspective on the problems of metaphor. In: Honeck, Richard P. / Hoffman, Robert R. (eds.): Cognition and figurative language. Hillsdale: Lawrence Erlbaum Associates, p. 47-68.

Labov, William (1978): Denotational structure. In: Farkas, Donka et al. (eds.): Papers from the parasession on the lexicon. Chicago: Chicago Linguistics Society, p. 220-260.

Lakoff, George (1972): Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. In: Peranteau, Paul M. et al. (eds.): Papers from the Eighth Regional Meeting of the Chicago Linguistics Society. Chicago: Chicago Linguistics Society, p. 458-508.

Lakoff, George / Johnson, Mark (1980): Metaphors we live by. Chicago: University of Chicago Press.

Margolis, Eric / Laurence, Stephen (1999): Concepts: core readings. Cambridge: MIT Press.

Mel'čuk, Igor A. (1996): Lexical functions: a tool for the description of lexical relations in the lexicon. In: Wanner, Leo (ed.): Lexical functions in lexicography and natural language processing. Amsterdam: John Benjamins, p. 37-102.

Moerdijk, A. M. F. J. [Fons] (1987): Lexicale semantiek en compositavorming. In: Forum der Letteren 28, p. 194-213.

Moerdijk, A. M. F. J. [Fons] (1988): Lexicaal-semantische vormingspatronen voor samenstellingen. In: Jaarboek van de Stichting Instituut voor Nederlandse Lexicologie. Overzicht van het jaar 1987. Leiden: INL, p. 49-65.

Moerdijk, A. M. F. J. [Fons] (1989): Benaderingen van metonymie. In: Forum der Letteren 30, p. 115-134.

Moerdijk, A. M. F. J. [Fons] (1990a): Metonymie uit een ander vaatje. In: Traditie en Progressie. Handelingen van het 40ste Nederlands Filologencongres. 's-Gravenhage: SDU Uitgeverij, p. 111-122.

Moerdijk, A. M. F. J. [Fons] (1990b): Metaal aast op vrouwen. Het verschijnsel metonymie. In: Onze Taal 59, p. 68-70.

Moerdijk, A. M. F. J. [Fons] (2007): Definities, frames en semagrammen. In: Moerdijk, Fons / van Santen, Ariane / Tempelaars, Rob (eds.): Leven met woorden. Opstellen aangeboden aan Piet van Sterkenburg bij zijn afscheid als directeur van het Instituut voor Nederlandse Lexicologie en als hoogleraar Lexicologie aan de Universiteit Leiden. Leiden: Instituut voor Nederlandse Lexicologie / Koninklijke Brill, p. 63-75.

Moerdijk, A. M. F. J. [Fons] (2008): Frames and semagrams. Meaning description in the General Dutch Dictionary. In: Bernal, Elisenda / De Cesaris, Janet (eds.): Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008), Barcelona: Documenta Universitaria, p. 561-571.

Ortony, Andrew (1980): Some psycholinguistic aspects of metaphor. In: Honeck, Richard P. / Hoffman, Robert R. (eds.): Cognition and figurative language. Hillsdale: Lawrence Erlbaum Associates, p. 69-83.

Wierzbicka, Anna (1985): Lexicography and conceptual analysis. Ann Arboer: Karoma Publishers.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions

Jörg Didakowski, Berlin-Brandenburgische Akademie der Wissenschaften
Alexander Geyken, Berlin-Brandenburgische Akademie der Wissenschaften

1. Introduction

In this article, *Wortprofil 2012* is presented, the current version of a lexical profiling tool for German based on grammatical co-occurrences. *Wortprofil 2012* can be used twofold: it assists lexicographers in their work to compile collocations and it provides useful corpus-based syntagmatic information for users interested in improving their language production skills. It is implemented as an additional functionality of the lexical information system of the Digital Dictionary of the German Language ("Digitales Wörterbuch der deutschen Sprache", DWDS) which is accessible to all users via the internet (www.dwds.de). Wortprofil 2012 is a further development of the word profile system presented in Geyken et al. (2009).

Wortprofil 2012 provides separated co-occurrence lists for twelve different grammatical relations and links them to their corpus contexts where the node word and its collocate co-occur. The co-occurrence lists and their ordering are based on statistical computations over a fully-automatic annotated German corpus containing about 1.8 billion tokens.

Wortprofil 2012 can help to answer questions like "Which attributive adjectives are typically used for the noun *Vorschlag* 'proposal'?" or "Which active subject does a verb like *ausstoßen* 'emanate' usually take?" and it can be a good starting point for looking at a specific word.

The remainder of this article is structured as follows. In section 2 we shortly describe the specific challenges that the German language provides for the extraction of syntactically relevant co-occurrences. In section 3 we comment on related work in this field. In section 4 we describe the corpus basis used for the extraction of the co-occurrence. This process together with the statistical computations are described in section 5 and 6. Finally section 7 provides some concluding remarks as well as an outlook on future work.

2. Challenges of the German Language

There are several grammatical characteristics of German which make the extraction of grammatical co-occurrences particularly challenging.

German has a variable placement of the finite verb (verb-first, verb-second and verb-final) which depends on the clause type. Additionally, German has a relatively rich morphology with case marking which allows a relatively flexible phrase ordering. These flexible orderings can cause discontinuous verb chains and they can cause separable prefix verbs where the prefix is far away from its target. Furthermore, in newspaper text it can be shown that the ambiguity rate with regard to the morphological case information is very high (Evert 2004).

These features of German make it very difficult to extract grammatical co-occurrences by sequence based formalisms if one is interested in syntactic relations on sentence level.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

3. Related Work

Church / Hanks (1991) show that lexical statistics are very useful for summarizing concordance data of a corpus by representing a sorted list of the statistically most salient collocates. In order to extract collocates they use natural language processing tools like a part-of-speech tagger and a partial phrasal parser. Their statistical calculation is based on mutual information.

Kilgarriff / Tugwell (2002) build on this idea in their word sketch approach. Word sketches are "summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2004) and are provided by a so called Sketch Engine, a corpus query system. For languages with fixed word order the Sketch Engine uses patterns over part-of-speech sequences to detect grammatical relations in form of a sketch grammar.

Ivanova et al. (2008) develop a sketch grammar for German. Their main problem is to achieve high precision and high recall at the same time. Their results show that richer linguistic analysis is necessary to obtain a better overall performance.

Horák / Rychlý (2009) use a robust syntactic parser in order to extract grammatical relations for the Czech Language, which has a rich morphology and a relatively free word order. The result of the deeper analysis is used as input to the Sketch Engine.

4. Corpus Selection and Corpus Size

The base of Wortprofil 2012 is the corpus collection shown in table 1. It consists of renowned German daily and weekly newspapers for which legal arrangements are obtained for the use within the DWDS project together with the balanced reference corpus "DWDS Kernkorpus", the core of our corpus collection. The corpora range from 1900 up to now.

corpus	tokens	sentences	documents
Süddeutsche Zeitung	453,945,194	29,125,790	1,099,920
DIE ZEIT	417,422,714	23,631,230	499,520
Berliner Zeitung	242,046,373	15,951,701	869,023
DIE WELT	238,403,711	15,787,624	600,007
Der Tagesspiegel	184,202,717	10,392,257	394,465
DWDS-Kernkorpus	125,990,080	7,046,937	79,312
Bild	121,520,037	12,629,828	548,181
total	**1,783,530,826**	**114,565,367**	**4,090,428**

Table 1: The base of *Wortprofil 2012*

In table 2 a more compact listing of the collection is presented. It shows that the major part of the collection consists of daily newspapers. The total size of the corpus bases amounts to 1.78 billion tokens in 4.09 million documents.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

corpus	tokens	sentences	documents
daily newspaper	1,240,118,032	83,887,200	3,511,596
weekly newspaper	417,422,714	23,631,230	499,520
balanced corpus	125,990,080	7,046,937	79,312
total	**1,783,530,826**	**114,565,367**	**4,090,428**

Table 2: The base of *Wortprofil 2012* – compact list

The newspaper archives contain a substantial number of duplicates, thus leaving the user with a distorted view of the *Wortprofil 2012* statistics. Therefore, we decided to remove all duplicates from the whole corpus collection. After the removal our collection comprises 90,806,646 sentences and 1,594,223,632 tokens.

Linguistic preprocessing for the relation extraction comprises the following steps: the corpus texts are automatically tokenized, split up into sentences and annotated with part-of-speech tags by the moot tagger (Jurish 2003). This is the standard preprocessing of the DWDS corpora.

