Erschienen in: Zeitschrift für germanistische Linguistik, Bd. 45, Heft 1, S. 161-170
ZGL 2017; 45(1): 161–170

Linguistik im Internet

Harald Lüngen*

DeReKo – Das Deutsche Referenzkorpus

DeReKo – the German Reference Corpus.

Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim

Text Corpora of Contemporary German at the Institute for the German Language (IDS) in Mannheim.

DOI 10.1515/zgl-2017-0008

1 Einleitung

Das am Institut für Deutsche Sprache in Mannheim beheimatete Deutsche Referenzkorpus (DeReKo) ist vermutlich das weltweit größte sprachwissenschaftlich motivierte Archiv deutschsprachiger Texte. Seit 1964 wird es mit dem Anspruch, den deutschen Schriftsprachgebrauch zeitbegleitend möglichst repräsentativ abzubilden, kontinuierlich ausgebaut und enthält heute (Stand 31.12.2016) über 30 Milliarden Tokens (Abbildung 1). Es dient primär dazu, die Verwendung der deutschen Sprache und ihre Entwicklung zu dokumentieren und damit allen interessierten Sprachwissenschaftlern weltweit eine empirische Grundlage für Forschungen zur deutschen Gegenwartssprache zur Verfügung zu stellen, insbesondere für quantitative Untersuchungen, die sehr große Korpora erfordern.

Die Merkmale Gegenwartssprache und Schriftsprache charakterisieren DeReko als komplementär zu zwei weiteren großen langfristigen Korpusinitiativen zur

***Kontaktperson: Dr. Harald Lüngen**: IDS – Institut für Deutsche Sprache, R 5, 6-13, D-68161 Mannheim, E-Mail: luengen@ids-mannheim.de

deutschen Sprache, dem Archiv für gesprochenes Deutsch (AGD),[1] das ebenfalls vom IDS bereitgestellt wird, und dem Deutschen Textarchiv (DTA),[2] das von der Berlin-Brandenburgischen Akademie der Wissenschaften aufgebaut wird und deutschsprachige historische Texte des 17.–19. Jahrhunderts enthält.

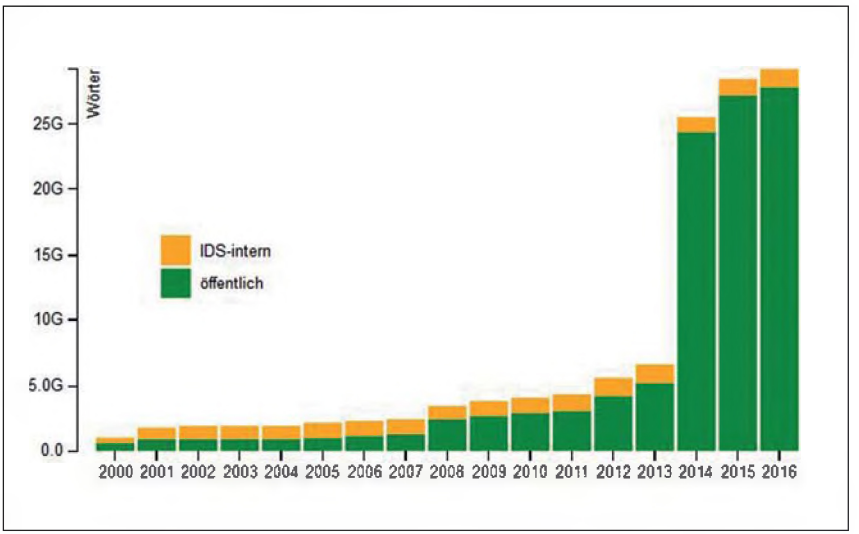

Abb. 1: DeReKo-Wachstum

2 Zusammensetzung und Urstichprobendesign

Im Bestand von DeReKo befinden sich zahlreiche verschiedene Teilkorpora, die zumeist definiert sind durch ein Genre (z. B. Literaturkorpora, Handbuchkorpus), eine bestimmte Quelle (z. B. Zeitungskorpora wie das *Mannheimer Morgen*-Korpus oder autorenbezogene Korpora wie das Goethe-Korpus) oder durch ein bestimmtes Projekt, in dem sie entstanden sind, z. B. das Mannheimer Korpus I und II, das PolMine-Plenardebattenkorpus oder das Dortmunder Chatkorpus (s. u.). Den Hauptanteil an DeReKo bilden nach wie vor Zeitungskorpora, zusätzlich sind viele weitere schriftsprachliche Genres wie Literatur, Biographien, Fachtexte, wissenschaftliche Texte, Agenturmeldungen, Interviews und Protokolle vertreten; aus-

