Abstracts – Postersession Sektion Computerlinguistik

Peter Adolphs / Xiwen Cheng / Tina Klüwer / Hans Uszkoreit / Feiyu Xu
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

QA on Structured Data – Chitchatting with Gossip Galore (Demo)

Within the projects Rascalli (EU IST-27596-2004) and KomParse (cofunded by EFRE fonds and ProFIT), DFKI developed the embodied conversational agent "Gossip Galore". Gossip Galore is a virtual female character who knows (almost) all biographic and career-oriented facts about celebrities in the pop music area and beyond. Users interact with her by typing utterances into a chatbox. She answers multimodally with a verbal answer, which is realized using our speech synthesis system MARY, with gestures and optionally also with further multimodal content, which is displayed in a dedicated screen in her environment. Furthermore, the question answering functionality is not provided in isolation but rather in a smooth and connected dialogue between agent and user.

The strength of the system lies in its knowledge base. Whereas current Question Answering systems are typically based on selecting snippets from Web document retrieved by a search engine, Gossip Galore utilizes structured knowledge sources, which allows for providing natural-language answers that are tailored to the current dialog context. We use different data sources for populating the knowledge base: i) data retrieved with Information Wrappers of high-quality knowledge sources, ii) existing data sets acquired with Information Extraction methods applied to (semi-)structured parts of Web documents, and iii) data actively acquired with state-of-the-art Information Extraction methods on free text using the minimally supervised relation extraction system DARE. Based on just a few trusted examples for a target semantic relation, DARE automatically learns linguistic patterns for expressing instances of the relation in a bootstrapping process. Once these patterns are learned, they can be used to extract further, previously unknown relation instances from free text and thereby continuously extend the knowledge base.

Tafseer Ahmed† / Tina Bögel† / Miriam Butt† / Sarmad Hussain‡ /
Muhammad Kamran Malik‡ / Ghulam Raza† / Sebastian Sulger†

Universität Konstanz†, CRULP, FAST-NUCES‡

Ein System zur Transliteration von Urdu/Hindi innerhalb der Urdu ParGram-Grammatik

An der Universität Konstanz wird im Rahmen des ParGram-Projekts (Parallel Grammar; Butt et al. 1999, 2002) eine Computergrammatik für die südasiatische Sprache Urdu entwickelt. Die Grammatik verwendet den Formalismus der Lexical-Functional Grammar (LFG; Dalrymple 2001). Es wird beabsichtigt, auch Hindi abzudecken, welches sehr nah mit Urdu verwandt ist. Urdu, stark beeinflusst vom Persischen, verwendet eine Variation arabischer Schrift. Vokale, dargestellt durch Diakritika, werden in Urdu normalerweise nicht ausgeschrieben. Hindi hingegen wird in Devanagari-Schrift dargestellt. Das Poster diskutiert die Besonderheiten der Urdu-Schrift und präsentiert ein System zur Transliteration von in Unicode kodierter Urdu-Schrift zu einer ASCII-basierten Umschrift.

Das System verwendet einen modul-basierten Ansatz mit mehreren einzelnen Programmen, die gemeinsam eine Pipeline bilden; die Programme können jedoch auch abseits des Transliterators Verwendung finden. Es werden die einzelnen Module vorgestellt, unter anderem das Programm “Diacritize”, welches nicht vorhandene Vokal-Diakritika anhand eines Wortformenlexikons hinzufügt. Eine Auswertung bezüglich Effizienz und Genauigkeit des Transliterators wird angegeben. Das Poster zeigt weiterhin die Integrierung des Systems in das XLE-Programm, eine Plattform zur Entwicklung von LFG-Grammatiken (Crouch et al. 2008). Der Transliterator fügt sich nahtlos in die XLE-Pipeline ein, sodass der ASCII-basierte Output des Systems den Input für das Morphologie-Modul (Bögel et al. 2007) und die Syntax (Butt / King 2007) der Grammatik darstellt.

Die ASCII-Umschrift im Output des Transliterators ist so konzipiert, dass sie als eine neutrale Schnittstelle zwischen Urdu und Hindi fungiert. Aufwand und Größe des Lexikons werden so minimiert; die Möglichkeit, parallel Urdu und Hindi verarbeiten zu können, wird für die Zukunft beibehalten. Es werden kurz Probleme mit dem präsentierten System angesprochen und künftige Absichten zur Verbesserung diskutiert.

Bögel, Tina / Butt, Miriam / Hautli, Annette / Sulger, Sebastian (2007). “Developing a Finite-State Morphological Analyzer for Urdu and Hindi.” In: *Proceedings of the Sixth International Workshop on Finite-State Methods and Natural Language Processing*, Potsdam.

Butt, Miriam / King, Tracy H. (2007). “Urdu in a Parallel Grammar Development Environment.” In: *Language Resources and Evaluation* 41(2), 191–207.

Butt, Miriam / Dyvik, Helge / King, Tracy H. / Masuichi, Hiroshi / Rohrer, Christian. (2002). “The Parallel Grammar Project.” In: *Proceedings of COLING, Workshop on Grammar Engineering and Evaluation*, Taipei, 1–7.

Butt, Miriam / King, Tracy H. / Niño, María-Eugenia / Segond, Frédérique (1999). *A Grammar Writer’s Cookbook*. Stanford: CSLI Publications.

Crouch, Dick / Dalrymple, Mary / Kaplan, Ronald M. / King, Tracy Holloway / Maxwell III, John T. / Newman, Paula (2008). *XLE Documentation*. Palo Alto Research Center.

Dalrymple, Mary (2001). *Lexical Functional Grammar*. New York: Academic Press.