5. Relation Extraction

In the approach presented in this article syntactic parsing is used as backbone for the extraction of grammatical co-occurrences. That is, the extraction is divided into two tasks: first the sentences of a corpus are syntactically analysed via a robust dependency parser; second the syntactic relations of interest are extracted from the parsing results. An example of this procedure is given below.

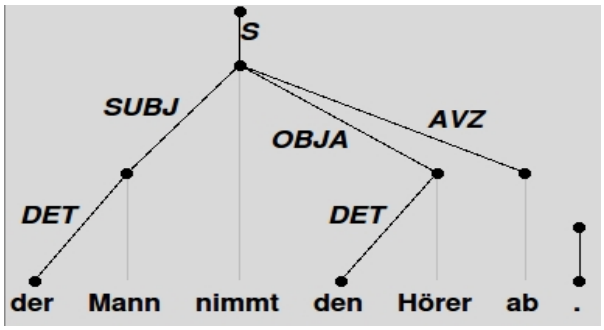

Figure 1: Example of a parsing result

In figure 1 the parsing result of the sentence “Der Mann nimmt den Hörer ab” (‘The man picks up the handset’) is shown. Via the dependency links and the dependency labels (SUBJ, OBJA, AVZ) three binary grammatical relations relevant for collocation extraction can be extracted:

- verb prefix: <*nehmen* ‘to pick’, *ab* ‘up’>
- active subject: <*abnehmen* ‘to pick up’, *Mann* ‘man’>
- accusative object: <*abnehmen* ‘to pick up’, *Hörer* ‘handset’>

For this task, it is not important which kind of syntactic parser is chosen. It could be a dependency parser or a phrase structure parser or any other parser providing sufficient information for the relation extraction task.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

5.1 The Syntax Parser

For syntactic annotation the rule based dependency parser SynCoP (Syntactic Constraint Parser, Didakowski 2008a; 2008b) is used. The parser is implemented with the help of finite state techniques and is based on the high-coverage morphology TAGH (Geyken / Hanneforth 2006).

A grammar for the SynCoP parser is developed which is designed for the specific relation extraction task. Therefore, issues like the attachment of sub-clauses or specific rare syntactic phenomena are not dealt with in this grammar. Also, it does not cover long distance dependencies, it provides only weak lexicalisation information and it does not use subcategorisation information. The parser for this grammar is similar to what Grefenstette (1998) called an approximation to full parsing.

In the following a few features of the SynCoP parser are mentioned which are essential for the quality and quantity of extracted relations and for the parsing time.

The parser only assigns dependency structures which are allowed by its grammar. There is a distinction between grammatical and ungrammatical in contrast to grammarless data-driven parsers (cf. McDonald / Nivre 2011). We attempt to avoid evident annotation mistakes which emerge from phenomena such as agreement errors or uniqueness violations.

It is unrealistic to require a parser grammar to fully cover the syntactic structure of all sentences of a corpus because corpora generally contain fragments and ungrammatical sentences and because a grammar does not cover all grammatical phenomena. Therefore the parser allows partial analyses in order to annotate as much as possible. Thus the relevant phenomena for our extraction tasks can be annotated separately, including subordinate clauses, noun groups, and prepositional groups.

It is possible to assign weights to the grammar rules. In this way lexical and structural preferences can be defined. Note that only one dependency structure for a sentence is used for the relation extraction step and that ambiguities can only be resolved by the parser with the help of heuristics. With the weighted rules it can for example be expressed that a subject of a sentence is expected to precede its corresponding finite verb. It is also possible to model more general preferences like a longest match strategy implementing the late closure parsing principle (see Didakowski 2008b).

The parser does not rely completely on results of the part-of-speech tagger but uses them as preferences only. This approach is compatible with the observation that part-of-speech tagging is a considerable source of errors (Kilgarriff et al. 2004).

The corpus collection used in Wortprofil 2012 comprises a large amount of data and has to be annotated in adequate time. Therefore, the parser implements some mechanisms that speed up parsing. Left and right embeddings of sentences are replaced by iteration and centre embeddings of sentences are restricted to a depth of one. This approach follows the observation that there exists an absolute limit on centre-embeddings in written and spoken language (Karlsson 2010).[1] Furthermore, pruning is performed if the search space is too large to parse a sentence completely.

[1] Karlsson (2010) studies different types of recursion and iteration in written and spoken language empirically and examines empirical determinable constraints on the number of recursive and iterative cycles.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

5.2 The Covered Syntactic Relations

The syntactic annotation of our corpus collection is followed by the extraction of binary and ternary syntactic relations from the parsing analyses. Ten different syntactic binary relations are extracted. These relations are listed in table 3. In the second column of table 3 tuples of parts-of-speech are shown which are included by a syntactic relation. The coordination is the sole symmetric relation and the verb prefix relation emerges from morphological analysis and syntax analysis.

syntactic relation	**part-of-speech tuples**
accusative object	{<verb,noun>}
active subject	{<verb,noun>}
adjective attribute	{<noun,adjective>}
coordination	{<verb,verb>,<noun,noun>,<adjective,adjective>}
dative object	{<verb,noun>}
genitive attribute	{<noun,noun>}
modifying adverbial	{<verb,adverb>,<adjective,adverb>}
passive subject	{<verb,noun>}
predicative complement	{<noun,noun>,<noun,adjective>}
verb prefix	{<verb,prefix>}

Table 3: Extracted binary syntactic relations

In addition two ternary relations are considered. They are listed in table 4.

syntactic relation	**part-of-speech tuples**
comparative conjunction	{<noun,conjunction,noun>,<verb,conjunction,noun>}
prepositional group	{<noun,preposition,noun>,<verb,preposition,noun>}

Table 4: Extracted ternary syntactic relations

To parse one corpus of our corpus collection and to extract the syntactic relations took five days in average on five processors working in parallel. The overall extraction took about one month.

Frequency information about the extracted relations is given in table 5 in the second column. In total 381 million relations are extracted from the annotated corpora.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

5.3 Filtering of Extracted Syntactic Relations

In some cases the parser produces systematic errors. In order to increase the quality of the extracted relations a filter is applied taking the context into account in which an extracted relation is found. In this approach we distinguish between safe and unsafe relations depending on the context. That is, safe syntactic relations must be contained in a safe context. On the basis of this assumption a threshold calculation is applied. All relations are removed if they do not occur at least twice in a safe context.

In the following an example for an unsafe context is given. The parser has a problem with the differentiation of active perfect and static passive. This is shown by the sentences (1) and (2):

(1) *Der Mann ist gerudert*. (active perfect)
'The man had paddled.'

(2) *Der Baum ist gefällt*. (static passive)
'The tree is felled.'

Because no subcategorisation information is used in the grammar, the parser cannot distinguish between the two different analyses because the parser has to know whether the verb is transitive or intransitive. Therefore, both sentences would get the same annotation and this would cause the problem of quality in the extraction of the syntactic relations "active subject" and "passive subject".

There are also some more general criteria the context of safe relations has to meet:

- Nouns within a requested syntactic relation must have a determiner.
- All words involved in a requested syntactic relation have to be within an analysable subordinate clause or main clause.
- A sentence must start with an uppercase letter and must end with a final punctuation mark.
- A sentence must not contain any unknown word.

The frequency information about the extracted relations after the application of the filter is given in table 5 in the third column. In total 257 million relations remain.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

syntactic relation	frequency	frequency (applied filter)
prepositional group	93,640,099	45,165,932
adjective attribute	68,658,904	58,297,991
modifying adverbial	63,341,241	45,392,107
active subject	51,968,759	37,824,884
accusative object	29,458,909	19,251,695
Coordination	21,886,952	21,685,018
genitive attribute	22,051,327	10,975,398
verb prefix	8,488,938	8,142,960
predicative complement	7,531,807	3,895,181
dative object	5,263,222	2,685,383
passive subject	5,017,454	2,784,627
comparative conjunction	3,806,860	1,300,991
total	**381,114,472**	**257,402,167**

Table 5: Frequency information about the extracted relations

6. Statistical Computations

After the relation extraction a statistical computation is applied in order to determine the attraction between words of a specific syntactic relation.

Two statistics are used in *Wortprofil 2012*: logDice which is based on the Dice coefficient (see Rychlý 2008) and MI-log-Freq which is based on mutual information (see Kilgarriff / Tugwell 2002). These statistics can be used as a quantitative measure of the attraction between words where high scores indicate high correlation and where a distinction between negative (<0) and positive (>0) association is made (see Evert 2008).

In order to generate a sorted candidate set of collocations a ranking approach and a threshold approach are combined. First the candidate set is determined by using a threshold value of zero for logDice and MI-log-Freq and the threshold 5 for absolute frequency of word-occurrence is set. Then the candidate set is sorted by their association score, thus providing an ordering by collocational strength.