1 http://agd.ids-mannheim.de/
2 http://www.deutsches-textarchiv.de/

drücklich nicht akquiriert werden lediglich Lyrik und nicht im Original auf Deutsch entstandene Texte. Alle Teilkorpora sind im Rahmen von DeReKo nach dem einheitlichen Kodierungsschema I5 erfasst und aufbereitet (s. Abschnitt 4) und mit dem Korpusrecherchesystem (derzeit COSMAS II, s. Abschnitt 5) einzeln oder in Kombination abfragbar.

Ein wichtiges Merkmal von DeReKo ist sein Design als Urstichprobe des deutschen Schriftsprachgebrauchs, also als eine flexible, übergeordnete Stichprobe, aus der in einer Benutzungsphase weitere, spezialisierte Sub-Stichproben, sogenannte virtuelle Korpora, gezogen werden können (vgl. Kupietz et al. 2010). Anders als frühere National- oder Referenzkorpora wie das wegweisende *British National Corpus* (BNC) wird für DeReKo keine ausgewogene oder gar „repräsentative" Zusammensetzung bezüglich Genres oder anderer Dimensionen angestrebt; denn was als ausgewogen zu gelten hat, hängt letztlich von einer Forschungsfrage und einer zu untersuchenden Sprachdomäne ab. Welche sprachlichen Dimensionen und Strata relevant sind, welche Zeiträume durch ein Korpus abgedeckt werden sollen, ob ein Anteil von 20 % von Texten aus Österreich relevant ist und weitere derartige Kriterien – das sollte jede Forscherin, jedes Projekt selbst definieren dürfen und die Möglichkeit nutzen, aus der Urstichprobe virtuelle Korpora zu ziehen, die möglichst repräsentativ bezüglich ihrer jeweiligen Fragestellung sind. Ein aktuelles Beispiel eines auf ein spezifisches Forschungsvorhaben zugeschnittenen virtuellen Korpus ist das Paronymkorpus des IDS-Projekts *Paronymwörterbuch* (vgl. Storjohann 2016), welches u. a. anhand der Entstehungszeiten, der regionalen Abdeckung und der geschätzten Reichweite von Tageszeitungen mit Hilfe des Korpusrecherchesystems COSMAS II definiert wurde und dort auch öffentlich zur Verfügung steht, d. h. auch von projektexternen Nutzern verwendet werden kann.[3]

Das Urstichproben-Design gewährleistet eine optimale Nutzbarkeit der Daten für die größtmögliche Anzahl potenziell relevanter Fragestellungen, wobei die relative Besetzung einzelner Strata im Gesamtbestand irrelevant ist – entscheidend sind die absoluten Größen. Beispielsweise machen die Belletristik-Korpora in DeReko mit rund 17,6 Millionen Tokens derzeit nur einen relativ kleinen Anteil am Gesamtbestand aus, dennoch kann man daraus leicht ein für eine Forschungsfrage angemessenes virtuelles Korpus mit einem Belletristik-Anteil von z. B. 16 % (das entspricht etwa dem Anteil der *imaginative written texts* im BNC, vgl. Burnard 2007) definieren.

Das Urstichproben-Design ermöglicht auch den kontinuierlichen Ausbau von DeReKo, ohne dass z. B. angebotene Textdaten zurückgewiesen werden müssten,

3 Vgl. http://www.ids-mannheim.de/lexik/paronymwoerterbuch/dasparonymkorpus.html

nur weil die vorgesehene Quote eines Genres bereits erreicht ist. Eine Maximierung der Korpusgröße ist aus korpuslinguistischer Perspektive grundsätzlich wünschenswert, denn je größer das Korpus, desto verlässlichere Aussagen über mehr verschiedene und seltenere Phänomene können anhand des Korpus getroffen werden. Beim Ausbau von DeReKo wird außerdem auf die Maximierung der Stratifizierung (Diversität) geachtet. Selbstverständlich spielen auch langfristige Akquisitionsstrategien und Prognosen eine Rolle, wie aber auch Angebot (von Rechte-Inhabern) und Nachfrage (von IDS-internen und -externen Projekten und Nutzern) sowie Kosten/Nutzen-Abwägungen bezüglich des finanziellen und technischen Aufwands der Akquisition und Kuratierung von Textbeständen und -archiven. Unter diesen Vorzeichen wurden in den letzten fünf Jahren schwerpunktmäßig in den folgenden Bereichen neue Textdaten akquiriert:

Presse: Die Archive großer überregionaler Tageszeitungen (*Süddeutsche Zeitung, Die ZEIT, Der SPIEGEL*) wurden für DeReKo erschlossen, und durch eine Kooperation mit einem kommerziellen News-Datenbank-Provider wurden zahlreiche weitere regionale und überregionale Pressequellen akquiriert, und zwar in der Größenordnung von insgesamt rund 20 Milliarden Tokens. Dadurch ist der deutsche Sprachraum mit Pressequellen mittlerweile recht gleichmäßig abgedeckt.

Konzeptionelle Schriftlichkeit bei medialer Mündlichkeit: Das *PolMine*-Korpus deutscher Plenardebattenprotokolle von Andreas Blätte (Uni Duisburg-Essen)[4] und das *German Political Speeches*-Korpus von Adrien Barbaresi (Barbaresi 2012) wurden integriert mit insgesamt 316 Millionen Tokens.

Belletristik: Als Spenden von Verlagen wurden seit 2011 neu eingeworbene vollständige Buchtexte im Umfang von circa 8,5 Millionen Tokens in DeReKo integriert. Die vergleichsweise geringe Tokenanzahl dieses Genres erklärt sich dadurch, dass die Akquisition von Buchpublikationen am aufwändigsten ist, sowohl was Verhandlungen mit den Rechte-Inhabern als auch die technischen Anforderungen an die Korpusaufbereitung betrifft.

Konzeptionelle Mündlichkeit bei medialer Schriftlichkeit: Internetbasierte Kommunikation (IBK): In einem CLARIN-D-Kurationsprojekt wurde das Dortmunder Chatkorpus (Beißwenger 2013, ca. eine Million Tokens) für Korpusinfrastrukturen im Forschungsverbund CLARIN-D (u. a. DeReKo) kuratiert (vgl. Lüngen et al. 2016). In Eigenregie des IDS wurden außerdem folgende IBK-Korpora aufgebaut:

a. Wikipedia-Korpora. In einem Turnus von zwei Jahren wird die jeweils aktuelle deutschsprachige Wikipedia als Korpus in DeReKo zur Verfügung gestellt. Die Konvertierung von 2015 umfasst sämtliche Artikel (die allerdings nicht unter IBK fallen), alle Artikel-Diskussionen und erstmals auch alle

4 http://polmine.sowi.uni-due.de/

Nutzer-Diskussionen, mit einem Gesamtumfang von 1.378 Milliarden Tokens (vgl. Margaretha/Lüngen 2014).

b. Ein Usenet-News-Korpus mit den Texten aller deutschen Newsgruppen seit 2013 wurde für DeReKo aufgebaut (Schröck/Lüngen 2015). Dieses Korpus wird in Zukunft sowohl aktualisierend als auch mit weiter zurückliegenden Jahrgängen erweitert.

Derzeit erscheint alle sechs Monate eine neue, aktualisierte Ausgabe von DeReKo (DeReKo-Release).

3 Rechtliche Situation

Das IDS ist nicht Eigentümer der Texte in DeReKo, sondern hat mit über 200 Rechte-Inhabern, d. h. Inhabern von Urheber-, Verwertungs- oder Datenbankschutzrechten, Lizenzvereinbarungen über eingeschränkte, übertragbare Nutzungsrechte abgeschlossen. Zu den gängigen Einschränkungen gehört, dass die Korpora in DeReKo ausschließlich für nicht-kommerzielle, wissenschaftliche Zwecke genutzt werden dürfen, dass Nutzer sich unter Angabe ihrer Affiliation beim IDS registrieren müssen und dass die Nutzung nur über ein Korpusrechercheystem erfolgen darf, das als Treffer für linguistische Suchanfragen Textstellen maximal in Zitatgröße anzeigt und ein Herunterladen von Volltexten technisch ausschließt. Bei einigen Korpora (derzeit rund 4 % des Bestandes) kommt hinzu, dass sie nur vor Ort in den Räumlichkeiten des IDS, also nicht weltweit über das Web, genutzt werden dürfen. Die genannten Nutzungseinschränkungen kommen dadurch zustande, dass Buch- und Zeitungsverlage eine gewisse Garantie dafür haben möchten, dass die Korpusangebote des IDS nicht in Konkurrenz zu ihren eigenen Verwertungsinteressen stehen.

