Marc Kupietz

DER PROGRAMMBEREICH KORPUSLINGUISTIK AM IDS: GEGENWART UND ZUKUNFT

Von Anfang an war als eine wichtige Aufgabe des IDS festgeschrieben, im Rahmen seiner Forschungsaufgaben die Möglichkeiten eines empirisch orientierten Arbeitens zu nutzen und voranzutreiben (s. z.B. Kupietz et al. 2010b sowie den Beitrag von Teubert/Belica in diesem Band). Der Programmbereich Korpuslinguistik, der 2004 entstanden ist und Anfang 2009 direkt dem Vorstand untergeordnet wurde, ist heute am IDS dafür verantwortlich, die dazu notwendigen Grundlagen zu schaffen und für die Zukunft zu sichern. Zu diesen Grundlagen gehört erstens der Ausbau und die Pflege der Korpora geschriebener Gegenwartssprache – insbesondere des Deutschen Referenzkorpus DeReKo, zweitens die Entwicklung von Methoden zur Korpusanalyse und ihrer linguistischen Erschließung und drittens zunehmend auch wieder die Entwicklung der Technologien, die notwendig sind, um die aufgebauten Korpora und entwickelten Methoden für die eigene und externe sprachwissenschaftliche Forschung nutzbar und handhabbar zu machen.

Hintergrund

Das wissenschaftliche Programm der Korpuslinguistik am IDS ist es, durch die explorative Analyse von sehr großen Sammlungen natürlichsprachlicher Daten neue Einsichten in die Strukturen, Gesetzmäßigkeiten, Eigenschaften und Funktionen von Sprache zu erlangen. Vor diesem Hintergrund wird einerseits eine Reihe von methodologischen Forschungszielen formuliert, die auf Fortschritte bei der Entwicklung von strukturentdeckenden korpusanalytischen Methoden ausgerichtet sind und verschiedene grundlegende Fragestellungen der Sprachwissenschaft aufgreifen. Andererseits wird durch systematische Generalisierungen der so gewonnenen Erkenntnisse die Beurteilung bestehender und die Formulierung neuer, empirisch fundierter linguistischer Hypothesen und formaler Modelle angestrebt. Die in Korpora aufgezeichneten Resultate von Kommunikationsprozessen werden dabei als empirische Grundlage sowohl für die explorative Analyse als auch für die induktive, auf Theoriebildung zielende Generalisierungsstrategie verstanden. Obwohl der Ansatz von lexikalischen Einheiten und deren Kontexten ausgeht, sind hier die lexikalische, syntaktische und semantische Ebene nicht voneinander getrennt: Eine fundamentale Rolle im postulierten Lexikon-Syntax-Kontinuum fällt dabei dem mit Hilfe von mathematisch-statistischen, musterorientierten Methoden in empirischen Sprachdaten operationalisierten

und um Varianz und Vielgliedrigkeit erweiterten Begriff der Kookkurrenz zu. Diese Herangehensweise bezweckt das Aufdecken präferenzrelationaler Gesetzmäßigkeiten, die unter anderem dadurch gekennzeichnet sind, dass sie in Abhängigkeit von pragmatischen, sprachlichen und außersprachlichen Faktoren nicht primär regelbasiert variieren. Es können außerdem auch subtile sprachliche Strukturen aufgespürt werden, die dem Sprachgefühl individueller Sprachteilnehmer unzugänglich sind und erst durch die Analyse großer Datenmengen erschlossen werden können.
Basierend auf diesen allgemeinen Voraussetzungen hat der Programmbereich ein *Empirisch-linguistisches Forschungsprogramm* formuliert (Kupietz/Keibel 2009), das u.a. von Hoppers Emergent Grammar (1987) inspiriert ist, dem sogenannten *usage-based framework* (z.B. Langacker 2000; Tomasello 2003) zuzurechnen ist und die Grundlage für u.a. die eigenen zukünftigen Forschungsarbeiten legt. Wichtige Eckpfeiler des Programms sind eine möglichst unvoreingenommene und prämissenarme Vorgehensweise, eine explanatorische Ausrichtung, eine emergentistische Sichtweise auf Sprache, ein Verzicht auf den Anspruch globaler Widerspruchsfreiheit der Modellierungen und die zentrale Rolle eines weitgefassten Konzepts von Ähnlichkeit (Belica 2011). Ausgehend von der Sichtweise, dass Grammatik und in der Sprache beobachtbare Strukturen schlechthin im Wesentlichen ein Epiphänomen des Sprachgebrauchs sind, die aus der Interaktion zwischen Sprachteilnehmern und ihren Spracherfahrungen evolvieren, soll vor allem basierend auf Korpora und korpuslinguistischer Methodik aber auch unter Einbeziehung anderer Datentypen latentes sprachliches Wissen rekonstruiert bzw. im Sinne einer fein-granularen Erfassung kontextsensitiver, sprachlicher Regularitäten modelliert werden. Kernidee dieser Rekonstruktion oder Modellierung ist dabei, das Korpus als ein unvollkommenes, aber der empirischen Forschung zugängliches Modell des Ergebnisses (vgl. Belica et al. 2010) der Spracherfahrungen eines Sprechers/Hörers aufzufassen.