As a result of the statistical computations the database contains 11,980,910 different collocation candidates of specific syntactic types where it is possible to query 104,704 different lemma / part-of-speech pairs.

An example for a query in Wortprofil 2012 is shown in figure 2. It shows the query results for the noun *Vorschlag* (‘proposal’) with the syntactic relation “adjective attribute”. The result set is presented as a word cloud where the font size of the adjectives occurring with the query word correlates with the score of the statistical computations.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

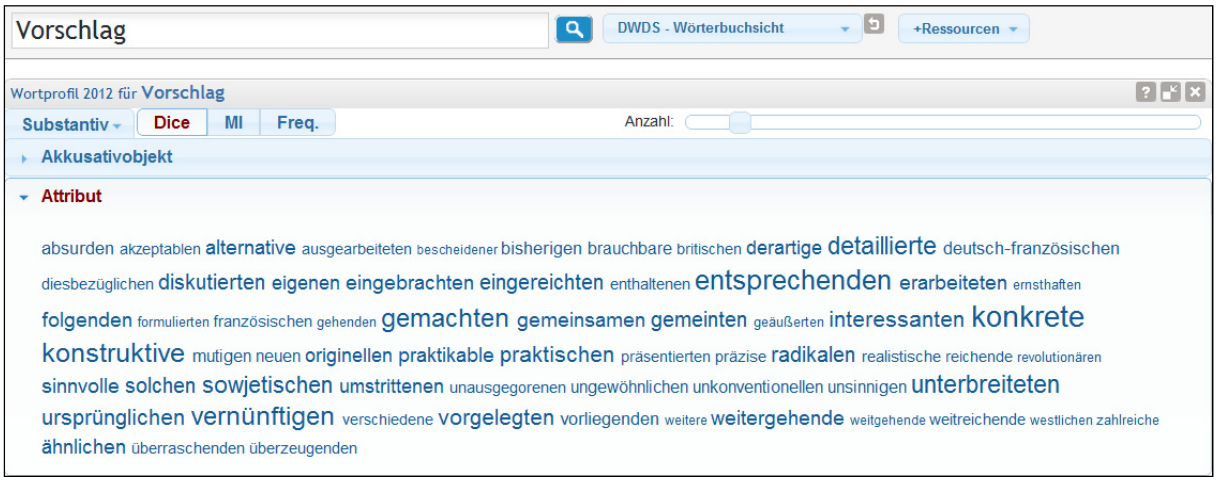

Figure 2: Example for the query result for the noun *Vorschlag* in *Wortprofil 2012* (DWDS, http://www.dwds.de/?qu=Vorschlag&view=1, Stand: 10.05.2013).

7. Conclusion

In this paper we have presented *Wortprofil 2012*, a lexical profiling tool for the extraction of statistically salient and syntactically relevant co-occurrences. We have presented the corpus base as well as the extraction method that relies on syntactic and statistical analysis. We have also shown that additional heuristics to reduce syntactic complexity as well as computational optimizations such as the introduction of thresholds are necessary in order to obtain satisfying results in a reasonable time.

Further work will focus on the evaluation of the results and on the evaluation of the heuristics and thresholds. In addition we will implement a more intuitive user interface in order to make the *Wortprofil 2012* useful for a broader public. Finally, we will apply the *Wortprofil* method to a diachronic corpus of German.

8. References

Church, Ken / Hanks, Patrick (1991): Word Association Norms, Mutual Information and Lexicography. In: Computational Linguistics 16, 1, p. 22-29.

Didakowski, Jörg (2008a): Local syntactic tagging of large corpora using weighted finite state transducers. In: Storrer, Angelika et al. (eds.): Text resources and lexical knowledge: selected papers from the 9th Conference on Natural Language Processing. KONVENS 2008. Berlin: Mouton de Gruyter, p. 65-78.

Didakowski, Jörg (2008b): SynCoP – combining syntactic tagging with chunking using weighted finite state transducers. In: Hanneforth, Thomas / Würzner, Kay-Michael (eds.): Finite-state methods and natural language processing. 6th International Workshop. FSMNLP 2007. Potsdam: Universitätsverlag, p. 107-118.

Evert, Stefan (2004): The statistical analysis of morphosyntactic distributions. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. LREC 2004. Lisbon, Portugal, p. 1539-1542.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke / Kytö, Merja (eds.): Corpus linguistics. An international handbook of the science of language and society. Berlin / New York: Mouton de Gruyter.

Geyken, Alexander et al. (2009): Generation of word profiles for large German corpora. In: Kawaguchi, Yuji et al. (eds.): Corpus analysis and variation in linguistics. (= Tokyo University of Foreign Studies, Studies in Linguistics 1). Amsterdam: John Benjamins, p. 141-157.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Geyken, Alexander / Hanneforth, Thomas (2006): TAGH: a complete morphology for German based on weighted finite state automata. In: Yli-Jyrä, Anssi et al. (eds.): Finite-state methods and natural language processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, Revised Papers. (= Lecture Notes in Artificial Intelligence 4002). Berlin/Heidelberg: Springer, p. 55-66.

Grefenstette, Gregory (1998): The future of linguistics and lexicographers: will there be lexicographers in the year 3000? In: Fontenelle, Thierry et al. (eds.): EURALEX 1998 Proceedings. Liège: University of Liège, p. 25-41. Internet: http://www.euralex.org/elx_proceedings/Euralex1998_1/ (last visited: 10.05.2013).

Horák, Aleš / Rychlý, Pavel (2009): Discovering grammatical relations in Czech sentences. In: Proceedings of the RASLAN Workshop 2009. Vyd. první. Brno: Masaryk University, p. 81-90. Internet: http://nlp.fi.muni.cz/raslan/2009/papers/16.pdf (last visited: 10.05.2013).

Ivanova, Kremena et al. (2008): Evaluating a German sketch grammar: a case study on noun phrase case. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the 6th Conference on Language Resources and Evaluation. LREC 2008. Marrakech, Morocco. p. 2101-2107. Internet: http://www.lrec-conf.org/proceedings/lrec2008/pdf/537_paper.pdf (last visited: 10.05.2013).

Jurish, Bryan (2003): A hybrid approach to part-of-speech tagging. Final report. Project "Kollokationen im Wörterbuch". Berlin-Brandenburgische Akademie der Wissenschaften. 16 Sept. 2012. Internet: http://kollokationen.bbaw.de/doc/report_jurish.pdf (last visited: 10.05.2013).

Karlsson, Fred (2010): Recursion and iteration. In: Hulst, Harry van der (ed.): Recursion and human language. (= Studies in Generative Grammar 104). Berlin / New York: de Gruyter, p. 43-47.

Kilgarriff, Adam et al. (2004): The sketch engine. In: Williams, Geoffrey / Vessier, Sandra (eds.): EURALEX 2004 Proceedings. Lorient: UBS, p. 105-116. Internet: http://www.euralex.org/elx_proceedings/Euralex2004/ (last visited: 10.05.2013).

Kilgarriff, Adam / Tugwell, David (2002): Sketching words. In: Corréard, Marie-Hélène (ed.): Lexicography and natural language processing: a festschrift in honour of B. T. S. Atkins. EURALEX, Manchester: St. Jerome Publishing, p. 125-137.

McDonald, Ryan / Nivre, Joakim (2011): Analyzing and Integrating Dependency Parsers. In: Computational Linguistics 37, 1, p. 197-230.

Rychlý, Pavel (2008): A lexicographer-friendly association score. In: Sojka, Petr / Horák, Aleš (eds.): Proceedings of recent advances in Slavonic natural language processing. RASLAN 2008. Brno: Tribun EU, p. 6-9.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Towards a Firthian Notion of Collocation

Sabine Bartsch, Technische Universität Darmstadt
Stefan Evert, Friedrich-Alexander-Universität Erlangen-Nürnberg

1. Introduction

Collocations are pervasive in language. According to Altenberg (1991: 128), "roughly 70% of the running words in the corpus form part of recurrent word combinations of some kind." The investigation of such word combinations in corpora of authentic language dates back to the earliest studies of collocations by J. R. Firth (1957), who is commonly credited with introducing the concept within British Contextualism. However, serious corpus-based exploration of collocations on a larger scale has only become feasible with the arrival of the computer in the linguist's workspace in the late 20th century. Since then, a substantial number of corpora of different sizes have become available, opening up new possibilities for collocation studies and many other linguistic applications. Progress has been made in particular by harnessing ever larger corpora, a growing range of statistical measures of association (cf. Evert 2004), and state-of-the-art software tools for automatic linguistic annotation and analysis.