Grundsätzlich überträgt das IDS alle übertragbaren Nutzungsrechte so weit wie möglich. Folglich werden einige Korpora, die unter freien Lizenzen stehen, auch zum Herunterladen angeboten.[5]

5 http://www.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html

4 Kodierung, Metadaten und linguistische Annotationen

Alle Korpora in DeReKo werden nach dem sogenannten *IDS-Textmodell* kodiert, in dem die grundlegende Textstruktur (Markierung von Kapiteleinteilungen, Überschriften, Absätzen, Listenstrukturen, Postings, hervorgehobenen und fremdsprachigen Bereichen und vielem mehr) sowie Metadaten zum Text ausgezeichnet werden. Die Metadaten umfassen zunächst vollständige bibliografische Angaben inklusive der nachweisbaren Entstehungszeit bzw. des Datums der Erstveröffentlichung (damit die Nutzer gefundene Textstellen wissenschaftlich zitieren können), aber auch textlinguistische Kategorisierungen, die entweder anhand von in den Quellen vorhandenen Angaben ausgefüllt werden (Schlagwort, Ressort, Textart, Textsorte) oder auch durch automatische Textklassifikation bestimmt werden (Textthema) (vgl. Klosa et al. 2012).

Seit 2013 ist das IDS Textmodell als I5, eine Kustomisierung des P5-Standards der *Text Encoding Initiative* (TEI), realisiert. Demnach enthält I5 derzeit 148 Original-Elemente aus TEI P5 sowie 48 zusätzlich definierte Elemente, von denen die meisten aus (X)CES (Ide et al. 2000) stammen, doch einige auch IDS-spezifisch sind. I5 ist formal und inhaltlich in dem TEI-Formalismus ODD definiert und kann somit auch von externen Projekten verwendet werden (Lüngen/Sperberg-McQueen 2012).[6] I5 wird von Zeit zu Zeit erweitert, insbesondere wenn neue Textsorten und -genres in DeReKo aufgenommen werden. Beispielsweise wurden anlässlich der Integration von IBK-Korpora in DeReKo einige Elemente adaptiert, die in den Special Interest Groups (SIGs) „Computer-mediated Communication (CMC)" und „Correspondence" der TEI entwickelt wurden. Das IDS beteiligt sich dabei insbesondere an der Arbeit der SIG „CMC", die sich unter anderem zum Ziel gesetzt hat, neue Elemente und Attribute, die für die Auszeichnung von IBK-Korpora benötigt werden, auch im offiziellen TEI-Standard zu etablieren.[7] Die Kodierung der DeReKo-Korpora in TEI, dem internationalen Standard zur Kodierung von Texten in den Geisteswissenschaften, ermöglicht Interoperabilität mit anderen Korpora, die den gleichen Standard verwenden, insbesondere im Rahmen des europäischen Forschungsverbundes CLARIN (CLARIN-D 2012).

Neben den I5-kodierten Metadaten werden deskriptiv-statistische Metadaten wie die Anzahl der Sätze, Wörter, Stoppwörter, Absätze, Umbrüche etc. in einer Metadatenbank festgehalten. Außerdem werden Arten von Textduplikaten ermit-

6 http://www.ids-mannheim.de/kl/projekte/korpora/textmodell.html

7 Vgl. Webseite der TEI CMC SIG: http://www.tei-c.org/Activities/SIG/CMC/

telt und als Relationskategorien zwischen Texten kodiert. Alle Arten von Metadaten können zur Definition virtueller Korpora herangezogen werden.