Methoden der Korpusanalyse und -erschließung

Im gleichnamigen Projekt des Programmbereichs findet korpuslinguistische Grundlagenforschung statt, die einerseits der Korpora bedarf und andererseits die Grundlagen für ihre linguistische Interpretation legt, die dann z.T. als Analyse- und Erschließungsmethoden in die Weiterentwicklung des IDS-Recherche- und Analyse-Systems eingehen. In Fortführung der *Approaching-Grammar*-Reihe (Keibel/Kupietz 2009; Keibel et al. 2008, 2011) soll z.B. das Konzept der emergenten Grammatik durch neue Methodologien zur Erkundung der Verzahnung von kohäsiven syntagmatischen Strukturen und zur Untersuchung von paradigmatischen Variationen innerhalb dieser

weiterentwickelt werden (Keibel et al. 2011). Um die oben genannten graduellen Regularitäten und Ähnlichkeiten adäquat abbilden zu können, wird es dabei wichtig sein, neue Repräsentationsformalismen zu entwickeln und zu etablieren. Diese können zum einen in Form von Informationsvisualisierungen direkt der linguistischen Erschließung von statistischen Analysen dienen und aus großen Datenmengen etwa gezielt bestimmte Aspekte herausheben. Zum anderen werden aber Repräsentationen auch unabhängig von einer Visualisierung benötigt, um die Elemente einer linguistischen Modellierung, z.B. Konzepte wie *syntagmatische Muster*, fassen zu können. Hier sind teilweise neue Wege zu gehen und neue Ebenen in der Erklärungshierarchie zu betreten (s. Smolensky 1988; Kupietz 1996, S. 12ff.), da die herkömmliche Terminologie (z.B. der Begriff Lesarten) und herkömmliche Formalismen (z.B. Phrasenstrukturgrammatiken) nicht feingranular genug sind und solche aus der statistischen maschinellen Sprachverarbeitung allein zu feingranular sind, um mit ihrer Hilfe korpuslinguistisch relevante Phänomene adäquat zu erfassen und auf der konzeptuellen Ebene die Abduktion neuer Hypothesen zu erleichtern (s. Kupietz 1996, S. 22f. sowie 2002). Ein Beispiel sowohl für den Aspekt der Visualisierung als auch den Aspekt der Modellierung sind die von Cyril Belica (2006, 2011) entwickelten Modelle semantischer Beziehungen. Diese können zum einen als ausgesprochen nützliche Visualisierung der im Korpus vorgefundenen Ähnlichkeitsbeziehungen zwischen Kookkurrenzprofilen aufgefasst und entsprechend interpretiert werden (s. z.B. Marková 2012). Darüber hinaus können sie aber auch entsprechend ihrer ursprünglichen Intention als (psycho-)linguistische Modelle von Aspekten des latenten lexikalischen Wissens betrachtet werden. Als solche können aus ihnen zahlreiche Hypothesen abgeleitet werden, z.B. zum Erlernen von semantischen Beziehungen anhand des Auftretens von Wörtern in ähnlichen Verwendungskontexten sowie zur Ähnlichkeit von Verwendungskontexten und der semantischen Nähe selbst. Um an dieser Stelle weitere Erkenntnisse zu gewinnen und die verschiedenen Komponenten des Modells zu verbessern, sollen zukünftig verstärkt experimentell gewonnene Daten, wie z.B. Ergebnisse aus Experimenten zum lexikalischen Priming und zur Satzverarbeitung (s. z.B. Elman 2011), einbezogen werden.

