Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining (KobRA)

Durchführende Forschungseinrichtungen

TU Dortmund (Germanistik)	**TU Dortmund (Informatik)**
Prof. Dr. **Angelika Storrer** (Koordination) Institut für deutsche Sprache und Literatur, Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik	Prof. Dr. **Katharina Morik** Fakultät Informatik, Lehrstuhl für Künstliche Intelligenz

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)	**Eberhard-Karls-Universität Tübingen, Seminar für Sprachwissenschaft (SfS)**	**Institut für Deutsche Sprache Mannheim (IDS)**
Dr. **Alexander Geyken** Zentrum Sprache / Projekt: Digitales Wörterbuch der deutschen Sprache (DWDS)	Prof. Dr. **Erhard Hinrichs** Seminar für Sprachwissenschaft (Computerlinguistik)	Dr. **Marc Kupietz** Dr. **Andreas Witt** Programmbereiche Korpuslinguistik und Forschungsinfrastruktur

Disziplinäre Verortung

Beteiligte Disziplinen: Germanistische Sprachwissenschaft / Informatik / Computerlinguistik

Methoden und Anwendungsbereich: Data-Mining-Methoden zur Verbesserung der korpus-basierten Recherche und Analyse in großen strukturierten Textkorpora (mit Schwerpunkt auf Deutsch als Untersuchungssprache)

Wissenschaftliche Anwendungsfelder („*use cases*“):

Korpus-basierte Linguistik hat sich in den letzten Jahren zu einem wichtigen Gebiet der Sprachforschung entwickelt. In Infrastrukturprojekten wie CLARIN werden umfangreiche, strukturierte Sprachressourcen (Textkorpora, Baumbanken, lexikalische Wortnetze) bereitgestellt, die hervorragende Möglichkeiten für die empirische Untersuchung sprachlicher Phänomene eröffnen. Das Projekt setzt auf diesen Infrastrukturen auf und nutzt innovative Data-Mining-Verfahren (insbesondere Lernverfahren), die über die reine Suche hinausgehen, indem sie die Suchergebnisse filtern, sortieren oder strukturieren und ggf. die weitere Aufbereitung der Daten für eine konkrete Fragestellung erleichtern. Um die Nutzer bei der Exploration verschiedener strukturierter Datenbestände zu unterstützen, sollen auch innovative Formen der Visualisierung für typische sprachbezogene Forschungsfragen erprobt werden. Durch diese Verfahren sollen korpus-basiert arbeitende Linguisten und Lexikographen bei typischen Routineaufgaben unterstützt werden, sodass sie diese künftig schneller und mit besserem Ergebnis durchführen können.

Die zu entwickelnden Verfahren werden an Fallstudien aus drei linguistischen Anwendungsfeldern erprobt und evaluiert:

Varietätenlinguistik/Internetbasierte Kommunikation: Untersuchung von Sprachmerkmalen in Genres internetbasierter Kommunikation, auch im Vergleich zu standardkonformer redigierter Schriftlichkeit in anderen Textsortenbereichen (Belletristik, Zeitung, Wissenschaft, Gebrauchstexte). Studien zur sprachlichen Variation zwischen verschiedenen Genres der internetbasierten Kommunikation; Studien zum Einfluss diatopischer Varietäten und sprachvergleichende Untersuchungen.

Lexikographie: Unterstützung des lexikographischen Arbeitsprozesses, z.B.: Suche nach interessanten, ungewöhnlichen Belegen (Metapher, Metonymie); Frequenzdaten zu disambiguierten lexikalischen Einheiten; Rekonstruktion und Visualisierung von Bedeutungswandel (z.B. *billig*, *toll*, *zeitnah*) und von Prozessen der Ausdifferenzierung von Teilbedeutungen über Zeiträume und Textsortenbereiche hinweg (z.B. *Ampel* als Hängelampe, als Lichtzeichenanlage, als politische Koalition etc.).

Diachronische Sprachforschung: Entwicklung von Wortschatz, Syntax, Morphologie in einem bestimmten Untersuchungszeitraum. Z.B. Entwicklung und Ausdifferenzierung des Systems deutscher Stützverbgefüge (*zur Anwendung bringen*, *zur Anwendung kommen* und *Anwendung finden*). Studien zu Prozessen des lexikalischen Wandels (Metapher, Metonymie, Grammatikalisierung etc.). Einfluss von Kontaktsprachen und diatopischen Varietäten.