Jasmine Bennöhr
Humboldt-Universität zu Berlin and
Landesinstitut für Lehrerbildung und Schulentwicklung, Hamburg

How much variance in writing competence can be explained by morphology?

This poster presents initial results of research in the project ‘Identification of Indicators for Competence Assessment of Students’ Essays and Development of a Prototype for Computer-Assisted Text Analysis’.

Writing competence is composed of different aspects which interact in a complex way. Much research has been dedicated to automating the grading of essays with varying degrees of success (cf. Burstein et al. 2003 and Valenti et al. 2003). Our goal is not to assign a holistic rating to an essay, but rather to identify indicators or subdomains for competence which supply teachers with more precise, yet manageable and categorized, information on their students’ strengths and weaknesses.

Using statistical methods, the relation of students’ morphology and their writing competence is investigated. This study is based on a corpus of students’ essays in the German language from the city of Hamburg’s longitudinal KESS study and additional language test results for validation. The core of the essay dataset is comprised of approximately 9000 texts which were rated along several dimensions.

In order to assess how much variance in essay writing competence is explained by morphology, indicators on the morphological level are identified. After automatically tagging the essays (Lüdeling 2008) the indicators are quantified and used in regression analysis with writing competence as the dependent variable (cf. readability formulae).

Another multidimensional approach (Biber 2009) is carried out on the same quantification of indicators. A factor analysis determines which morphological dimensions are formed by the indicators. Using a qualitative analysis we test hypotheses about these dimensions and what they represent. We calculate values for each text along all dimensions and show how the values differ according to the assumed writing proficiency. Results for each dimension can be reported to teachers who can consider them when giving feedback to students or planning lessons.

Biber, Douglas (2009). “Multi-dimensional Approaches.” In: Lüdeling, Anke / Kytö, Merja (eds), *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin, 822–855.

Burstein, Jill / Chodorow, Martin / Leacock, Claudia (2003). “CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays.” *IAAI 2003*, 3–10.

Lüdeling, Anke (2008). „Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora.“ In: Walter, Maik / Grommes, Patrick (eds), *Fortgeschrittene Lerner-varietäten*. Tübingen: Niemeyer, 119–140.

Valenti, Salvatore / Neri, Francesca / Cucchiarelli, Alessandro (2003). “An Overview of Current Research on Automated Essay Grading.” *Journal of Information Technology Education* 2, 220–230.

Stefanie Dipper / Lara Kresse / Martin Schnurrenberger / Seong Cho
Ruhr-Universität Bochum

OTTO: A Tool for Diplomatic Transcription of Historical Texts

We present OTTO ("Online Transcription TOol"), a new transcription tool which is designed for diplomatic transcription of historical language data.

Early manuscripts (or prints) exhibit a large amount of peculiarities (special letters, punctuation marks, abbreviations, etc.), which are not easily encoded by, e.g., the ASCII encoding standard. Hence, an important issue with historical corpora is the transcription and encoding of special characters. Basically, there are two options: (i) To use a virtual keyboard, which can support various character sets simultaneously and is operated by the mouse. (ii) To use special characters, such as "$", "@", "#", as substitutes for historical characters. Virtual keyboards are often preferred by casual and new users, because they are “wysiwyg” (in that their keys are labeled by the special characters) and, hence, do not require any extra knowledge. However, the drawback is that “typing” with a computer mouse is rather slow and tedious and, hence, not a long-term solution. By contrast, regular and advanced users usually prefer a system that provides character substitutes, because once the user knows the substitutes, typing them becomes very natural and fast.

OTTO combines the advantages of the two methods. It provides two parallel, synchronized windows for transcribing. In one window, the transcriber types characters and substitutes, while the other window does a live (hence “online”) transformation of the entered line into its actual diplomatic transcription form, using a set of user-defined mapping rules. The tool thus supports easy and fast typing and, at the same time, renders the transcription as close to the original manuscript as possible.

OTTO is used through a standard web browser and supports distributed, collaborative working of multiple parties. It also allows for the annotation of rich, user-defined header (meta) information.

Kerstin Eckart
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Representing Underspecification in a Relational Database

Representing syntactic ambiguities explicitly has different advantages. On the one hand disambiguation can be used where indeed needed and on the other hand ambiguity phenomena can be analysed in detail. A data model supporting this needs to preserve ambiguity without the drawback of having to store an exponential amount of alternative analyses. As a solution, we explore the combination of two representation concepts, namely underspecification and relational databases. Underspecified formats efficiently represent possible alternatives, without storing all analyses separately. Relational databases are optimized for efficient queries on large amounts of data. The combined representation therefore allows us to query different types of ambiguities in large corpora.

For the underspecified representation an enhanced version of the LAF data model (ISO/DIS 24612, Ide/Romary 2006), presented by Kountz et al. (2008) was chosen. Kountz et al. define constraints to encode structural and labelling ambiguities as well as interdependencies between these ambiguity types. For the modelling of the database, the XML-serialisation of the LAF pivot format (GrAF, Ide/Suderman 2007) and the data model presented by Kountz et al. were each transferred into an entity-relationship model and integrated to a conceptual schema for the database. The resulting data structures were implemented as an extension to an existing relational database, the B3DB[1] (cf. Eberle et al. 2009).

In the poster we give an introduction to the database extension and present examples for ambiguity-related queries within the database. We also want to describe some advantages of using the concept of underspecification as a representation format together with a relational database and the possibilities arising from the use of a data model based on an upcoming ISO-standard.