Am IDS treffen pro Tag durchschnittlich etwa 20.000 neue Texte ein, und zwar in allen denkbaren Formaten (z. B. Word, PDF, RTF, XML, HTML, InDesign). Für viele Formate existieren Aufbereitungs-Toolchains, die die benötigten Daten und Metadaten aus den Dateien extrahieren und in das Zielformat I5 konvertieren. Nicht selten muss eine Toolchain auch neu angepasst werden, da jede Textquelle eigene formattechnische Besonderheiten aufweist.

Zusätzlich zur Kodierung der Basisstruktur und der Metadaten in I5 werden DeReKo-Texte auf Tokenebene und auf Satzebene segmentiert sowie auf weiteren linguistischen Ebenen annotiert, um den Nutzern zu ermöglichen, in ihren Korpusanfragen auch Bedingungen über linguistische Kategorien wie Wortarten zu formulieren. Da Annotationstools immer einen gewissen Anteil von Fehlern produzieren, sind mit ihrer Anwendung auch Fallstricke verbunden (Belica et al. 2011). Um diese abzumildern, wird für DeReKo die Maxime verfolgt, möglichst viele Annotationswerkzeuge zu verwenden, die unterschiedliche Funktionsweisen haben, um möglichst viele Interpretationen zu erhalten. Das Ziel dabei ist also eine Minimierung von Typ-II-Fehlern in Anfrageergebnissen bzw. eine Maximierung des Recalls in der Treffermenge. Entsprechend wird DeReKo derzeit mit vier verschiedenen Werkzeugen annotiert:

1. TreeTagger (Schmid 1994): Lemmatisierung, Wortarten-Tagging mit dem TagSet: STTS (Schiller et al. 1999)
2. Connexor Machinese Phrase Tagger:[8] Wortarten-Tagging, Phrasen-Tagging
3. CoreNLP (Manning et al. 2014): Lemmatisierung, Wortarten-Tagging, Syntaxanalyse, Named Entity-Erkennung
4. OpenNLP:[9] Lemmatisierung, Wortarten-Tagging

Das aktuelle Korpusrecherchesystem COSMAS II kann allerdings nur einen kleinen Teil der DeReKo-Annotationen indizieren. Die TreeTagger und Connexor-Annotationen stehen in COSMAS II zur Verfügung, sind aber nur einzeln (also nicht als multiple Annotationen) abfragbar und werden auch nicht für jedes DeReKo-Release neu bereitgestellt. Die CoreNLP und OpenNLP-Annotation (sowie noch weitere Annotationen) werden erst in dem zukünftigen neuen Recherchesystem KorAP (Diewald et al. 2016) zur Verfügung stehen.