Eine wichtige Zielgruppe für die Projektergebnisse sind Nachwuchswissenschaftler und fortgeschrittene Studierende: Als wichtige Multiplikatoren bei der Verbreitung korpus-basierter Zugänge zur Sprachforschung werden sie in die linguistischen Fallstudien und in die Erprobung der Verfahren mit eingebunden. In Verbindung mit Konzepten des „forschenden Lernens“ sollen die Projektergebnisse auch der Sprachvermittlung in der Schule zugutekommen. Darüber hinaus können auch andere Fachbereiche profitieren, in denen das Aufspüren interessanter und ungewöhnlicher Sprachverwendungen eine wichtige Rolle spielt (z.B. die Literaturwissenschaft).

Im Rahmen des Projekts wird erprobt, welche Routineaufgaben mithilfe welcher Data-Mining-Verfahren beschleunigt und/oder im Ergebnis verbessert werden können. Dabei kommen verschiedene Lernverfahren in Verbindung mit strukturierten Daten und Annotationen vielfältiger Art zum Einsatz. Auf diese Weise lassen sich Einsichten im Hinblick auf die Frage gewinnen, welche Merkmale in welchen Repräsentationen für welche Lernaufgaben am besten geeignet sind. Diese Frage ist für Informatik, Linguistik und Sprachtechnologie gleichermaßen interessant.

Methode

Im Projekt arbeiten Partner aus Informatik, Linguistik und Sprachtechnologie zusammen: Die Data-Mining-Methoden kommen aus der Informatik und beziehen sich auf korpus-basierte Forschungen der Linguistik. Als Wissensbasis dienen strukturierte Sprachressourcen der Sprachtechnologie-Partner (BBAW, IDS, SfS Tübingen), die im Rahmen von CLARIN-D Infrastrukturen für Sprachressourcen bereitstellen. Die Data-Mining-Verfahren des Projektes setzen auf diesen Infrastrukturen auf. Dabei ergibt sich einerseits eine Schnittstelle zu den linguistischen Anwendern und andererseits eine interne Schnittstelle zwischen der Data-

Mining-Komponente und der Infrastruktur. Das Schaubild in Abbildung 1 soll diese Verzahnung verdeutlichen.

Die Ergebnisse der Suchanfragen aus den Korpusinfrastrukturen werden als Datensätze behandelt, aus denen maschinell gelernt werden soll. Dabei werden folgende, für viele linguistische Untersuchungen relevante Verfahren an konkreten Fallstudien erprobt:

– Die Klassifikation der Ergebnisliste nach verschiedenen Bedeutungen (Disambiguierung);
– das Clustering der Ergebnisliste, sodass eine übersichtliche Struktur auch visuell dargestellt werden kann;
– das Erkennen von „ungewöhnlichen“ Belegen (Ausreißern);
– die aufgabenbezogene linguistische Annotation.

Die im Projekt entwickelten Verfahren werden im dritten Projektjahr in die Korpusinfrastrukturen der Sprachtechnologie-Partner integriert. Durch diese Integration ist der nachhaltige Nutzwert der Projektergebnisse über die Projektlaufzeit hinaus gesichert.