ISO/DIS 24612 Language resource management - Linguistic annotation framework (LAF)

Ide, Nancy / Romary, Laurent (2006). “Representing Linguistic Corpora and Their Annotations.” In: *Proceedings of the Fifth Language Resources and Evaluation Conference, LREC 2006*. Genoa, Italy.

Ide, Nancy / Suderman, Keith (2007). “GrAF: A Graph-based Format for Linguistic Annotations.” In: *Proceedings of the Linguistic Annotation Workshop*, held in conjunction with ACL 2007, Prague.

Eberle, Kurt / Eckart, Kerstin / Heid, Ulrich (2009) „Eine Datenbank als multi-engine für Sammlung, Vergleich und Berechnung möglichst verlässlicher unterspezifizerter syntaktisch/semantischer Satzrepräsentationen.“ Poster presentation at the DGfS 2009, Osnabrück. 2009

Kountz, Manuel / Heid, Ulrich / Eckart, Kerstin (2008). “A LAF/GrAF based Encoding Scheme for underspecified Representations of syntactic Annotations.” In: *Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008*. Marrakech, Morocco.

[1] The B3DB was created in the B3 project of the collaborative research centre 732: *http://www.uni-stuttgart.de/linguistik/sfb732/*

Ralf Gehrke* / Roland Mittmann* / Julia Richling+
* Goethe-Universität Frankfurt am Main
+ Humboldt-Universität Berlin

Referenzkorpus Altdeutsch

In diesem Poster wird das Projekt Referenzkorpus Altdeutsch (HU Berlin, Frankfurt, Jena) vorgestellt. Im Rahmen dieses Projekts werden die überlieferten Texte des Althochdeutschen und des Altsächsischen (etwa aus der Zeit 750–1050, ca. 650.000 Wörter) digitalisiert, annotiert und in einer Online-Datenbank der Wissenschaft zugänglich gemacht.

1. *Erfassung.* Die in Manuskripten tradierten Texte liegen normalisiert in Referenzausgaben vor. Der Großteil der Texte ist bereits über das TITUS-Projekt im HTML-Format verfügbar, die verbleibenden werden gescannt, korrigiert, getaggt und ebenfalls auf den TITUS-Server überführt, wo sie der Öffentlichkeit zur Verfügung stehen.
2. *Wörterbuch-Digitalisierung.* Zu den meisten Referenzausgaben liegen Wörterbücher vor, die unter den Lemmata die einzelnen Belegstellen samt ihrer morphologischen Bestimmung aufführen. Diese werden ebenfalls digitalisiert.
3. *Konvertierung 1.* Die Texte werden dann ins Format der Annotationssoftware ELAN übertragen – samt Sprachbestimmung und hierarchischer Struktur des Textmanuskripts und der Referenzausgabe. Auch Angaben zur Gestalt oder Überlieferungsqualität einzelner Buchstaben bleiben erhalten.
4. *Lemmatisierung und grammatische Bestimmung.* Zu jedem belegten Wort werden mit Hilfe der digitalisierten Wörterbücher morphologische Bestimmung sowie Standardform des Lemmas und Übersetzung per Skript annotiert.
5. *Tagging.* Zur Angabe der Wortart und ggf. grammatischer Form von Lemma und Beleg werden auf dem STTS basierende Tags (DDDTS, DeutschDiachronDigital-Tagset) vergeben.
6. *Standardisierte Textform.* Neben Manuskript- und Referenzfassung kann eine dritte Version der Texte – in der Standardform der jeweiligen Sprachstufe – mit Hilfe der Lemmata und der morphologischen Bestimmung automatisiert ergänzt werden.
7. *Annotierung.* Die Annotatoren prüfen in ELAN die Referenz- sowie die Standardform der Texte, disambiguieren eventuelle Mehrfachzuordnungen bei Lemmata, Übersetzungen und grammatischen Bestimmungen und ergänzen fehlende Angaben. Außerdem fügen sie Daten zu Reimformen und Gliedsätzen hinzu.
8. *Konvertierung 2.* Zusammen mit ebenfalls im Projekt erhobenen philologischen Informationen über die Texte werden die ins PAULA-Format überführten Daten schließlich in die Datenbank ANNIS importiert. Über das ANNIS-Webinterface können die tief annotierten Texte auf allen Ebenen durchsucht werden.

Donhauser, Karin (2007). „Zur informationsstrukturellen Annotation sprachhistorischer Texte“. In: Gippert, Jost / Schmitz, Hans-Christian (Hrsg.). *Sprache und Datenverarbeitung* 31(1-2) (Themenheft Diachrone Corpora, historische Syntax und Texttechnologie), 39–45.

Gehrke, Ralf (2009). „TITUS: datenbank- und internetorientierte Konzepte“. In: Hofmeister, Wernfried / Hofmeister-Winter, Andrea (Hrsg.). *Wege zum Text. Überlegungen zur Verfügbarkeit mediävistischer Editionen im 21. Jahrhundert. Grazer Kolloquium 17. - 19. September 2008.* (Beihefte zu Editio 30.) Tübingen: Niemeyer, 43–51.

Gippert, Jost (1999) “Language-specific Encoding in Multilingual Corpora: Requirements and Solutions”. In: Gippert, Jost (Hrsg.). *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung*. Prag: Enigma, 371–384.

Lüdeling, Anke / Poschenrieder, Thorwald / Faulstich, Lukas C. (2009). „DeutschDiachronDigital – Ein diachrones Korpus des Deutschen“. In: Braungart, Georg / Eibl, Karl / Jannidis, Fotis (Hrsg.) *Jahrbuch für Computerphilologie* 6. Paderborn: Mentis-Verlag, 119–136.