8 http://www.connexor.com/nlplib/
9 https://opennlp.apache.org/

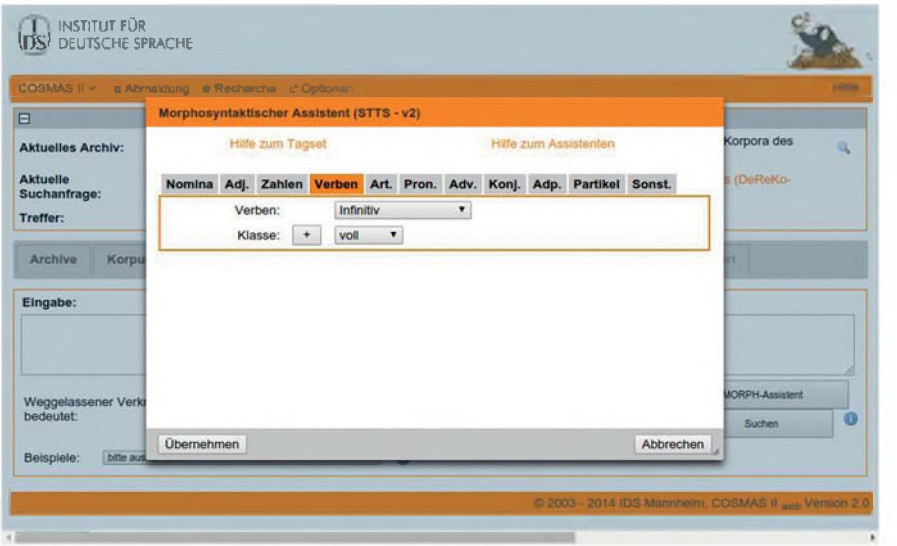

Abb. 2: Morph-Assistent in COSMAS II

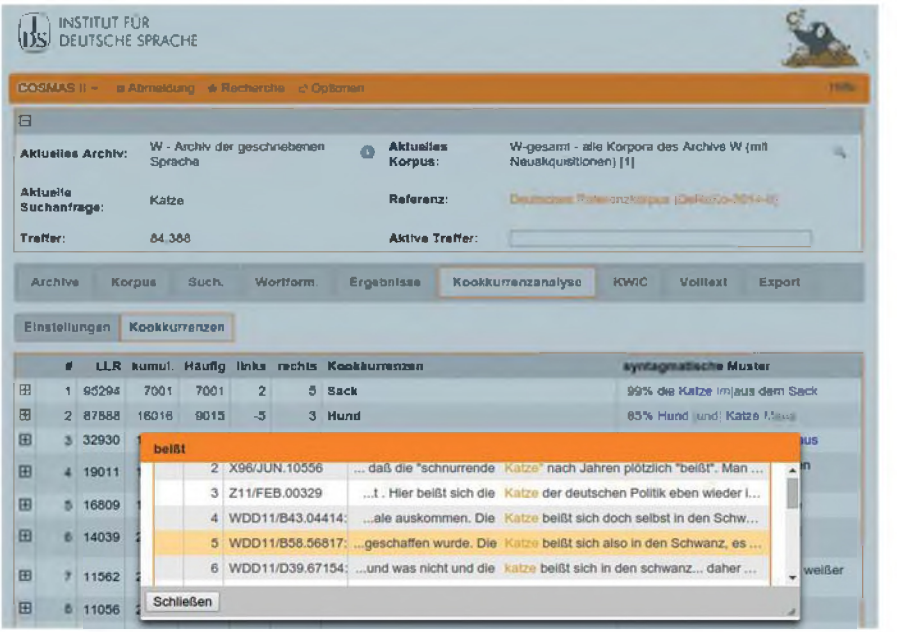

Abb. 3: Kokkurrenzanalyse in COSMAS II

5 Zugriff

Die wichtigste und bekannteste Zugriffsmöglichkeit auf DeReKo erfolgt über das web-basierte Korpusrecherchesystem COSMAS II (Corpus Search, Management and Analysis System) des IDS (Bodmer 2014). Für COSMAS sind derzeit weltweit 39.000 Benutzer registriert, wobei nicht alle davon fortlaufend aktiv sind. COSMAS wurde bereits 1994 konzipiert und bietet viele korpuslinguistische Funktionalitäten, u. a. eine linguistische Suchanfragesprache und eine Methode zur Definition virtueller Korpora. Ein Morph-Assistent erleichtert die Formulierung von Anfragen über Wortarten-Annotationen (Abbildung 2). Als typische Ergebnispräsentation gilt die KWIC-Ansicht von Treffermengen. Über den Treffermengen kann zusätzlich eine Kookkurrenzanalyse durchgeführt werden (Abbildung 3).

Durch das Anwachsen von DeReKo stößt COSMAS II seit einiger Zeit an seine Grenzen. DeReko ist beispielsweise in COSMAS II mittlerweile auf vier verschiedene Archive verteilt, wobei eine Rechercheanfrage immer nur auf maximal einem Archiv ausgeführt werden kann. Dass die DeReKo-Annotationen in COSMAS nicht vollständig zur Verfügung gestellt werden können, wurde oben bereits erwähnt. Als Nachfolger von COSMAS II wird daher am IDS seit 2011 das neue Recherchesystem KorAP entwickelt (*Korpusanalyseplattform der nächsten Generation,* vgl. Diewald et al. 2016). Korpora in KorAP können über beliebig viele Rechner verteilt sein und daher beliebig groß werden. Neben der COSMAS-II-Suchanfragesprache können auch die in der Korpuslinguistik außerdem gebräuchlichen Suchanfragesprachen Poliqarp und ANNIS-QL verwendet werden. KorAP bietet außerdem neue Möglichkeiten der Suche über multiplen linguistischen Annotationen und der Visualisierung von Annotationen. Ab Sommer 2017 wird DeReKo seinen Nutzern auch in KorAP zur Verfügung stehen, wobei es für eine Übergangszeit von mindestens zwei Jahren einen Parallelbetrieb von COSMAS II und KorAP geben wird.