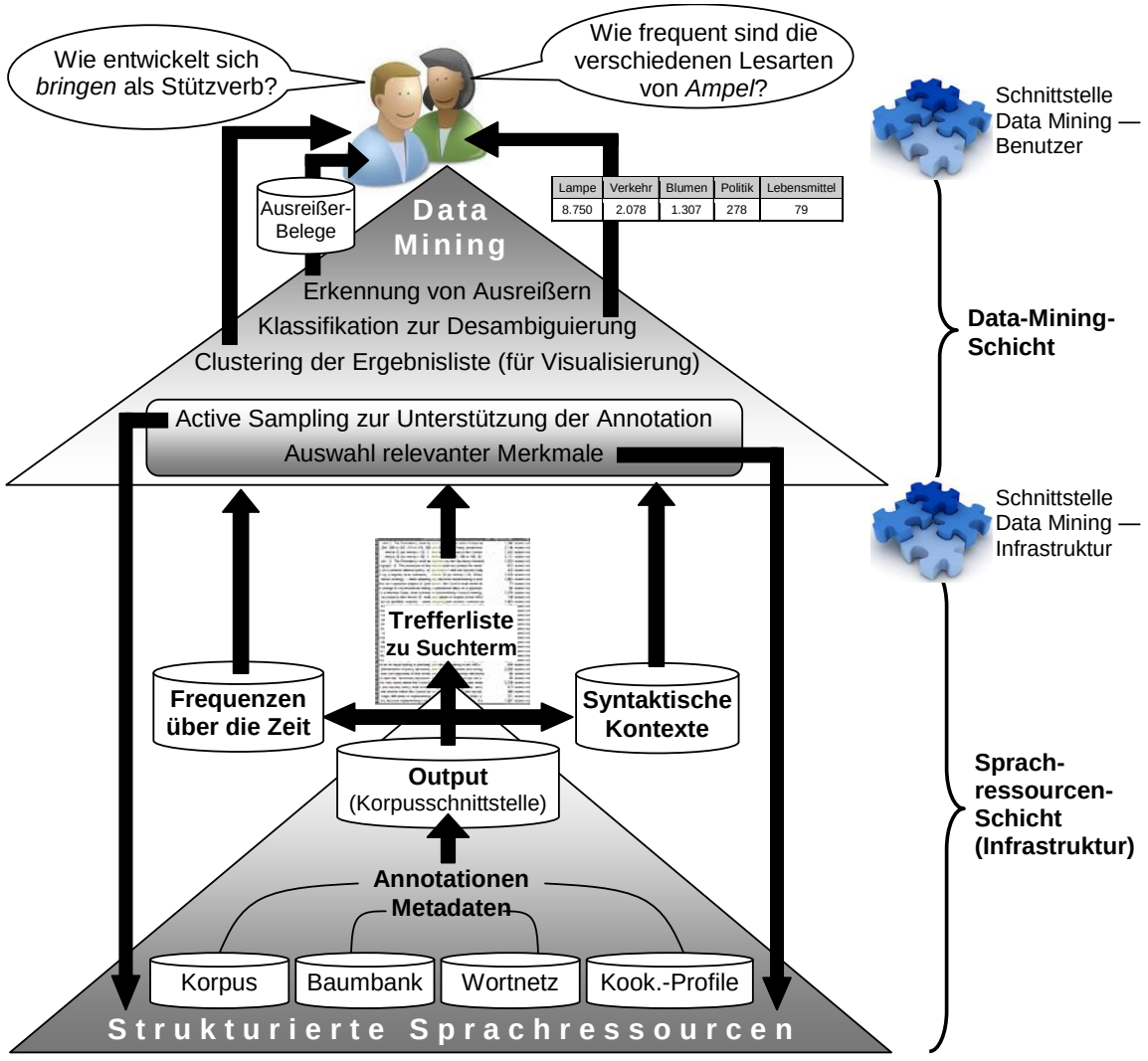

Abbildung 1: Data-Mining zur Verbesserung der korpus-basierten Sprachanalyse.

Genutzte Ressourcen

Verfahren: Die am Lehrstuhl für künstliche Intelligenz der TU Dortmund entwickelten Softwaresysteme SVMlight und RapidMiner befinden sich weltweit im Einsatz und werden auch für die Entwicklung innovativer Verfahren im vorliegenden Projekt genutzt. Bei beiden Systemen handelt es sich um für wissenschaftliche Zwecke frei verfügbare und erweiterbare Software (ggf. nach Rücksprache mit den Herausgebern).

Daten: Datengrundlage für die Lernverfahren und linguistischen Fallstudien sind strukturierte, linguistisch annotierte große Korpora/Baumbanken und lexikalische Ressourcen, die von den Sprachtechnologie-Partnern bereitgestellt werden und über Abfragewerkzeuge und/oder Webservices in CLARIN-D zur Verfügung stehen. Die folgende Tabelle gibt einen Überblick über diese Sprachressourcen und ihre Merkmale:

Ressource	Anbieter	Typ	Daten	Strukturierung	Umfang
DWDS-Kernkorpus	BBAW	annotiertes Korpus	deutschsprachige Texte (1900-2000), ausgewogen in Bezug auf Textsorten und Dekaden	lemmatisiert, wortartenannotiert, Metadaten zu Textsortenbereich und Erscheinungsdatum	100 Mio. Tokens
Referenzkorpus des deutschen Textarchivs (DTA)	BBAW	annotiertes Korpus	deutschsprachige Texte (aktuell 1780-1900), ausgewogen in Bezug auf Textsorten und Dekaden	lemmatisiert, wortartenannotiert, Metadaten zu Textsortenbereich und Erscheinungsdatum	aktuell 532 Bücher, wird erweitert
Deutsches Referenzkorpus (DeReKo)	IDS	annotiertes Korpus	deutschsprachige Texte (ca. 1900-2012) aus unterschiedlichen Textsorten	lemmatisiert, wortartenannotiert, Metadaten zu Textsortenbereichen, Erscheinungsdatum, Thema	> 4 Milliarden Tokens
Wikipedia-Korpus	IDS	annotiertes Korpus	Artikel- und Diskussionsseiten der deutschsprachigen Wikipedia	lemmatisiert, wortartenannotiert, Metadaten zu Erscheinungsdatum, Thema	> 1 Milliarde Tokens
Tübinger Baumbank des Deutschen / Schriftsprache (TüBa-D/Z)	SfS	Baumbank	deutschsprachige Zeitungstexte	lemmatisiert, wortartenannotiert, morphologisch und syntaktisch annotiert, Koreferenzen, klassifizierte Eigennamen	> 65.000 Sätze (> 1.164.000 Tokens)
Tübinger Baumbank des Deutschen / Spontansprache (TüBa-D/S)	SfS	Baumbank	spontansprachliche Dialoge (deutsch)	syntaktische Annotation, Konstituenten auf lexikalischer und phrasaler Ebene, auf Ebene der topologischen Felder sowie auf Satzebene	38.000 Sätze (360.000 Tokens)
Tübinger Baumbank des Deutschen / diachron (TüBa-D/DC)	SfS	Baumbank	deutschsprachige Texte der Sammlung „Projekt Gutenberg“ (diachron; 1210 bis Anfang 20. Jh.)	lemmatisiert, wortartenannotiert, Annotation von Named Entities, Parsebäume	knapp 12 Mio. Sätze (> 258 Mio. Tokens)

Ressource	Anbieter	Typ	Daten	Strukturierung	Umfang
Tübinger partiell geparstes Korpus des Deutschen / Schriftsprache (TüPP-D/Z)	SfS	Baumbank	deutschsprachige Zeitungstexte	Annotationen für Morphologie, Syntax, Satzstruktur, topologische Felder, Chunks	> 200 Mio. Tokens
Dortmunder Chat-Korpus	TU Dortmund	annotiertes Korpus	deutschsprachige Chats aus unterschiedlichen sozialen Handlungsbereichen	Annotationen für chattypische Elemente, z.B. Emoticons, Adressierungen, Asterisk-Ausdrücke oder Zuschreibungen („action messages“)	478 Mitschnitte (140.000 Nutzerbeiträge, > 1 Mio. Tokens)
GermaNet	SfS	Wortnetz	Nomen, Verben und Adjektive. (Deutsches Pendant zum englischen Princeton Word-Net)	Semantische Relationen (Hyponymie, Hyperonymie, Antonymie, Meronymie etc.)	> 93.000 lexikalische Einheiten
Kookkurrenzdatenbank CCDB	IDS	Kookkurrenzen	Kollokationsprofile von Wörtern der geschriebenen Gegenwartssprache	Kookkurrenzen und Belege	Profile zu 220.000 lexikalischen Einheiten
DWDS Wortprofil	BBAW	Kookkurrenzen	Kookkurrenzdaten zu den lexikalischen Einheiten des DWDS-Kernkorpus	Kookkurrenzen und Belege	Profile zu 90.000 lexikalischen Einheiten

Entstehende Ressourcen

Alle entwickelten und erprobten technischen Verfahren werden nach Ablauf des Projekts in Form weiter entwickelbarer Open-Source-Software zur Verfügung stehen. Weiterhin werden die Verfahren in die Infrastrukturen der Sprachtechnologie-Partner eingebunden; konkret sind folgende Integrationsarbeiten geplant:

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW):
Integration als APIs auf der Arbeitsoberfläche der am Ausbau des DWDS-Wörterbuchs beteiligten Lexikographen.