Zeldes, Amir / Ritz, Julia / Lüdeling, Anke / Chiarcos, Christian (2009) “ANNIS: A Search Tool for Multi-Layer Annotated Corpora”. In: *Proceedings of Corpus Linguistics 2009, July 20–23, Liverpool, UK.*

Anton Karl Ingason
University of Iceland

Context-sensitive proofreading for a minority language

Spellcheckers and grammar checkers have been around for a while and in recent years more advanced context-sensitive proofreading tools have started surfacing for large languages such as English. For example the latest version of Microsoft Word can detect errors such as “a nice pear of shoes” (where *pair* was intended). For a less-resourced language like Icelandic, on which the current work focuses, there is limited interest in commercial development of such tools because of the small market size and even those tools that do exist (word-by-word spellchecking) are mostly limited to the most widespread word processor (MS Word).

In this poster+demo we present a fully open-source system (all components LGPL-licensed) for rule-based context-sensitive proofreading for Icelandic built on the LanguageTool framework (Naber 2003; Miłkowski submitted). The system makes use of the IceNLP toolkit (Loftsson / Rögnvaldsson 2007) which includes a PoS-tagger (Loftsson 2008) and a lemmatizer (Ingason et al. 2008). Sentence segmentization is based on SRX, an open standard processed by open libraries. This is the first context-sensitive proofreading tool for Icelandic and because of its integration with an international open-source project the development effort enjoys advantages over commercial systems that are critical in a setting where resources are limited. The LanguageTool team maintains integration with OpenOffice.org so that our system is already available for OOo. Free software developers around the world are already integrating the LanguageTool framework with the Firefox web browser (in Brazil) and the Thunderbird mail client (in Russia), thus making Icelandic proofreading available for those systems, while our efforts are focused on the Icelandic-specific backend.

Thus, we argue that for any less-resourced language like Icelandic to be able to keep up with the ever progressing world of language technology – a free and open-source approach can be the key to success.

Ingason, Anton Karl / Helgadóttir, Sigrún / Loftsson, Hrafn / Rögnvaldsson, Eiríkur (2008). “A Mixed Method Lemmatization Algorithm Using Hierarchy of Linguistic Identities (HOLI).” In: Nordström, Bengt / Rante, Aarne (eds). *Advances in Natural Language Processing, Proceedings of the 6th International Conference on NLP, GoTAL 2008*. Gothenburg, Sweden.

Loftsson, Hrafn (2008). “Tagging Icelandic Text: A Linguistic Rule-based Approach.” *Nordic Journal of Linguistics* 31(1), 47–72.

Loftsson, Hrafn / Rögnvaldsson, Eiríkur (2007). “IceNLP: A Natural Language Processing Toolkit for Icelandic.” In: *Proceedings of InterSpeech 2007, Special session: ‘Speech and language technology for less-resourced languages’*. Antwerp, Belgium.

Miłkowski, Marcin (Submitted). Developing a Rule-based Proofreading Tool as Open Source. *Software Practice and Experience*.

Naber, Daniel (2003). *A Rule-Based Style and Grammar Checker*. Master’s thesis, Universität Bielefeld, Bielefeld, Germany.

Katja Keßelmeier / Tibor Kiss / Antje Müller / Claudia Roch / Tobias Stadtfeld / Jan Strunk
Sprachwissenschaftliches Institut, Ruhr-Universität Bochum

Developing an Annotation Scheme and a Reference Corpus for Preposition Senses in German

We present an annotation scheme for preposition senses in preposition-noun combinations (PNCs) and PPs in German. PNCs are combinations of prepositions with determinerless nominal projections such as *auf Anfrage* ('after being asked'), *unter Voraussetzung* ('under prerequisite'), *mit Vorbehalt* ('with reservation'). They represent an anomaly in the grammar because they violate the rule on the realization of countable singular nouns, viz. that such nouns have to appear with a determiner. For some time, PNCs have been treated as exceptions, but recent research has shown that they are indeed productive and no more idiomatic than other phrasal combinations (Dömges et al. 2007, Kiss 2007). In search of licensing conditions for PNCs, it turns out that not every preposition can appear in a PNC, probably also depending on the meaning of the preposition.

In order to investigate the distribution of preposition senses, we have developed an annotation scheme which will facilitate manual annotation of the meaning of prepositions in these constructions. We plan to construct a reference corpus of preposition senses which will serve as input for annotation mining with the ultimate goal of identifying the pertinent licensing conditions for PNCs. This corpus will be large enough to permit the training of an automatic sense annotator eventually. Automatic annotation will not only rely on the annotations of preposition meaning, but also on various annotations on the lexical, syntactic, and conceptual level, including in particular a complex conceptual annotation of nominal complements provided by HaGenLex (Hartrumpf et al. 2003).

Dömges, Florian / Kiss, Tibor / Müller, Antje / Roch, Claudia (2007). "Measuring the Productivity of Determinerless PPs." In: *Proceedings of the ACL 2007 Workshop on Prepositions*, Prague 31–37.

Hartrumpf, Sven / Helbig, Hermann / Osswald, Rainer (2003). "The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment." *Traitement automatique des langues* 44(2), 81–105.

Kiss, Tibor (2007). „Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen." *Zeitschrift für Sprachwissenschaft* 26, 317–345.

Valia Kordoni / Yi Zhang
German Research Centre for Artificial Intelligence (DFKI GmbH) &
Department of Computational Linguistics, Saarland University

A Dynamic Annotation of the Wall Street Journal sections of the Penn Treebank

In this poster, we present an on-going project whose aim is to produce rich syntactic and semantic annotations for the Wall Street Journal (WSJ) sections of the Penn Treebank (PTB; Marcus et al. 1993). In doing so, we are not only focusing on the various stages of the semi-automated annotation process we have adopted, but we are also showing that rich linguistic annotations, which can apart from syntax also incorporate semantics, may ensure that treebanks are guaranteed to be truly sharable, re-usable and multi-functional linguistic resources.