Korpusausbau

Das vom Programmbereich gepflegte und laufend erweiterte Deutsche Referenzkorpus DeReKo (Kupietz et al. 2010a) ist mit 6 Milliarden Wortformen (IDS 2013) eine der größten Korpussammlungen überhaupt. Es ist als *very large general purpose corpus* konzipiert, mehrfach morphosyntaktisch annotiert (s. Belica et al. 2011) und dient dem Programmbereich selbst, dem IDS und

der synchronen germanistischen Sprachwissenschaft insgesamt als empirische Grundlage. Im Unterschied zu anderen bekannten Korpora strebt DeReKo keine wie auch immer geartete Ausgewogenheit an, sondern soll vielmehr als Ur-Stichprobe des Schriftsprachgebrauchs dienen, aus der sich möglichst viele Nutzer selbst bzgl. ihrer Fragestellungen möglichst repräsentative virtuelle Korpora zusammenstellen können (s. Kupietz/Keibel 2009 sowie den Beitrag von Teubert/Belica in diesem Band).

Oberste Maxime des DeReKo-Ausbaus ist entsprechend dieser Konzeption die Maximierung von Umfang und Dispersion bzgl. potenziell relevanter Strata. In der Praxis müssen bei der Auswahl neuer Texte jedoch (leider) weitere Kriterien einfließen, nämlich insbesondere die Kosten für die Beschaffung der notwendigen Nutzungsrechte und die Kosten für die Aufbereitung der Rohdaten. Anders als noch zu Zeiten der Konzeption des British National Corpus (Aston/Burnard 1998) gab es vor allem in der jüngeren Vergangenheit je nach Textsorte und Quelle große Diskrepanzen bzgl. dieser Kriterien (vgl. Kupietz i. Vorb.). Während Zeitungsverlage in der Regel früh auf eine elektronische Vermarktung gesetzt haben und ihre gesamte Produktionskette auf ein *Single-Source-Publishing* beginnend mit einem elektronischen Redaktionssystem für die Autoren und endend mit der traditionellen Print-Ausgabe, verschiedenen e-Paper- und einem Export-Format zur Weiterverarbeitung durch verschiedene elektronische Archive, das auch vom Korpusausbauprojekt leicht genutzt werden kann, umgesetzt haben, vollzieht sich ein solcher Trend auf dem deutschen Buchmarkt erst seit kurzem und noch recht langsam. Unter anderem dieser Umstand führt dazu, dass das Verhältnis der Kosten für Akquisition und Aufbereitung von Zeitungen im Vergleich zu Romanen zurzeit pro Wort gerechnet bis zu 1:25.000 beträgt. Für die Zukunft ist zwar eine Verkleinerung dieser Diskrepanz zu erwarten, aber nicht nur im positiven Sinne: Während insbesondere größere überregionale Zeitungen bereits jetzt in der Lage und auch darauf angewiesen sind, mit der elektronischen Verwertung Geld zu verdienen – was sie bereits jetzt für DeReKo sehr oder auch zu teuer machen – werden in einigen Jahren nicht nur ein breites Spektrum an Neuerscheinungen sondern auch ältere Bücher als E-Books in leichter verarbeitbaren Formaten verfügbar sein. Zu hoffen bleibt dabei, dass sich damit auch der Produktionsweg von einer nachträglichen Aufbereitung durch Drittfirmen in Richtung eines Single-Source-Publishing verändern wird, was die Datenqualität verbessern und – durch die Vermeidung von Rechten Dritter an der elektronischen Aufbereitung – die Lizensierung vereinfachen würde.

Die Lizensierungsproblematik betrifft leider auch und besonders die unerschöpfliche Textquelle des World Wide Web. Da Konzepte aus der US-amerikanischen Rechtsprechung wie *fair use* und *implied license* dem deutschen und europäischen Rechtssystem fremd sind, muss im Prinzip mit jedem