Seminar für Sprachwissenschaft (SfS), Universität Tübingen:
Integration in die Nutzerschnittstelle von WebLicht.

Institut für deutsche Sprache (IDS) Mannheim:
Integration in die Infrastruktur zur Pflege und zum Ausbau des Deutschen Referenzkorpus (DeReKo), in COSMAS bzw. in die gerade im Aufbau befindliche Korpusanalyse-Plattform KorAP sowie über die vom IDS angebotenen Web-Service-APIs in CLARIN-D; DeReKo wird zudem um die von den Klassifikationstools erzeugten Metadaten bzw. Annotationen erweitert.

Da alle Sprachtechnologiepartner auch als CLARIN-D-Zentren fungieren, sichert diese Integration die nachhaltige Verwertbarkeit der Verfahren über die Projektlaufzeit hinaus. Die Ergebnisse der im Rahmen des Projekts durchgeführten linguistischen Fallstudien werden auf der projektbegleitenden Website unter http://www.kobra.tu-dortmund.de dokumentiert werden.

Für die korpus-basierte Erforschung der internetbasierten Kommunikation gibt es bislang keine Referenzkorpora (vgl. Beißwenger/Storrer 2008). Die Dortmunder Linguistik und die DWDS-Arbeitsgruppe an der BBAW arbeiten deshalb gemeinsam am Aufbau eines deutschen Referenzkorpus zur IBK (vgl.

Beißwenger u.a. 2012a) und kooperieren dabei mit Partnern aus anderen europäischen Initiativen, die ähnliche Zielsetzungen verfolgen. Um die Nachhaltigkeit der Ressourcen und die Interoperabilität der dafür entwickelten Werkzeuge zu sichern, werden gemeinsam mit diesen Partnern im Rahmen der *Text Encoding Initiative* (TEI) Annotationsrichtlinien erarbeitet, die auch die Basis für geplante KobRA-Verfahren bilden (vgl. Beißwenger u.a. 2012b, [CMC-WG]).

Kooperationen

Durch die Beteiligung der Sprachtechnologie-Partner ist ein unmittelbarer Bezug zum Infrastrukturprojekt CLARIN-D gegeben: Der Projektpartner in Tübingen ist Koordinator des CLARIN-D-Gesamtverbunds, die BBAW, das IDS und die Universität Tübingen sind als CLARIN-D-Zentren wichtige Anbieter und Multiplikatoren für die im Projekt entwickelten Verfahren. Weiterhin wird der Austausch mit einschlägigen fachspezifischen Arbeitsgruppen von CLARIN-D gepflegt: Kontakte bestehen insbesondere zur Arbeitsgruppe 7 „Angewandte Sprachwissenschaft – Computerlinguistik“, zur FAG 1 „Deutsche Philologie“ und zum Arbeitspaket 8 „Schulungen und Ausbildung“. Außerdem beteiligt sich die Dortmunder Germanistik an einer im Rahmen von CLARIN-D initiierten Arbeitsgruppe zur Optimierung des *Stuttgart-Tübingen Tagsets* (STTS), dem Standard für das deutsche POS-Tagging. Über die Dortmunder Informatik bestehen zudem Kontakte und Kooperationen im Rahmen des DFG-Sonderforschungsbereichs 876 „Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung“, in dem die Dortmunder Informatikpartnerin als Sprecherin fungiert.

Die Fallstudien in der Dortmunder Germanistik sind in laufende Kooperationsprojekte integriert: (1) Das Projekt „Bericht zur Lage der deutschen Sprache“, durchgeführt von der Union der Deutschen Akademien der Wissenschaften und der Deutschen Gesellschaft für Sprache und Dichtung (Fallstudien zu „Stützverbgefügen“); (2) Das DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation“ (Fallstudien im Bereich „Internetbasierte Kommunikation“); (3) Das DFG-Netzwerk „Internet-Lexikographie“ (Fallstudien zur korpus-basierten Lexikographie). Im Bereich internetbasierte Kommunikation/Varietätenlinguistik ist die Dortmunder Germanistik an internationalen Initiativen zum Aufbau und zur Standardisierung von IBK-Korpora beteiligt [vgl. CMC-WG].