The task is being carried out with the help of the English Resource Grammar (ERG; Flickinger 2002), which is a hand-written grammar for English in the spirit of the framework of Head-driven Phrase Structure Grammar (Pollard/Sag 1994). To aid the treebank development we use automatic parsing outputs as guidance. Many state-of-the-art parsers are nowadays able to efficiently produce large amounts of annotated syntactic structures with relatively high accuracy. This approach has changed the role of human annotation from a labour-intensive task of drawing trees from scratch to a more intelligence-demanding task of correcting parsing errors, or eliminating unwanted ambiguities (cf. the Redwoods Treebank). In other words, the treebank under construction in this project is in line with the so-called dynamic treebanks (Oepen et al. 2002). We rely on the HPSG analyses produced by the ERG, and manually disambiguate the parsing outputs with multiple annotators. The development is heavily based on the DELPH-IN (http://www.delph-in.net) software repository and makes use of the ERG, PET (Callmeier 2001), an efficient unification-based parser which is used in our project for parsing the WSJ sections of the PTB, and [incr tsdb()] (Oepen 2001), the grammar performance profiling system we are using, which comes with a complete set of GUI-based tools for treebanking.

Callmeier, Ulrich (2001). *Efficient Parsing with Large-scale Unification Grammars*. MA thesis, Universität des Saarlandes, Saarbrücken, Germany.

Flickinger, Dan (2002). “On building a more efficient grammar by exploiting types.” In: Oepen, Stephan / Flickinger, Dan / Tsujii, Jun’ichi / Uszkoreit, Hans (eds). *Collaborative Language Engineering*. Stanford, CA: CSLI Publications, 1–17.

Marcus, Mitchell P. / Santorini, Beatrice / Marcinkiewicz, Mary Ann (1993). “Building a large annotated corpus of English: The Penn Treebank.” *Computational Linguistics*, 19(2), 313–330.

Oepen, Stephan / Toutanova, Kristina / Shieber, Stuart / Manning, Christopher / Flickinger, Dan / Brants, Thorsten (2002). “The LinGO Redwoods Treebank: motivation and preliminary applications.” In: *Proceedings of COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*. Taipei, Taiwan.

Johann-Mattis List / Sven Sommerfeld
Heinrich-Heine-Universität Düsseldorf

Cultural Phylogenies: From Words to Weddings
Basic Vocabulary and the Quest for a Universal Test List

Recently, cultural phylogenetic studies have enjoyed a steady growth in popularity. Often, such studies involve mapping some cultural traits on phylogenetic trees (e.g. Fortunato et al. 2006, Jordan et al. 2009). These “reference” trees are based on linguistic data and are claimed to resemble more or less the true genealogical history of the cultures in comparison (cf. e.g. Mace/Pagel 1994). As a consequence, the validity of these analyses is crucially dependent on the quality of the linguistic data upon which the respective reference trees are built.

When Morris Swadesh at the end of the 1940s first proposed the idea that a certain non-cultural part of human languages’ lexicon is universal, stable and relatively resistant to borrowing, he initialized a controversial debate regarding the validity of this claim, which is still waiting for a proper solution. Since then, many scholars have tried to modify the original test list which Swadesh used for his lexicostatistical calculations.

In our presentation we shall introduce a database of about 50 basic vocabulary lists, proposed by various scholars, which we have been collecting and compiling for our research project on evolution and classification in linguistics, biology and history of science. These lists are interlinked in different ways, enabling direct comparison of different lists, network presentation of interlinked items, and checking for root repetition and semantic connections between the basic vocabulary items. Accompanied by statistical evidence we show that lexicostatistic analysis in principle lacks reliability and objectivity in terms of item selection and item translation that may render the results inaccurate if not useless. Used as a tool in phylogenetic studies in and outside linguistics (cf. e.g. Gray/Atkinson 2003, Mace/Holden 2005, Gray et al. 2009), findings based on this method should be taken with extreme care.

Fortunato, Laura / Holden, Clare / Mace, Ruth (2006). "From Bridewealth to Dowry? A Bayesian Estimation of Ancestral States of Marriage Transfers in Indo-European Groups." *Human Nature* 17(4), 355–376.

Gray, Russell D. / Atkinson, Quentin D. (2003). "Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin." *Nature* 426, 435–439.

Gray, Russell D. / Drummond, Alexei J. / Greenhill, Simon J. (2009). "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement." *Science* 323, 479–483

Jordan, Fiona M. / Gray, Russell D. / Greenhill, Simon J. / Mace, Ruth (2009). "Matrilocal Residence is Ancestral in Austronesian Societies." In: *Proc. R. Soc. B* 276, 1957–1964.

Mace, Ruth, / Holden, Clare J. (2005). "A Phylogenetic Approach to Cultural Evolution." *Trends in Ecology and Evolution* 20, 116–121.

Mace, Ruth / Pagel, Mark (1994). "The Comparative Method in Anthropology." *Current Anthropology* 35(5), 549–564.