einzelnen Rechteinhaber ein Lizenzvertrag abgeschlossen werden, in dem der Lizenzgeber auch zusichern müsste, überhaupt die notwendigen Rechte zu besitzen. Um DeReKo's Abdeckung spezieller Hypertextsorten nachhaltig zu verbessern, wird die Strategie des Projekts daher weiterhin sein, sich primär auf solche Textsammlungen zu konzentrieren, die bereits unter ausreichend liberalen Lizenzen veröffentlicht sind oder deren Nutzungsrechte – etwa über große Foren- oder Blog-Provider – zentral verhandelbar sind.[1]
Um die Situation langfristig zu verbessern, wird außerdem die Lobby-Arbeit zur Verbesserung der urheberrechtlichen Rahmenbedingungen auf deutscher Ebene fortgeführt und auf europäischer Ebene u.a. über Kooperationen im CLARIN-Kontext intensiviert (vgl. Kupietz/Bankhardt 2009, 2010; Witt 2013).
Weitere Schwerpunkte der Arbeiten des Projekts sind die Verbesserung der Dokumentation und Auszeichnung von DeReKo-Texten, um die Möglichkeiten zur Definition virtueller Korpora zu verbessern und die Verallgemeinerbarkeit DeReKo-basierter Befunde für seine Nutzer besser überprüfbar zu machen, die Erleichterung der Integrierbarkeit IDS-extern aufgebauter Korpora (Lüngen et al. 2011), die Erweiterung des Konzepts virtueller Korpora auf standort-übergreifende virtuelle Kollektionen (Broeder et al. 2007 sowie den Beitrag von Teubert/Belica in diesem Band), die Erweiterung der linguistischen Annotationen u.a. um Dependenzannotationen und die Ermöglichung nutzerdefinierter Textklassifikatoren (Klosa et al. 2012).

Korpustechnologie
Wie bereits eingangs angesprochen, wird zurzeit die Entwicklung von Korpustechnologie, die bisher zum größten Teil an der Arbeitsstelle ZDV angesiedelt war, in zunehmendem Maße wieder in den Programmbereich Korpuslinguistik integriert und damit enger an die korpuslinguistische Forschung angebunden.
Dafür gibt es mehrere Gründe: Zum einen hat sich die Anwendung anspruchsvoller korpuslinguistischer Methoden in den letzten Jahren auch in den anderen Abteilungen des IDS und der germanistischen Sprachwissenschaft insgesamt sehr stark etabliert. Dies hat dazu geführt, dass der so entstandene Bedarf nicht immer durch bereits in COSMAS II implementierte Funktionalitäten gedeckt werden konnte und häufig Lösungen gefunden werden mussten, bei denen es vor allem auf eine schnelle Umsetzbarkeit ankam und daher auf eine nachhaltige Wiederverwendbarkeit in anderen Kontexten nicht immer Rücksicht genommen werden konnte. Zum anderen hat das rasante Wachstum von

[1] Zu methodischen Aspekten und nach wie vor bestehenden Herausforderungen siehe Belica et al. (2007).

DeReKo und seiner Annotationsschichten auch im Programmbereich selbst dazu geführt, dass die Entwicklung von Prototypen für spezifische Forschungsaufgaben immer mehr Zeit in Anspruch nahm. Da außerdem COSMAS II bereits Anfang der neunziger Jahre konzipiert wurde[2] und daher der Aufwand für Weiterentwicklungen und die Integration neuer Funktionalitäten immer größer wurde, wurde 2010 zusammen mit dem Programmbereich Forschungsinfrastrukturen[3] im Rahmen des Senatsausschusswettbewerbsverfahrens der Leibniz-Gemeinschaft ein Drittmittel-Projekt zur Anschubfinanzierung der Neuentwicklung einer Korpusanalyseplattform der nächsten Generation *KorAP* beantragt (s. Bański et al. 2012), das nach einer Übergangsphase im Parallelbetrieb COSMAS II ablösen soll. Über ein solides Kernsystem, das für die nächsten 15-20 Jahre tragfähig ist, und über die COSMAS-II-Funktionalitäten hinaus soll KorAP vor allem eine beliebige Anzahl potenziell konkurrierender (auch nutzerdefinierter) Annotationen erlauben, die Definition virtueller Korpora besser unterstützen und die Nutzer in die Lage versetzen, die Reichhaltigkeit der Daten und Mächtigkeit des Systems möglichst vollständig ausschöpfen zu können. Letzteres soll auch über eine *Mobile Code Sandbox* erreicht werden, in der Nutzer selbstentwickelte Analysekomponenten innerhalb des KorAP-Systems ausführen können. Mittelfristig soll so, ähnlich wie ein App-Store für Smartphones, ein Open-Source-*KorApp*-Store mit von Nutzern entwickelten, wiederverwendbaren Analyse- und Visualisierungsapplikationen entstehen. Frei nach Jim Grays Postulat „put the computation near the data" (2003, S. 6), das umso zutreffender ist, wenn Daten nicht nur aufgrund ihres Umfangs kaum bewegt werden können, sondern sie darüber hinaus nicht bewegt werden dürfen (vgl. Kupietz/Frick 2013), ist dies auch gleichzeitig der eingeschlagene Weg, um trotz der unvermeidbaren urheber- und lizenzrechtlichen Auflagen, die die Weitergabe von Textdaten ausschließen, eine möglichst weitreichende und flexible Nutzung der Daten zu ermöglichen, ohne dabei die Interessen von Rechteinhabern zu verletzen. Zumindest der KorApp-Store ist allerdings im Moment noch Zukunftsmusik, da die produktreife Umsetzung der Kernfunktionalitäten noch einige Herausforderungen mit sich bringt und im Überlappungsbereich von fundierter empirischer Methodik (Unterscheidung von Beobachtungen und Interpretation, Nachvollziehbarkeit und Replizierbarkeit, Kontrollierbarkeit von Ergebnisreihenfolgen, ...) und Software-Entwicklung Lösungen, wie sie bei großen Web-Suchmaschinen und generell im Information-Retrieval eingesetzt werden, nicht 1:1 übernommen werden können und daher vielfach informatisches Neuland betreten werden musste und muss.
Um Nachteile und Risiken einer Insellösung zu vermeiden und die Aufgabe