The purpose of the research presented in this paper is to enhance our understanding of the role played by (i) the size and composition of the corpus (ranging from reasonably sized clean, balanced reference corpora to huge, messy Web collections), (ii) automatic linguistic annotation (part-of-speech tagging, syntactic parsing, etc.), and (iii) the mathematical properties of statistical association measures in the automatic extraction of collocations from corpora. In contrast to most prior comparative evaluation studies, which focused on the extraction of lexicalised multiword expressions relevant for traditional paper dictionaries, the present study builds upon a strictly Firthian (1951/1957) definition of collocation as the habitual and recurrent juxtaposition of words with particular other words. By this approach, we hope to complement work towards data acquisition for electronic dictionaries of the future with a closer look at a type of word combinatorics that despite some considerable progress so far has proven quite difficult to grasp.

The research presented in this paper was driven by three questions: Are bigger corpora always better, or are a balanced composition and clean data more important? To what extent does automatic linguistic annotation improve collocation identification and what annotation levels are most beneficial? Can we find evidence for the postulated presence of syntactic relations between collocates (Bartsch 2004: 79), in contrast to the traditional window-based operationalization (Sinclair 1966, 1991) of the Firthian notion of collocation?

2. The notion of collocation revisited

The concept of collocation is most commonly defined as a characteristic co-occurrence of lexical items, although definitions differ in a number of details. The definition advocated by J. R. Firth, who can be credited with systematically establishing the concept in modern linguistics, holds that collocations are to be defined as the habitual and recurrent juxtaposition of semantically related words. Definitions, furthermore, differ in terms of the number of postulated lexical items assumed as constituents of collocations. Hausmann (1985), for example, assumes a binary and directional relation and permits as constituents of collocations content

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

words only to the exclusion of function words. However, function words are subsumed under the notion of collocation in Renouf/Sinclair's (1991: 128ff.) "collocational frameworks" exemplified by constructions such as 'a + … + of' as in 'a pride of lions', 'a pair of scissors' etc.

Halliday/Hasan (1976) deviate in their definition by describing collocations as "semantically related lexical items" which are more commonly interpreted in the sense of semantic field relations such as "doctor – hospital – nurse" co-occurring within the same context. Their definition rests on a tendency of lexical items to occur in the same context because they belong to the same semantic field. The common underlying assumption is that collocations are characteristic co-occurrences of related lexical items, a notion that can also be identified in Eugenio Coseriu's (1967) concept of "lexical solidarity".

Early approaches to the empirical study of collocation rest upon the manual identification of collocations in relatively small amounts of text (cf. Firth 1957), which are necessarily limited in scope and coverage. Since then, new research methods and tools as well as data have become available. The study of collocations has received fresh impetus through new computational approaches and the availability of large electronic text corpora, especially since the early 1990s. With the successively wider availability of ever larger corpora, studies of collocations have become feasible on a previously unknown scale, reaching a wider coverage of empirical data than ever before. Yet, in order to make such corpus-based studies possible and fruitful, the notion of collocation had to be not only defined, but also had to be operationalized. These new developments have brought about a further aspect in the definition of collocation, namely the definition and operationalization in terms of window-based approaches within the constrained context of a typically 3:3 or 5:5 key-word in context concordance window as proposed for example by Sinclair (1966, 1991). This approach has paved the way for the automatic investigation of collocations on the basis of relatively little linguistic pre-processing other than part of speech tagging, and by means of statistical methods modelling the characteristic co-occurrence of lexical items in terms of significance of co-occurrence as well as statistical measures of association. The current state of the art in statistical collocation identification will be discussed in the following section 3.

3. Statistical identification of collocations

Collocations are best studied on the basis of suitably large corpora of authentic language data. Their identification in corpora rests on linguistic hypotheses regarding the nature of collocations in terms of the co-occurrence of their constituents and the qualitative and quantitative relations between them. These hypotheses have to be operationalized so they can be systematically applied to corpus data; furthermore, suitable parameters have to be chosen in order to be able to distinguish instances of genuine collocations from false positives. These parameters are informed by features employed in linguistic definitions of collocations such as frequency of co-occurrence etc. In order to implement a Firthian definition of collocations, the explicitly mentioned parameter of a recurrent co-occurrence of lexical items translates directly into co-occurrence frequency in a corpus, where the context is usually taken to be a collocational span of 3 or 5 words to either side (we refer to this type of context as surface co-occurrence, following Evert 2008). For a meaningful interpretation, observed co-occurrence frequency has to be put in relation to the expected frequency of co-occurrence by chance, which can be computed from the individual frequencies of the two lexical items. For this purpose, a large number of mathematical formulae have been suggested as measures of statistical association.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Questions arising towards a corpus-based and quantitative definition of collocation entail the question of which association measure best captures our intuition of habitual, recurrent word combinations. More explicitly, the question which statistical measures are most suitable for the identification of Firthian collocations in corpora requires an answer. The parameters employed towards an operationalization of a Firthian notion of collocation as described above are thus to be implemented in terms of a suitable search space as exemplified most pervasively in terms of so-called window-based approaches which typically identify collocations as lexical co-occurrences within a 3:3 or 5:5 window; maximally, the sentence boundary is assumed as the upper limit for a collocational relation (textual co-occurrence); some approaches assume syntactic co-occurrence and postulate a syntactic relation as the context of co-occurrence (Bartsch 2004). Statistical measures are employed for the identification of collocations which, put in simple terms, rest on gauging the frequency of the co-occurrence of the collocation in relation to the independent frequencies of the constituent lexical items. A number of statistical measures have come to be used for these purposes over the years, among them widely applied measures such as the log-likelihood ratio, t-score, the Dice coefficient, and mutual information (MI), which is widely used in lexicographic contexts, as well as a number of variations of the MI formula (see Evert 2004 for a more detailed discussion). Despite the widespread use and discussion of statistically based studies of collocation, there has not, to our knowledge, been any systematic large-scale study resting on a Firthian notion of collocation. Studies typically take as their vantage point specific types of multi-word expressions (such as support verb constructions or verb-particle constructions, e.g. Baldwin 2008), or rely completely on intuitions of annotators (e.g. lexicographers' judgements). This study tries to avoid such initial assumptions and aims to evaluate association measures solely based on a Firthian notion of collocation, i.e. it works on the basis of co-occurrence in context without phraseological or lexicographical assumptions guiding the experiment. It studies the effects and suitability of the different association measures in correlation to their performance in different research settings concerning size and composition of the corpora employed, thus challenging the sometimes bluntly put assumption that when it comes to corpora bigger is always better. The research also incorporates an investigation of the effects of different levels of linguistic pre-processing and annotation on collocation identification and extraction quality. These latter sets of factors, corpus size and composition and linguistic pre-processing and annotation are of special relevance in order to gain a better understanding of properties of collocation as entailed in notions of collocation postulating that constituents of collocations must be assumed to be in a direct syntactic relation with one another (e.g. Daille 1994; Bartsch 2004).

The different statistical measures of collocations have, as yet, to be tested in terms of their performance as well as the impact of different factors concerning the corpora under study such as corpus size and composition and different types of corpus pre-processing and annotation ranging from lemmatized and part of speech tagged corpora to corpora that have undergone syntactic parsing and thus allow testing the above mentioned hypothesis of a direct syntactic relation as a constraint on relations obtaining between constituents of collocations.

The next section discusses the ways in which these statistical measures were tested on corpora of different size and composition and what types of linguistic pre-processing and annotation have an impact on the performance of different statistical measures for collocation identification.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

4. Research set-up

The present study rests upon a study of collocations in authentic text corpora of different size and composition, which are annotated at different levels of linguistic organisation. The amount and depth of annotation range from entirely unannotated plain text corpora, over corpora with part of speech tagging and lemmatization, to syntactically parsed corpora. The corpora range in size between the 100-million-word British National Corpus and a Web corpus of ca. 2 billion words (ukWaC). Linguistic annotations include part-of-speech tagging, lemmatization and syntactic analysis by means of a dependency parser.