Rainer Osswald / Thomas Gamerschlag / Brigitte Schwarze
Heinrich-Heine-Universität Düsseldorf

A Lexical Resource of Stative Verbs

Existing lexical-semantic resources such as WordNet, FrameNet, or VerbNet do not pay particular attention to the systematic analysis and classification of stative verbs, with consequential gaps and inconsistencies in this domain. We describe an ongoing project that aims at a detailed morpho-syntactic and semantic description of stative verbs, including stative readings of non-stative verbs. The specific focus of the project is on what we call *attributive* (or *dimensional*) *verbs*, which are stative verbs that characterize an entity by specifying the value of one of its attributes. The notion of attribute is taken here in a broad sense that subsumes common cases such as weight, duration, name, and price but also location, meaning, function, etc.

We compiled a lexical database of about 2000 stative German verbs, primarily based on an exhaustive analysis of the verb entries in the *Duden - Deutsches Universalwörterbuch*. Each of the verbs has been classified as to whether it is attributive and, if so, which attribute it encodes. The attributes are formalized in a type hierarchy that has been developed in the course of the classification process. Each entry contains information on whether the attribute value is incorporated by the verb or, if not, how the value is realized syntactically. Moreover, attributive verbs that have a basic non-stative reading are classified with respect to the underlying stativization mechanism such as reflexivization, metonymic shift, or generic interpretation. Since it is also part of the project to compare the lexicalization of attributive verbs cross-lingually, we are currently expanding the database to include French and English entries following the same classification scheme.

Marc Reznicek* / Cedric Krummes+ / Anke Lüdeling* / Hagen Hirschmann* /
Astrid Ensslin+ / Jia Wei Chan*
*Humboldt-Universität zu Berlin
+Bangor University

„Dass wenn man etwas will, muss man dafür arbeiten“- Herleitung und Anwendung von Zielhypothesen im Lernerkorpus Falko

Unser Poster präsentiert eine Studie zur Fehlerannotation in Lernerkorpora. Diese Studie ist Teil des dreijährigen DFG/AHRC-Projekts „What's hard in German?“, das seit Juli 2009 als internationale Kooperation der Humboldt-Universität zu Berlin und der Bangor University (Wales) durchgeführt wird. Hauptschwerpunkt des Projekts ist die Ermittlung solcher sprachlichen Strukturen, die beim Erwerb des Deutschen als Fremdsprache besondere Schwierigkeiten bereiten. Neben der Untersuchung von Under- und Overuse bezüglich einer muttersprachlichen Kontrollgruppe interessieren uns dabei die Fehlerannotation und ihre Grundlagen.

Fehleranalysen in Lernerkorpora basieren immer auf der Hypothese eines Korrektors über die Struktur, die der Lerner ausdrücken wollte. Diese „Zielhypothese“ von unterschiedlichen Annotatoren gleich bearbeiten zu lassen, ist nicht trivial (vgl. Lüdeling 2008). Komplexitäten beruhen zum einen auf unterschiedlichen Entscheidungen der Annotatoren, zum anderen auf unterschiedlichen Ebenen linguistischer Beschreibung.

Aus diesem Grund wurden für die systematische Annotation mit Zielhypothesen im Lernerkorpus FALKO (Lüdeling et al. 2008) Richtlinien erarbeitet und auf das gesamte Korpus angewandt. Das Poster stellt diese Richtlinien vor und zeigt anhand eines Beispiels alle Schritte der Normalisierung und Herleitung einer Zielhypothese auf mehreren Ebenen.

Auf einer Ebene werden Abweichungen rein morphosyntaktischer Art vermerkt. Eine weitere Ebene bezieht lexikalische und pragmatische Abweichungen mit ein, deren Annotation durch den großen Interpretationsspielraum besondere Herausforderungen birgt. Dazu stellen wir eine Pilotstudie zum Inter-Rater-Agreement vor.

Aus der Zielhypothese können automatisch Fehlerannotationen generiert werden, sodass ohne eine zusätzliche manuelle Interpretation z.B. bestimmte Wortstellungsphänomene bei Lernern untersucht werden können. Dies wird in einer zweiten Pilotstudie zu mehrfach eingebetteten Konjunktionalsätzen illustriert.

Lüdeling, Anke (2008). „Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora“. In: Walter, Maik / Grommes, Patrick (Hrsg.). *Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung*. (Linguistische Arbeiten 520.). Tübingen: Niemeyer / Deutsche Gesellschaft für Sprachwissenschaft, 119–140.

Lüdeling, Anke / Doolittle, Seanna / Hirschmann, Hagen / Schmidt, Karin / Walter, Maik (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache* (2), 67–73.

Julia Ritz / Christian Chiarcos / Heike Bieler
Sonderforschungsbereich 632 „Informationsstruktur“

On the Information Structure of Expletive Sentences: An Empirical Study across Multiple Layers of Annotation

It is obvious that the *semantic* content of an expletive sentence like (1a.) can be conveyed more efficiently, e.g. in (1b./c.). This study investigates possible *pragmatic* motivations for violating the Gricean principle of efficiency.

(1) a. Es liegen zur Zeit etwa 4.500 Bewerbungen vor.
b. Zur Zeit liegen etwa 4.500 Bewerbungen vor.
c. Etwa 4.500 Bewerbungen liegen zur Zeit vor.