Kontaktinformationen

Projektwebsite: http://www.kobra.tu-dortmund.de

Projektkoordination Germanistik TU Dortmund:
Thomas Bartz (thomas.bartz@tu-dortmund.de)
PD Dr. Michael Beißwenger (michael.beisswenger@tu-dortmund.de)
Nadja Radtke (nadja.radtke@tu-dortmund.de)

Ansprechpartner Informatik TU Dortmund:
Christian Pölitz (christian.poelitz@tu-dortmund.de)

Ansprechpartner BBAW:
Dr. Alexander Geyken (geyken@bbaw.de)
Dr. Lothar Lemnitzer (lemnitzer@bbaw.de)

Ansprechpartner IDS:
Dr. Marc Kupietz (kupietz@ids-mannheim.de)
Dr. Andreas Witt (witt@ids-mannheim.de)

Ansprechpartner Uni Tübingen:
Kathrin Beck (kathrin.beck@uni-tuebingen.de)

Referenzen

[CLARIN] Common Language Resources and Technology Infrastructure: http://www.clarin.eu

[CLARIN-D] CLARIN-D: Web- und zentrenbasierte Forschungsinfrastruktur für Geistes- und Sozialwissenschaftler: http://weblicht.sfs.uni-tuebingen.de/

[CMC-WG] Working group „Building & Annotating CMC Corpora“: http://wiki.itmc.tu-dortmund.de/cmc

[COSMAS II] Corpus Search, Management and Analysis System: http://www.ids-mannheim.de/cosmas2/

[DCK] Dortmunder Chat-Korpus: http://www.chatkorpus.tu-dortmund.de

[DEREKO] Deutsches Referenzkorpus: http://www.ids-mannheim.de/dereko

[DWDS] Digitales Wörterbuch der deutschen Sprache (Wörterbuchverbund und Korpora): http://www.dwds.de

[DTA] Deutsches Textarchiv (BBAW): http://www.deutschestextarchiv.de/

[empirikom] Wissenschaftliches Netzwerk Empirische Erforschung internetbasierter Kommunikation: http://www.empirikom.net

[GermaNet]: Lexical semantic Network (SfS/CL Tübingen): http://www.sfs.uni-tuebingen.de/lsd/

[HyTex] Projekt Hypertextualisierung auf textgrammatischer Grundlage: http://www.hytex.tu-dortmund.de

[RapidMiner]: Open Source Umgebung für Data Mining und maschinelles Lernen: http://de.wikipedia.org/wiki/RapidMiner, http://www.rapidminer.com/

[STTS] Arbeitsgruppe zur Optimierung des Stuttgart-Tübingen-Tagsets: http://wiki.ims.uni-stuttgart.de/STTSSite/Home

[SVMlight]: Implementierung einer Support Vector Machine in C, die sich besonders für das maschinelle Lernen aus strukturierten Daten eignet: http://svmlight.joachims.org/

[TüBa-D/Z] Tübinger Baumbank des Deutschen/Zeitungskorpus: http://www.sfs.uni-tuebingen.de/tuebadz.shtml

[TüBa-D/S] Tübinger Baumbank des Deutschen/Spontansprache (Projekt Verbmobil): http://www.sfs.uni-tuebingen.de/tuebads.shtml

[WebLicht] Web-basierte LRT Services für Deutsch: https://weblicht.sfs.uni-tuebingen.de/

Beißwenger, M. & A. Storrer (2008): Corpora of Computer-Mediated Communication. In: A. Lüdeling & M. Kytö (Hrsg.): Corpus Linguistics. An International Handbook. Volume 1. Berlin: de Gruyter, S. 292-308.

Beißwenger, M. & A. Storrer (2011): Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen – Erfahrungen – Desiderate. In: Journal for Language Technology and Computational Linguistics (Themenheft „Language Resources and Technologies in E-Learning and Teaching“, Hrsg. Frank Binder, Henning Lobin & Harald Lüngen), 119-139. Online: http://www.jlcl.org/2011_Heft1/9.pdf

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012a): DeRiK: A German Reference Corpus of Computer-Mediated Communication. In: Proceedings of Digital Humanities 2012.