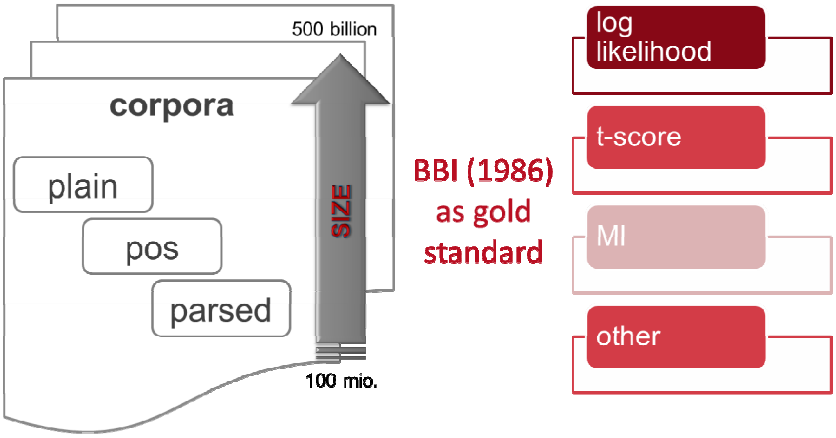

Figure 1: Research scenario

As a gold standard for our evaluation study, we use combinations of lexical words found in the entries of the BBI Combinatory Dictionary (Benson et al. 1986), a pre-corpus collocation dictionary that we consider to come very close to a Firthian definition of collocations. We compare (i) a standard range of well-known association measures (including log-likelihood, t-score, the Dice coefficient, co-occurrence frequency and different variants of mutual information), (ii) corpora of widely different sizes and composition (ranging from the 100-million-word British National Corpus to the ukWaC Web corpus comprising approx. 2 billion words), and (iii) different sizes of co-occurrence context (ranging from direct syntactic relations to co-occurrence within the same sentence) (see figure 1).

British National Corpus (BNC, Aston/Burnard 1998)	100 M
CLAWS tagger, lemmatised, C&C dependency parser (Curran/Clark/Bos 2007)	
Subset of Wackypedia (first 500 words of selected articles)	200 M
TreeTagger POS & lemmatisation (Schmid 1995), MaltParser (Nivre/Hall/Nilsson 2006)	
English Wikipedia @ 2009 (Wackypedia, available from wacky.sslmit.unibo.it)	850 M
TreeTagger POS & lemmatisation (Schmid 1995), MaltParser (Nivre/Hall/Nilsson 2006)	
Web corpus of British English (ukWaC, Baroni et al. 2009)	2,000 M
TreeTagger POS & lemmatisation (Schmid 1995), MaltParser (Nivre/Hall/Nilsson 2006)	

Table 1: Overview of corpora, corpus sizes and pre-processing and annotation

The quantitative task (section 5 below) entails the automatic identification of collocation candidates on the basis of well-known statistical association measures such as log-likelihood (G^2), t-score (t), Mutual Information (MI), and the Dice coefficient to name but some of the most widely used ones (see Evert 2008 for details). The aim of this aspect of the study is a more thorough investigation of the factors influencing results of collocation evaluation tasks as findings remain inconclusive as to which association measure is most useful depending on factors such as language, type of multiword expression as well as corpus size and composition. For example, MI is very popular among computational lexicographers despite a well-

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

known bias towards low-frequency data, while computational linguists prefer measures based on statistical hypothesis tests such as G^2 (following Dunning 1993). Recent versions of the SketchEngine (Kilgarriff et al. 2004) use a variant of the Dice coefficient for collocation identification, even though it has never been identified as a top-performing measure in comparative evaluation studies.

Another issue, which has not been systematically addressed yet, concerns the ideal corpus size and composition for statistical collocation identification. More bluntly put, we ask whether "bigger is better", as the recent trend towards large Web corpora presupposes, or whether representativeness and high-quality data preparation are more important than sheer size.

In terms of corpora suitable for collocation analysis an additional issue concerns the question of the impact of amount and level of corpus annotation, i.e. the question in how far collocation studies can benefit from linguistic annotations. A Firthian notion of collocations assumes mere "habitual" co-occurrence and has usually been implemented for very practical reasons as co-occurrences of lexical items that occur a minimum number of times, usually at least five times, the practical side of this decision also being the reduction of the amount of data coming under study. Most definitions of collocation tacitly assume that collocations are potentially specific to the morpho-syntactic class of the word and that different word forms at least potentially enter into similar sets of collocations, although this latter assumption has been called into question on occasion; in order to study collocations across the entire paradigm of a lexical item, most studies are based on lemmatized and part of speech tagged corpora. At the extreme end of definitions, syntactic relations are posited to obtain between the constituents of a collocation and that these, consequently, are best identified on the basis of suitable syntactic annotations, typically either parse trees or dependency parses.

In this research set-up, we are envisioning all of these possibilities comparatively with the aim and hope that we can not only test the Firthian notion of collocation, but also to check whether definitions assuming more intricate linguistic relations between the constituents of collocations might not actually improve the quality of collocation extraction and thereby help to improve the coverage of collocations in lexicographical works based on automatic and statistically driven approaches.

In this study, we follow the well-established paradigm of Evert/Krenn (2001) for the comparative evaluation of association measures. In this approach, a set of candidate word pairs is ranked according to different association measures. The quality of each ranking is then assessed based on how many true collocations (true positives, TPs) are found among the n highest-ranked candidates (an n-best list), compared to a gold standard of known collocations. The percentage of TPs in such an n-best list is called the n-best precision of the ranking. Similarly, n-best recall is the percentage of true collocations in the gold standard that are found among the n highest-ranked candidates. Since it is not clear in most cases what number n of ranked candidates should be considered, evaluation results are usually presented visually by plotting n-best precision against n-best recall for many different n-best lists, producing a precision-recall curve as shown in Fig. 2. For example, the black line (for a ranking based on the log-likelihood measure) shows that if enough highest-ranking candidates are considered to find 10% of the gold standard collocations, the n-best precision is 30%. With this intuitive visualization, it is easy to compare the performance of the different association measures with all other factors (corpus size-composition, collocational span, corpus annotation) being equal.

As a gold standard, we use lexical collocations from the BBI Combinatory Dictionary (Benson et al. 1986), which we believe to correspond well to the Firthian definition of collocations

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

and which avoid a bias in favour of a particular corpus or collocation identification method (unlike more recent corpus-based collocation dictionaries). Since the BBI is not available in electronic form, we manually selected 224 entries from the dictionary, based on corpus frequency of the headwords, examples from the literature and previous linguistic analyses of English collocations. All lexical collocates (i.e. nouns, verbs, adjectives and adverbs) listed in these entries were transcribed, resulting in a gold standard of 2,949 lexical collocations. Note that the collocations in the gold standard are directed, consisting of one of the 224 headwords and a lexical collocate. In total, there are 1,849 distinct collocates in the gold standard.

Candidate data were extracted from the four corpora listed in Table 1. In order to ensure a fair comparison between corpora of different size and composition, we prepared a list of 7,711 lexical words that are considered as potential collocates. This list includes all 1,849 collocates from the gold standard and was extended with nouns, verbs, adjectives and adverbs that fall into the same frequency range in the British National Corpus. Collocation candidates are thus all co-occurrences of one of the 224 headwords with one of the 7,711 potential collocates. No frequency thresholds were applied since the goal is to identify as many known collocations from the BBI as possible.

From each of the four corpora, we extracted collocation candidates for five different context settings:

- Surface co-occurrence with a collocational span of 3 words (L3 / R3)
- Surface co-occurrence with a collocational span of 5 words (L5 / R5)
- Surface co-occurrence with a collocational span of 10 words (L10 / R10)
- Textual co-occurrence within sentences
- Syntactic co-occurrence, where the two words must occur in a direct syntactic relation (according to the MaltParser or C&C analyses)

corpus / context	# candidates	coverage
BNC syntactic	201397	91.69%
BNC span 3 words	349372	95.56%
BNC span 5 words	455020	96.54%
BNC span 10 words	584723	97.49%
BNC sentence	735191	98.17%
Wackypedia 200M syntactic	239956	91.66%
Wackypedia 200M span 3 words	378468	94.17%
Wackypedia 200M span 5 words	483649	95.15%
Wackypedia 200M span 10 words	612099	96.41%
Wackypedia 200M sentence	766944	97.25%
Wackypedia syntactic	396504	96.64%
Wackypedia span 3 words	588229	97.73%
Wackypedia span 5 words	717994	98.27%
Wackypedia span 10 words	864365	98.68%
Wackypedia sentence	1034844	98.95%
ukWaC syntactic	544198	98.58%
ukWaC span 3 words	759158	99.15%
ukWaC span 5 words	898030	99.29%
ukWaC span 10 words	1048637	99.39%
ukWaC sentence	1256383	99.49%

Table 2: Number of candidates in each data set and coverage of the BBI-derived gold standard

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Table 2 shows the number of collocation candidates found in each setting and the corresponding coverage of the BBI gold standard, i.e. how many of the BBI collocations co-occurred at least once in the respective corpus and type of context. With all coverage values well above 90%, it is possible at least in principle to extract Firthian collocations in the sense of the BBI even from a relatively small corpus such as the BNC.