[2] Vgl. die Beiträge von Teubert/Belica und Bodmer Mory in diesem Band.
[3] Vgl. den Beitrag von Schonefeld/Witt im Anschluss.

der Weiterentwicklung von KorAP auf mehrere Schultern zu verteilen, soll diese zukünftig auch über Kooperationen auf europäischer Ebene vorangetrieben und in einem weiteren Schritt durch eine Realisierung als Open-Source-Projekt für die Allgemeinheit geöffnet werden. Über die Bündelung von Ressourcen bezüglich der Entwicklung hinaus soll damit auch auf der Seite der Anwendung der Weg für eine stärkere Kanonisierung von Methodologien und Verfahren zur Korpusanalyse geebnet und damit die Voraussetzungen für die Weiterentwicklung der empirischen Sprachwissenschaft insgesamt verbessert werden.

Literatur

- **Aston, Guy/Burnard, Lou** (1998): The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh.
- **Bański, Piotr et al.** (2012): The New IDS Corpus Analysis Platform: Challenges and Prospects. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), ELRA. Istanbul, Turkey, May 2012. Internet: http://www.lrec-conf.org/proceedings/lrec2012/pdf/789_Paper.pdf (abgerufen am 28.6.2013).
- **Belica, Cyril** (2006): Modellierung semantischer Nähe: Kontrastierung von nahen Synonymen. Korpusanalytische Methode. Internet: http://corpora.ids-mannheim.de/ccdb/.
- **Belica, Cyril** (2011): Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In: Abel, Andrea/Zanin, Renata (Hg.): Korpora in Lehre und Forschung. Freie Universität Bozen-Bolzano, S. 155-178.
- **Belica, Cyril et al.** (2007): Web as Corpus: Kooperation mit der Universität Bologna. In: Sprachreport Sonderheft/März 2007. Auslandskooperationen des Instituts für Deutsche Sprache, S. 21-25.
- **Belica, Cyril et al.** (2010): Putting corpora into perspective. Rethinking synchronicity in corpus linguistics, Proceedings of the Corpus Linguistics Conference (CL2009), University of Liverpool, UK, 20-23 July 2009, Edited by Michaela Mahlberg, Victorina González-Díaz, Catherine Smith. Internet: http://ucrel.lancs.ac.uk/publications/cl2009/342_FullPaper.doc.
- **Belica, Cyril et al.** (2011): The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Konopka, Marek et al. (Hg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.4.-24.9.2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen, S. 451-469.
- **Broeder, Dan et al.** (2007): Citation of Electronic Resources: proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Dokument N366. Internet: http://www.ttt.org/tc37/ISO%20Conference%202007_files/PeterW_Citation_of_Electronic_Resources.pdf (abgerufen am 28.6.2013).
- **Elman, J. L.** (2011). Lexical knowledge without a mental lexicon? In: The Mental Lexicon 60, S. 1-33.
- **Gray, Jim** (2003): Distributed Computing Economics. Technical report, Microsoft Research. MSR-TR-2003-24.