Beißwenger, M., M. Ermakova, A. Geyken, L. Lemnitzer & A. Storrer (2012b): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative, Issue 3 (November 2012): TEI and Linguistics. Online: http://jtei.revues.org/476 | DOI: 10.4000/jtei.476

Belica, C., M. Kupietz, A. Witt & H. Lüngen (2011): The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: M. Konopka et al. (Hrsg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.4.-24.9.2009. Tübingen: Narr, S. 451-469.

Geyken, A. & T. Hanneforth (2006): TAGH: A Complete Morphology for German Based on Weighted Finite State Automata. In: A. Yli-Jyrä, L. Karttunen und J. Karhumäki (Hrsg.): Finite-State Methods and Natural Language Processing. Berlin: Springer, S. 55–66. Online verfügbar unter http://dx.doi.org/10.1007/11780885_7.

Had, M., F. Jungermann & K. Morik (2010): Relation Extraction for Monitoring Economic Networks. In: D. Hutchison et al. (Hrsg.): Lecture Notes in Computer Science. Berlin: Springer, S. 103-114.

Henrich, V. & E. Hinrichs (2010a): GernEdiT - The GermaNet Editing Tool. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta.

Henrich, V. & E. Hinrichs (2010b): Standardizing Wordnets in the ISO Standard LMF: Wordnet-LMF for GermaNet. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Henrich, V., E. Hinrichs, M. Hinrichs & T. Zastrow (2010): Service-Oriented Architectures: From Desktop Tools to Web Services and Web Applications. In: D. Tufiş and C. Forăscu (Hrsg.): Multilinguality and Interoperability in Language Processing with Emphasis on Romanian. Bucharest: Romanian Academy Publishing House, S. 69-92.

Henrich, V. & E. Hinrichs (2011): Determining Immediate Constituents of Compounds in GermaNet. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2011), Hissar, Bulgaria.

Jungermann, F. & K. Morik (2008): Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining. In: Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems (NLDB 2008), London, UK.

Keibel, H. & C. Belica (2007): CCDB: A Corpus-Linguistic Research and Development Workbench. In: Proceedings of the 4th Corpus Linguistics Conference (CL 2007), Birmingham, UK. Online verfügbar unter: http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/132Paper.pdf.

Klein, W. (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In: J. Scharnhorst (Hrsg.): Sprachkultur und Lexikographie. Von der Forschung zur Nutzung von Wörterbüchern. Frankfurt am Main: Lang, S. 281-309.

Klein, W. & A. Geyken (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: U. Heid et al. (Hrsg.): Lexicographica. International Annual for Lexicography. Berlin: de Gruyter, S. 79-96.

Kupietz, M. & H. Keibel (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: M. Minegishi & Y. Kawaguchi (Hrsg.): Working Papers in Corpus-based Linguistics and Language Education 3 (2009). Tokyo: Tokyo University of Foreign Studies (TUFS), S. 53-59. Online verfügbar unter: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf.

Kupietz, M., C. Belica, H. Keibel & A. Witt (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta. Online verfügbar unter: http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf .

Morik, K., A. Kaspari, M. Wurst & M. Skirzynski (2011): Multi-objective frequent termset clustering. In: Knowledge and Information Systems.

Rössler, M. & K. Morik (2005): Using Unlabeled Texts for Named Entity Recognition. In: Proceedings of the ICML 2005 Workshop on Learning with Multiple Views, Bonn.

Storrer, A. (2007): Corpus-based Investigations on German Support Verb Constructions. In: C. Fellbaum (Hrsg.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum, S. 164-188.

Storrer, A. (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In: K. Knapp (Hrsg.): Angewandte Linguistik. Ein Lehrbuch. 3. vollst. überarb. und erweiterte Aufl. Tübingen: Francke, S. 216-239. Aufgaben und Lösungen online: http://www.studiger.tu-dortmund.de/images/Korpuslinguistik_aufgaben.pdf.

Storrer, A. (2013): Sprachstil und Sprachvariation in sozialen Netzwerken. In: B. Frank-Job, A. Mehler & T. Sutter (Hrsg.): Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW. Wiesbaden: VS Verlag für Sozialwissenschaften.

Tomanek, K. (2010): Resource-aware annotation through active learning. Dissertation, TU Dortmund.

Tomanek, K. & K. Morik (2011): Inspecting Sample Reusability for Active Learning. JMLR Workshop and Conference Proceedings 16 (2011), S. 169-181.