Not surprisingly, coverage increases for larger corpora and context windows with a coverage of up to > 99% for the ukWaC Web corpus. It likewise comes as no surprise that the syntactic dependency filter reduces the coverage, partly due to collocation candidates that do not form grammatical units after all or for which a direct grammatical relation does not show in the parse tree, partly due to parsing errors. It should be noted that the values shown here include all combinations, even if they just occur once and thus cannot be recognised as collocational by a quantitative analysis.

The statistical association measures tested in this study were selected based on the recommendations of Evert (2008):

- log-likelihood measure (G^2) (Dunning 1993)
- t-score (t) (Church et al 1991; *pace* Evert 2004: 82f.)
- Mutual Information (MI) (Church / Hanks 1990)
- Dice coefficient (*Dice*) (as employed in the Sketch Engine; Kilgarriff et al. 2004)
- Ranking by co-occurrence frequency (f) as a baseline

In addition, we considered several variants of MI that aim to reduce its low-frequency bias by raising the observed co-occurrence frequency to the k-th power. From this family of MI^k measures we selected MI^2, since it consistently gave the best results in preliminary experiments.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

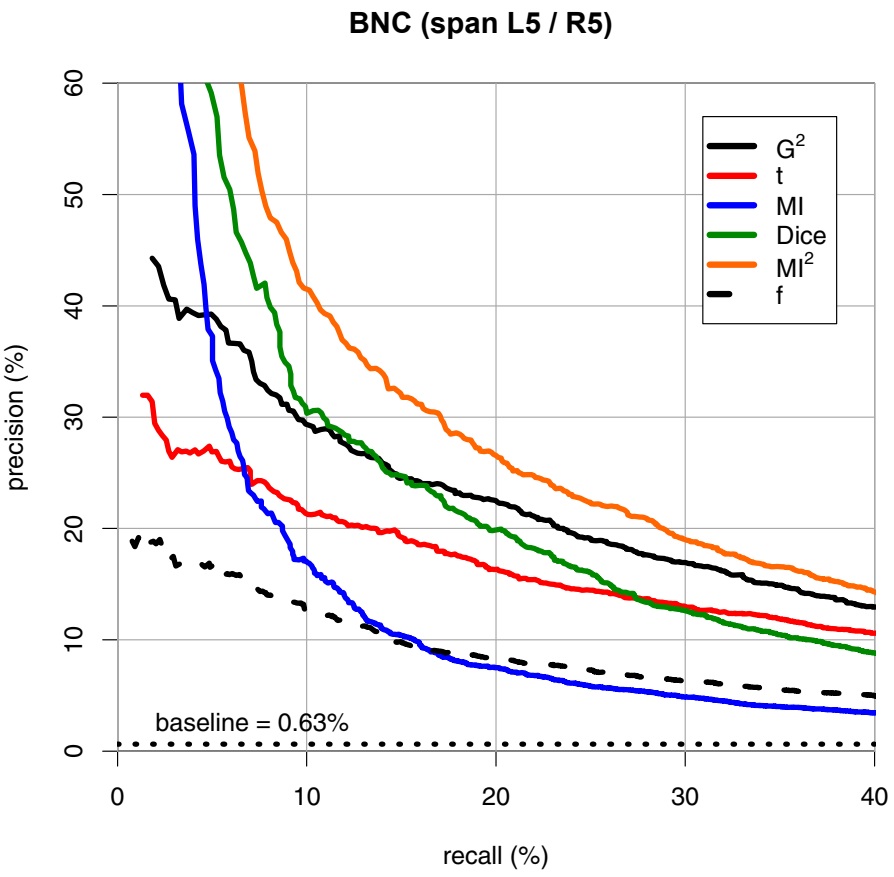

Figure 2: Evaluation of different association measures on British National Corpus with surface context (symmetric span of 5 tokens to the left and right)

Fig. 2 shows evaluation results for the British National Corpus, a "traditional" balanced reference corpus that has been widely used for collocation extraction and other tasks in computational lexicography. Candidates were obtained based on surface co-occurrence in a L5 / R5 window, which is a typical setting for "traditional" collocation studies. This evaluation generally conforms with our expectations and the results of previous studies: G^2 (black) and *t*-score (red) perform significantly better than plain co-occurrence frequency *f* (dashed black line). All association measures achieve much higher precision than the very low baseline corresponding to a random ranking of the candidates. The blue line once again confirms the well-known observation that MI performs poorly without frequency thresholds, due to its low-frequency bias. The Dice coefficient (green) displays surprisingly good performance and achieves higher accuracy than log-likelihood up to 15% recall; its performance drops drastically if a better coverage of the gold standard is required (above 25% recall). This observation explains why Dice was selected by the Sketch Engine developers: lexicographers reading the word sketches will typically focus on a few highly salient collocations. The uniformly best performance is achieved by MI^2 (orange) which outperforms the other association measures across all recall points. Results for other corpora and settings are qualitatively very similar; in almost all cases, MI^2 is the uniformly best association measure.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

5. Quantitative insights

Ever since quantitative and statistical approaches have come to be applied to electronic corpora of substantial size, the dictum of "bigger is better" has been treated almost as a natural law. However, there have also always been cautioning voices warning that size might come with a price, especially concerning corpora whose sheer size forbids careful manual intervention to improve the quality of the data and the annotation. Furthermore, the very large corpora available today are typically collected opportunistically from the web and thus, despite all cleaning efforts, are neither balanced nor can the text base be assumed to be as clean as that of smaller, carefully crafted corpora such as the BNC. Nevertheless, there is a certain appeal to corpora that are big enough to promise to fulfil the dream of being representative or at least overcoming the obvious gaps of coverage in corpora of lesser size.

The study of lexical phenomena is a classical case where, due to the highly skewed distribution of lexical items, large corpora are often assumed to be of paramount importance, and the same assumption is consequently made for the study of collocations.

In order to test the effects of corpus size and quality on the accuracy of collocation identification, the different association measures were tested on the two-billion-word Web corpus ukWaC, which is opposite to the BNC at the "large and dirty" end of the spectrum. Figure 3 shows the results of this evaluation.

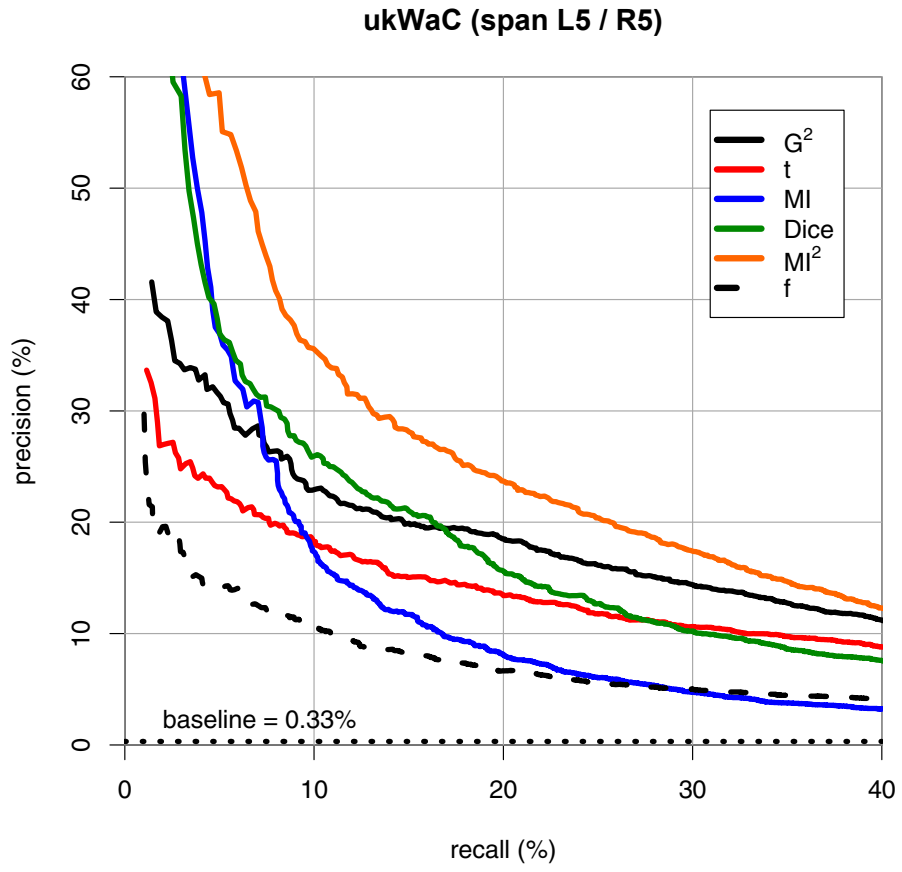

Figure 3: Evaluation of different association measures on the ukWaC Web corpus with surface context (symmetric span of 5 words to the left and right)

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

It turns out that the precision values are lower overall, but that the different association measures show a very similar pattern as compared to the BNC or, indeed, the other corpora under study. Again, MI^2 is uniformly the best-performing measure. These findings might lead one to modify the expectation of the impact of corpus size and quality to "bigger is worse" or "composition is more important than size". The following experiments take this modified hypothesis into account and focus on MI^2 as an association measure.