- **Hopper, Paul** (1987): Emergent Grammar. In: Berkeley Linguistics Society 13, S. 139-157.
- **IDS** (2013): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2013-I (Release vom 19.3.2013). Mannheim. Internet: http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html.
- **Keibel, Holger/Kupietz, Marc** (2009): Approaching grammar: Towards an empirical linguistic research programme. In: Minegishi, Makoto/Kawaguchi, Yuji (Hg.): Working Papers in Corpus-based Linguistics and Language Education, No. 3. Tokyo, S. 61-76. Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/061-076.pdf.
- **Keibel, Holger et al.** (2011): Approaching grammar: Detecting, conceptualizing and generalizing paradigmatic variation. In: Konopka, Marek et al. (Hg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.4.-24.9.2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen, S. 329-355.
- **Klosa, Annette/Kupietz, Marc/Lüngen, Harald** (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica. (= Lexicographica 28). Berlin/New York, S. 71-97.
- **Kupietz, Marc** (1996): Modellierung der Sprachproduktion: Perspektiven neuerer kognitionswissenschaftlicher Ansätze. Magisterarbeit, Fakultät für Linguistik und Literaturwissenschaften, Universität Bielefeld. Internet: ftp://ftp.ids-mannheim.de/kt/mds-1996.pdf.
- **Kupietz, Marc** (2002): Computersimulation von Sprachproduktion: Konnektionistische Syntaxmodellierung. In: Müller, Horst M. (Hg.): Arbeitsbuch Linguistik. Paderborn, S. 443-460.
- **Kupietz, Marc/Keibel, Holger** (2009): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In: Konopka, Marek/Strecker, Bruno (Hg.): Deutsche Grammatik – Regeln, Normen, Sprachgebrauch. (= Jahrbuch des Instituts für Deutsche Sprache 2008). Berlin/New York, S. 33-50.
- **Kupietz, Marc/Bankhardt, Christina** (Hg.) (2009): D-SPIN Report R7.1 – Legal Aspects in the Provision of Language Resources: The German Context. Internet: http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.1.pdf (abgerufen am 28.6.2013).
- **Kupietz, Marc/Bankhardt, Christina** (Hg.) (2010): D-SPIN Report R7.3 – Initial Localisation of CLARIN Best Practices and Business Models. Internet: http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.3.pdf (abgerufen am 28.6.2013).
- **Kupietz, Marc et al.** (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010), S. 1848-1854. Internet: http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (abgerufen am 28.6.2013).
- **Kupietz, Marc/Schonefeld, Oliver/Witt, Andreas** (2010): The German Reference Corpus: New developments building on almost 50 years of experience. In: Arranz, Eerten (Hg.): Language Resources: From Storyboard to Sustainability and LR Lifecycle Management, Workshop held at the seventh conference on International Language Resources and Evaluation (LREC 2010). Internet: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf (abgerufen am 28.6.2013).

- **Kupietz, Marc/Keibel, Holger** (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji (Hg.): Working Papers in Corpus-based Linguistics and Language Education 3. Tokyo, S. 53-59. Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf (abgerufen am 28.6.2013).
- **Kupietz, Marc/Frick, Elena** (2013): Korpusanalyseplattform der nächsten Generation. In: Kratochvilová, Iva/Wolf, Norbert Richard (Hg.): Grundlagen einer sprachwissenschaftlichen Quellenkunde. (= Studien zur Deutschen Sprache 66). Tübingen, S. 27-36.
- **Kupietz, Marc** (i. Vorb.): Constructing a Corpus. In: Durkin, Philip (Hg.): The Oxford Handbook of Lexicography. Oxford.
- **Langacker, Ronald W.** (2000): A dynamic usage-based model. In: Barlow, Michael/Kemmer, Suzanne (Hg.): Usage-based Models of Language. Stanford, S. 1-63.
- **Lüngen, Harald et al.** (2011): Strategien zur Weiterentwicklung von DeReKo 2012 bis 2020. (Tech. Rep. KT-2011-01). Institut für Deutsche Sprache.
- **Marková, Věra** (2012): Synonyme unter dem Mikroskop: Eine korpuslinguistische Studie. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 2). Tübingen.
- **Smolensky, Paul** (1988): On the proper treatment of connectionism. Behavioral and Brain Sciences 11, S. 1-74.
- **Tomasello, Michael** (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA.
- **Witt, Andreas** (2013): CLARIN-D AP6.2 Report: Legal and ethical issues. May 2012– April 2013. Internet: http://de.clarin.eu/images/jahrsberichte/CLARIND-R6.2.pdf (abgerufen am 28.6.2013).