The empirical basis of this study is TÜBA-D/Z, a German newspaper corpus annotated with parts of speech, syntax, topological fields, and coreference. In TÜBA-D/Z, 101 expletive sentences were found. An inspection of these examples raised the hypothesis that expletive constructions (as opposed to their 'canonical' counterparts, c.f. Example 1) have whole sentence or TAM (tense/aspect/mood) focus, i.e. the speaker emphasizes the event or state character, rather than a concrete object or referent. If this were the case, we expect constituents in expletive sentences to be co-referenced to a lesser degree than constituents of other sentences. We compared the first constituents of the Mittelfeld in expletive sentences to those in sentences with other material in the Vorfeld and found a significant difference between the two groups of sentences (χ^2=10.1, $p < .0025$; very few coreferential entities in expletive sentences). So far, one could also argue that expletive sentences were presentational constructions. To test this, we will present a comparison of whether constituents from the respective sentence groups are mentioned subsequently in the text. An equal or higher number of expletive constructions with subsequently mentioned constituents would be evidence for the 'presentational construction' hypothesis, a significantly lower number for the 'whole sentence/TAM focus' hypothesis. Results of this second experiment will be discussed on the poster, and the corpus will be available for further querying and inspection in a demo of ANNIS2 (http://www.sfb632.uni-potsdam.de/~d1/annis), a tool for querying and visualizing multi-layer corpora.

Christian Rohrer / Sina Zarrieß
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Right-node Raising (RNR) and Leftward Deletion (LD) in a Large-Coverage LFG Grammar

We present an implementation for two types of non-constituent coordination, RNR and LD, in a German LFG grammar. Non-constituent coordination poses problems for symbolic grammars because it allows coordination of material that does not correspond to syntactic constituents (Maxwell/Manning 1996, Forst/Rohrer 2009).

Our implementation of RNR accounts for coordinations where the raised material corresponds to DPs, PPs, ADVPs and sentential complements. We do not consider coordinations involving partial DPs or PPs.

(1) [Hans kauft] und [Maria verkauft] [$_{\text{RNR}}$ Aktien].
(2) [Lina will] und [Maria soll] [$_{\text{RNR}}$ abnehmen].

The depth of embedding of the raised constituent may differ between the conjuncts:

(5) [Wulf sagte] und [Merkel soll bestätigen], [$_{\text{RNR}}$ daß Steuern gesenkt werden].

Several constituents may be raised:

(7) [Hans kaufte] und [Maria verkaufte] [$_{\text{RNR}}$ gestern Aktien].

Our analysis allows incomplete constituents on the level of C-structure. The well-formedness of the construction is checked in F-structure. Sentence (1) is assigned a C-structure that coordinates two (incomplete) CPs. The F-structure of the raised constituent is distributed over the conjuncts.

The raised constituents are annotated with the function they have in the position from which they are "extracted". The annotation follows the same principles as the annotation of topicalized constituents. RNR can be considered an instance of LD (cf. Reich 2009):

(9a) ... weil Adam Äpfel und Maria Birnen als Vorspeise aß.

We introduce a rule allowing coordination of tuples of arguments/adjuncts in the *Mittelfeld*. The constituents to the right of the coordination are distributed in the same way as in RNR. Our analysis is implemented in a large-scale, reversible grammar. We can generate the examples discussed (demo available). The efficiency of the grammar is barely affected. Exact figures will be given.

Maxwell, John T. / Manning, Christopher D. (1996). "A theory of non-constituent coordination based on finite-state rules." In: Butt, Miriam / King, Tracy Holloway (eds). *Proceedings of the LFG '96 Conference.* Rank Xerox, Grenoble.

Reich, Ingo (2009). "Ellipsis." To appear in: Maienborn, Claudia / von Heusinger, Klaus / Portner, Paul (eds). *Semantics: An International Handbook of Natural Language Meaning.* Berlin: De Gruyter.

Forst, Martin / Rohrer, Christian (2009). "A flat structure for German VP coordination." To appear in: Butt, Miriam / King, Tracy Holloway (eds). *Proceedings of the LFG '09 Conference.* Cambridge, UK.

Karina Schneider-Wiejowski
Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft

Produktivität und Produktivitätswandel deutscher Derivationssuffixe

Derivationelle Produktivität ist graduell. Es gibt bestimmte Regeln der Wortbildung, die sehr häufig angewandt werden wie z.B. Wortbildungen mit dem Suffix *-ung* (Computerisier-*ung*, Digitalisier-*ung*). Produktivität ist aber auch ein Phänomen, das dem Sprachwandel ausgesetzt ist. Ein Suffix, das in älteren Sprachstufen produktiv war, aber bei heutigen Neubildungen keine Rolle mehr spielt, ist das Suffix *-sal* (Schick-*sal*). Es werden keine neuen Wortbildungen mehr geschaffen, obwohl natürlich noch lexikalisierte Bildungen in der Sprache verankert sind. Des Weiteren kann die Produktivität durch bestimmte Restriktionen eingeschränkt sein. Beispielsweise wird sich das nomenbildende Suffix *-chen* nicht an Stämme binden, die auf *-ch* enden.

In Zeiten der Korpuslinguistik ist es möglich, Produktivität und Wortbildungwandel anhand von Korpora empirisch zu untersuchen. Dazu wurden seit Anfang der 90er Jahre einige Produktivitätsmaße entwickelt (vgl. Baayen 1992). Insbesondere Baayens Formel P=V (1)/N (V (1)= hapax legomena, N= Tokens) wird häufig zur Produktivitätsmessung verwendet. So bekommt man die Möglichkeit, verschiedene Affixe hinsichtlich ihrer Produktivität miteinander zu vergleichen.