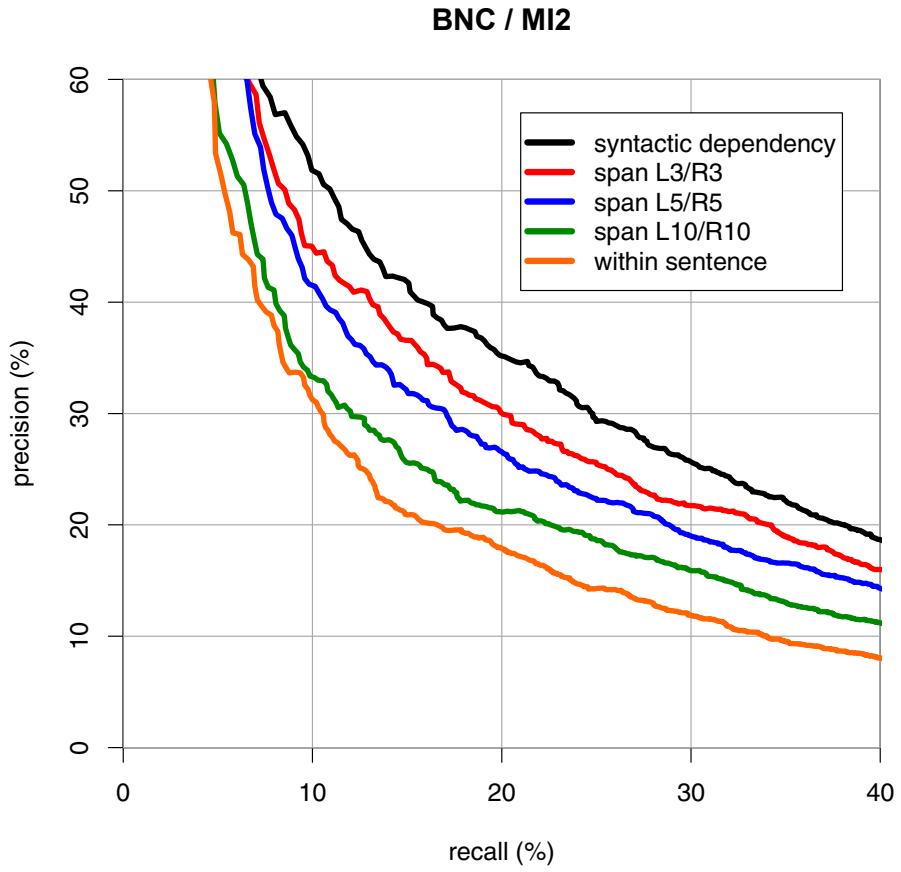

Figure 4: Comparison of different co-occurrence contexts for the British National Corpus and the best-performing association measure MI^2

The next question to be tested is the impact of the context type and size on the quality of the extraction task. In this set of experiments, contexts were tested ranging from narrow and specific (syntactic dependency) via increasing window sizes (L3 / R3, L5 / R5 and L10 / R10) to co-occurrence within a full sentence as the broadest context under consideration. Fig. 4 shows the results of this comparison for the British National Corpus.

These precision-recall graphs confirm that the larger the contexts become, the lower the precision drops. Thus, the smaller L3 / R3 span is better than the commonly used L5 / R5 span. Maybe surprisingly for some, the assumption of a syntactic relation between the constituents of collocations leads to even better results thus confirming the claims made by Bartsch (2004) that the Firthian notion of collocation should be supplemented with the additional criterion of a direct syntactic relation between the constituents.

The same pattern could be confirmed for all corpora, although not all plots are shown and discussed here in detail (due to space constraints): larger contexts lead to lower precision. It can thus be confirmed in turn that the component words of Firthian collocations tend to occur closely together and are usually in a direct syntactic relation. In the following experiment, we therefore focus on syntactic co-occurrence contexts for the comparison of different corpora.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

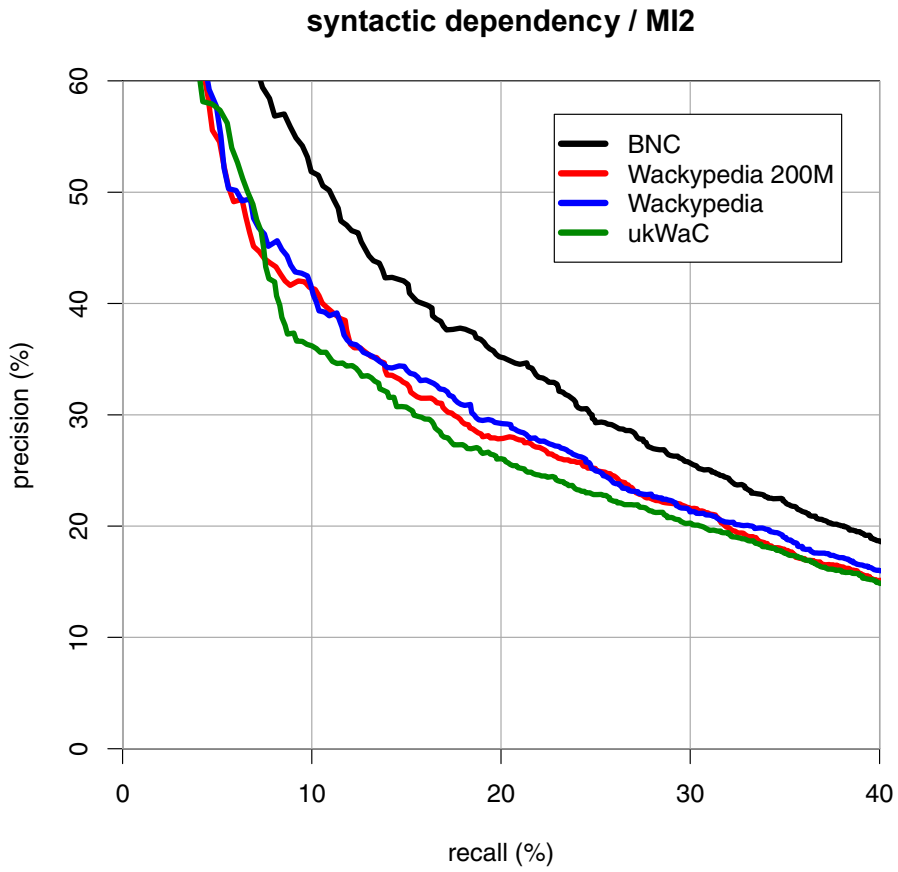

Figure 5: Comparison of different corpora for the best-performing combination of syntactic dependency co-occurrence and association measure MI^2

As already indicated above, initial observations suggest that increasing corpus size does not improve the performance of collocation extraction, especially if the larger corpora are less balanced and "clean". The graphs in Fig. 5 above confirm our new hypothesis that bigger is indeed worse. It appears that the larger and messier the corpora, the lower the precision drops. They also show that it is obviously composition and cleanliness that matter rather than size: the precision-recall curve for the full Wackypedia (approx. 850M words) and a subset of 200M words are practically identical, even though the former corpus is more than four times as large as the latter. The lowest precision is obtained from the large and messy Web corpus ukWaC.