DeReKo: http://www.ids-mannheim.de/dereko

COSMAS II: http://www.ids-mannheim.de/cosmas2/

Literatur

Barbaresi, Adrien (2012): *German Political Speeches, Corpus and Visualization.* http://purl.org/corpus/german-speeches.

Beißwenger, Michael (2013): „Das Dortmunder Chat-Korpus." In: *Zeitschrift für germanistische Linguistik,* 41(1), S. 161–164.

Belica, Cyril; Kupietz, Marc; Witt, Andreas; Lüngen, Harald (2011): „The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls." In: Konopka, Marek; Kubczak,

Jacqueline; Mair, Christian; Šticha, František; Waßner, Ulrich Hermann (Hrsg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.–24.9.2009.

Bodmer Mory, Franck (2014): „Mit COSMAS II »in den Weiten der IDS-Korpora unterwegs«". In: Institut für Deutsche Sprache (Hrsg.): *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Redaktion: Melanie Steinle, Franz Josef Berens. Mannheim: Institut für Deutsche Sprache, S. 376–385.

Burnard, Lou (2007): *Reference Guide for the British National Corpus (XML Edition)*. January 2007. http://www.natcorp.ox.ac.uk/docs/URG.xml

CLARIN-D AP-5 (2012): *CLARIN-D User Guide*. CLARIN. http://de.clarin.eu/de/sprachressourcen/benutzerhandbuch.html

Diewald, Nils; Hanl, Michael; Margaretha, Eliza; Bingel, Joachim; Kupietz, Marc; Banski, Piotr; Witt, Andreas (2016): „KorAP Architecture – Diving in the Deep Sea of Corpus Data." In: *Proceedings of LREC 2016*. http://www.lrec-conf.org/proceedings/lrec2016/index.html

Ide, Nancy; Bonhomme, Patrice; Romary, Laurent Romary (2000). „XCES: An XML-based Standard for Linguistic Corpora." In: *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*. Athens, Greece, S. 825–830.

Klosa, Annette; Kupietz, Marc; Lüngen, Harald (2012): „Zum Nutzen von Korpusauszeichnungen für die Lexikographie." In: Lexicographica 28, S. 71–97. Berlin/Boston: de Gruyter

Kupietz, Marc; Belica, Cyril; Keibel, Holger; Witt, Andreas (2010): „The German Reference Corpus DeReKo." In: Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odjik, Jan; Piperidis, Stelios; Rosner, Mike; Tapias, Daniel (Hrsg.): *Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA). S. 1848–1854.

Lüngen, Harald; Sperberg-McQueen, Michael (2012): „A TEI P5 Document Grammar for the IDS Text Model." In: Bański, Piotr/Modignani Picozzi, Eleonora Litta/Witt, Andreas (Hrsg.): *TEI and Linguistics. Journal of the Text Encoding Initiative* 3/2012. https://jtei.revues.org/508

Lüngen, Harald; Beißwenger, Michael; Ehrhardt, Eric; Herold, Axel; Storrer, Angelika (2016): „Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN." In: *Proceedings of KONVENS 2016*, Bochum. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf

Manning, Christopher D.; Surdeanu, Mihai; Bauer, John; Finkel, Jenny; Bethard, Steven J.; McClosky, David (2014): „The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, S. 55–60. http://acl2014.org/acl2014/P14-5/pdf/P14-5010.pdf

Margaretha, Eliza; Lüngen, Harald (2014): „Building Linguistic Corpora from Wikipedia Articles and Discussions." In: *Journal of Language Technology and Computational Linguistics* (JLCL) 29 (2), S. 59–82.

Schröck, Jasmin; Lüngen, Harald (2015): „Building and Annotating a Corpus of German-Language Newsgroups." In: *Proceedings of the Workshop NLP4CMC 2015*, Essen. https://sites.google.com/site/nlp4cmc2015/proceedings

Schmid, H. (1994): „Probabilistic part-of-speech tagging using decision trees." In: *International Conference on New Methods in Language Processing*, S. 44–49. Manchester, UK.

Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universitäten Stuttgart und Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

Storjohann, Petra (2016): „Vom Interesse am Gebrauch von Paronymen zur Notwendigkeit eines dynamischen Wörterbuchs." In: *Sprachreport* 4/2016, S. 32–43.