However, there is one confounding factor that needs to be mentioned here and that requires further testing: ukWaC and the Wackypedia corpora were analysed by means of the Malt-Parser (syntactic annotations are included in the official distribution of the corpus) while the BNC was parsed by means of C&C, a sophisticated Combinatory Categorial Grammar (CCG) parser (Clark/Curran 2004; Curran/Clark/Bos 2007). There is thus a possibility that C&C is simply more accurate or covers important direct relations that are not generated by the Malt-Parser. Further testing of annotation quality is thus planned for the near future. In particular, we intend to re-annotate ukWaC and Wackypedia with C&C so that we will be able to gauge the impact of differences between the parsers.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

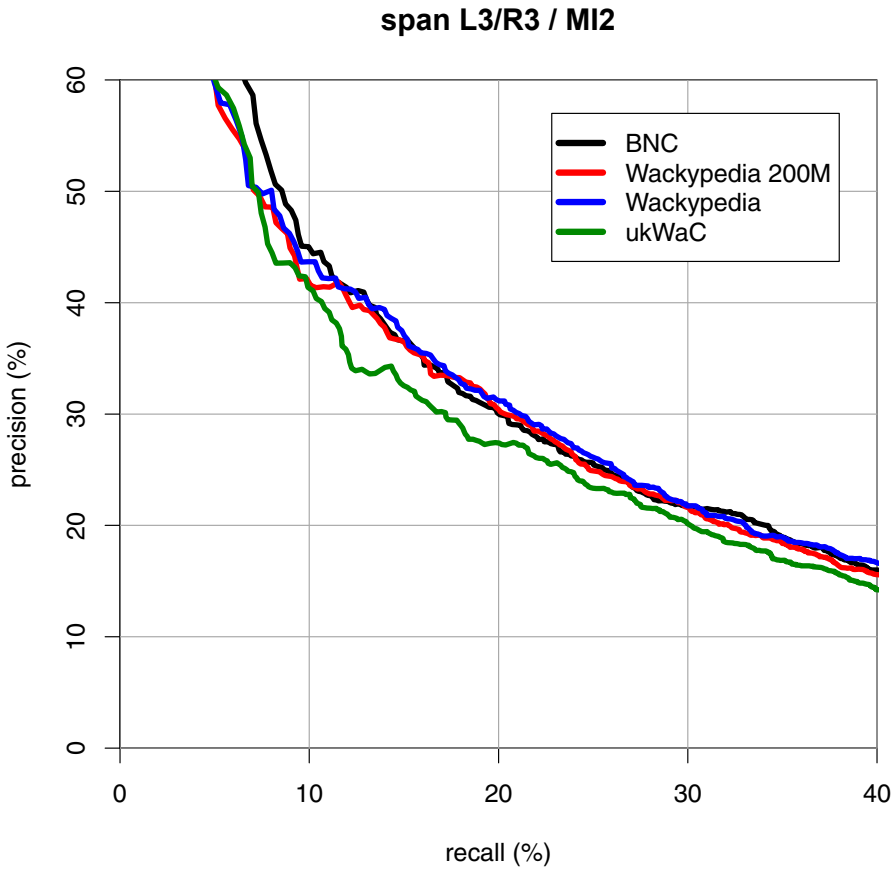

Figure 6: Comparison of different corpora for the best-performing collocational span (L3 / R3) and association measure MI^2

For the time being, in order to exclude the parser as a source of differences, we show a comparison on the basis of the smallest (and best-performing) surface context using a collocational span of 3 words (Fig. 6). In this experiment, there is no discernible difference between BNC, the Wackypedia subset and the full Wackypedia. The web corpus ukWaC even yields slightly worse results, despite its substantially larger size of approx. 2 billion words.

These findings lead to three conclusions: (i) increasing corpus size does not matter at all, at least for the identification of BBI collocations (since corpora ranging from 100M to 850M words yield virtually indistinguishable results); (ii) Wikipedia is a good replacement for the BNC as a reference corpus (at least with regard to Firthian collocations); (iii) the messiness of web data results in lower precision, even though the larger size should improve statistical analyses (and even though a web corpus might be expected to contain a broader range of genres than Wikipedia).

6. Conclusions

Based on the findings reported in this paper, we believe that some tentative conclusions can be drawn. The first one concerns corpus size where it could be shown that larger corpora do not necessarily lead to better results. This goes even for relatively simple statistical analyses (association measures) that should benefit from larger samples.

The second conclusion concerns corpus composition and suggests that composition and cleanness of a corpus are more important than corpus size. This is not to say that Web corpora

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

might not be useful, but it suggests that their usefulness is enhanced if clean and more balanced samples can be obtained without compromising size (i.e. ≥ 2G words).

Collocations have been shown by some studies to tend to form grammatical relations, thus, assuming a syntactic dependency context is optimal for an identification of Firthian collocations. However, the practical benefit of taking this approach depends on the accuracy of the parser and the set of syntactic dependency relations recognized. In our case, it looks as if the linguistically optimised C&C parser is much better suited to the task than the fast off-the-shelf MaltParser used by the WaCky team.

The experiments with recent corpora such as the 200M subset of Wackypedia suggest that it is on par with BNC in collocation studies based on surface context. This result is very encouraging because it suggests that such corpora might be a useful substitute for languages such as German for which no standard general reference corpus like the British National Corpus is publicly available.

7. References

Altenberg, Bengt (1991): Amplifier collocations in Spoken English. In: Johansson, Stig / Stenström, Anna-Brita (eds.): English computer corpora. Selected papers and research guide. Berlin / New York: de Gruyter, p. 127-147.

Aston, Guy / Burnard, Lou (1998): The BNC Handbook. Edinburgh: Edinburgh University Press. http://www.natcorp.ox.ac.uk/ (last visited: 10.05.2013).

Baldwin, Timothy (2008): A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Marrakech, Morocco, p. 1-2. Internet: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf (last visited: 10.05.2013).

Baroni, Marco et al. (2009): The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. In: Language Resources and Evaluation 43, 3, p. 209-226.

Bartsch, Sabine (2004): Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence. Tübingen: Narr.

Benson, Morton / Benson, Evelyn / Ilson, Robert (1986): The BBI Combinatory Dictionary of English: A Guide to Word Combinations. Amsterdam / New York: John Benjamins.

Church, Kenneth W. / Hanks, Patrick (1990): Word association norms, mutual information, and lexicography. In: Computational Linguistics 16, 1, p. 22-29.

Church, Kenneth / Gale, William A. / Hanks, Patrick / Hindle, Donald (1991): Using statistics in lexical analysis. In: Zernick, Uri (ed.): Lexical Acquisition: Using On-line Resources to Build a Lexicon. Hillsdale, NY: Lawrence Erlbaum, p. 115-164.

Clark, Stephen / Curran, James R. (2004): Parsing the WSJ using CCG and Log-Linear Models. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). Barcelona, Spain 2004, p. 104-111. Internet: http://aclweb.org/anthology-new/P/P04/#1000 (last visited: 10.05.2013).

Coseriu, Eugenio (1967): Lexikalische Solidaritäten. In: Poetica 1, p. 293-203.

Curran, James / Clark, Stephen / Bos, Johan (2007): Linguistically motivated large-scale NLP with C&C and Boxer. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions. Prague, Czech Republic. Madison: ACL, p. 33-36. Internet: http://www.aclweb.org/anthology-new/P/P07/P07-2.pdf (last visited: 10.05.2013).

Daille, Béatrice (1994): Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. Ph.D. thesis, Université Paris 7. Internet: http://www.bdaille.com/index.php?option=com_docman&task=doc_download&gid=8&Itemid (last visited: 10.05.2013).

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Dunning, Ted E. (1993): Accurate methods for the statistics of surprise and coincidence. In: Computational Linguistics 19, 1, p. 61-74.

Evert, Stefan / Krenn, Brigitte (2001): Methods for the qualitative evaluation of lexical association measures. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France, p. 188-195. Internet: http://aclweb.org/anthology-new/P/P01/ (last visited: 10.05.2013).

Evert, Stefan (2004): The statistics of word cooccurrences: word pairs and collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Veröffentlicht 2005. URN: urn:nbn:de:bsz:93-opus-23714.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke / Kytö, Merja (eds.): Corpus Linguistics. An International Handbook. Chapter 58. Berlin: de Gruyter.

Firth, John R. (1951/1957): Modes of meaning. In: Papers in Linguistics, 1934-1951. Oxford: Oxford University Press.

Halliday, MAK / Hassan, Ruqaiya (1976): Cohesion in English. London: Longman.

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, Henning / Mugdan, Joachim (eds.): Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch. (= Lexikographica, Series Maior 3). Tübingen: Niemeyer, p. 118-129.

Kilgarriff, Adam et al. (2004): The Sketch Engine. In: Williams, Geoffrey / Vessier, Sandra (eds.): EURALEX 2004 Proceedings. Lorient: UBS, S. 105-116. Internet: http://www.euralex.org/elx_proceedings/Euralex2004/ (last visited: 10.05.2013).

Nivre, Joakim / Hall, Johan / Nilsson, Jens (2006): MaltParser: a data-driven parser-generator for dependency parsing. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy, p. 2216-2219.

Renouf, Antoinette / Sinclair, John (1991): Collocational frameworks in English. In: Aijmer, Karin / Altenberg, Bengt (eds.): English corpus linguistics. New York: Longman, p. 128-143.

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT Workshop. Dublin, p. 47-50. Internet: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf (last visited: 10.05.2013).

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

Sinclair, John (1966): Beginning the study of lexis. In: Bazell, Charles E. et al. (eds.): In Memory of J. R. Firth. London: Longmans, p. 410-430.

© 2014 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.