Die Studie, die hier vorgestellt wird, beschäftigt sich mit der Frage der derivationellen Produktivität und ihrem Wandel im Deutschen. Untersucht wurde dabei eine Reihe von substantiv- und adjektivbildenden Suffixen des Deutschen. Als Datengrundlage konnte dazu u. a. das DWDS verwendet werden, das als Korpus der Gegenwartssprache bezeichnet werden kann. Die Auswertung der Daten zeigt zum einen, dass es in der deutschen Sprache Verschiebungen in der Produktivität von Derivationssuffixen gegeben hat, sodass ein Wortbildungswandel stattfindet. Zum anderen aber kann auch gezeigt werden, dass kategorische Aussagen, die man in der linguistischen Literatur zur Produktivität findet (vgl. dazu Fleischer/Barz 1995, Lohde 2006, Motsch 2004), Affixe seien "produktiv", "sehr produktiv", "aktiv", "inaktiv" oder "beschränkt aktiv", hinterfragt werden müssen, wenn die korpuslinguistische Untersuchung für zwei als produktiv beschriebene Affixe deutlich voneinander abweichende Werte hervorbringt.

Baayen, Harald (1992). "Quantitative Aspects of Morphological Productivity." *Yearbook of Morphology* 1991, 109–149.

Fleischer, Wolfgang / Barz, Irmhild (1995). *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Max Niemeyer Verlag.

Lohde, Michael (2006). *Wortbildung des modernen Deutschen. Ein Lehr- und Übungsbuch*. Tübingen: Narr Studienbücher.

Motsch, Wolfgang (2004). *Deutsche Wortbildung in Grundzügen. 2. überarbeitete Auflage*. Berlin: de Gruyter.

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Comparing Computational Models of Selectional Preferences – Second-order Co-Occurrence vs. Latent Semantic Clusters

Selectional preferences (i.e., semantic restrictions on the realisation of predicate complements) are of great interest to research in Computational Linguistics, both from a lexicographic and from an applied (w.r.t. data sparseness) perspective. This poster presents a comparison of three computational approaches to selectional preferences: (i) an intuitive distributional approach that uses second-order co-occurrence of predicates and complement properties; (ii) an EM-based clustering approach that models the strengths of predicate–noun relationships by latent semantic clusters (Rooth et al. 1999); and (iii) an extension of the latent semantic clusters by incorporating the MDL principle into the EM training, thus explicitly modelling the predicate–noun selectional preferences by WordNet classes (Schulte im Walde et al. 2008).

The motivation of our work was driven by two main questions: Concerning the distributional approach, we were interested not only in how well the model describes selectional preferences, but moreover which second-order properties were most salient. For example, a typical direct object of the verb *drink* is usually fluid, might be hot or cold, can be bought, might be bottled, etc. So are adjectives that modify nouns, or verbs that subcategorise nouns salient second-order properties to describe the selectional preferences of direct objects? Our second interest was in the actual comparison of the models: How does a very simple distributional model compare to much more complex approaches, especially with respect to model (iii) that explicitly incorporates selectional preferences?

Rooth, Mats / Riezler, Stefan / Prescher, Detlef / Carroll, Glenn / Beil, Franz (1999). “Inducing a Semantically-Annotated Lexicon via EM-Based Clustering.” In: *Proceedings ot the 37th Annual Meeting of the Association for Computational Linguistics.*

Schulte im Walde, Sabine / Hying, Christian / Scheible, Christian / Schmid, Helmut (2008). “Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences.” In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics.*

Einar Freyr Sigurðsson
University of Iceland

The Icelandic Diachronic Treebank

The poster describes ongoing work which is a part of a bigger project, IceBLARK (iceblark.wordpress.com). The goal is to construct a diachronic phrase structure treebank — from Old to Modern Icelandic — using methods that efficiently utilize existing resources. This is done in co-operation with the treebankers at the University of Pennsylvania. Our treebank annotation scheme is thus compatible with theirs, the Penn Treebank and the Penn Parsed Corpora of Historical English. The workflow relies in part on a substantial work that has already been developed, namely the open source IceNLP toolkit (http://sourceforge.net/projects/icenlp/) which contains a PoS-tagger for Icelandic and a shallow parser.

Icelandic has been the focus of much attention by syntacticians for the past decades. The evolution of finite verb agreement and the New Passive are a main focus of our inquiry at this stage. Finite verb agreement with nominative objects, as well as non-agreement, is found both in Old and Modern Icelandic (Jónsson, to appear). In order to advance beyond such basic observations and explain the phenomena one must be able to account for how the statistical distribution of such patterns evolves.

If non-agreement has become much more common in the development towards Modern Icelandic, this could explain the increasing — although rare — use of accusative objects instead of nominative ones (Árnadóttir/Sigurðsson, ms.). The much debated New Passive is a recent, ongoing change (e.g. Maling/Sigurjónsdóttir 2002, Eythórsson 2008). There have been various attempts to explain its origin but without concrete diachronic comparison against actual empirical facts, it is impossible to test such hypotheses.

We present results that address those areas of inquiry in the study of Icelandic syntax, and demonstrate the methods we have used to construct an information rich treebank by making efficient use of existing resources of Icelandic language technology.

Árnadóttir, Hlíf / Einar Freyr Sigurðsson (ms.) “The Glory of Non-agreement: The Rise of a New Passive.”

Eythórsson, Thórhallur (2008). “The New Passive in Icelandic Really is a Passive.“ In: Eythórsson, Thórhallur (ed.) *Grammatical Change and Linguistic Theory. The Rosendal papers*. Amsterdam: Benjamins, 173–219.

Jónsson, Jóhannes Gísli (to appear). “Samræmi við nefnifallsandlög [Agreement with nominative objects].” To appear in: Thráinsson, Höskuldur / Sigurðsson, Einar Freyr (eds). *Tilbrigði í íslenskri setningagerð [Variation in Icelandic syntax].*

Maling, Joan, / Sigurjónsdóttir, Sigríður (2002). “The ‘New Impersonal’ Construction in Icelandic.” *Journal of Comparative Germanic Linguistics* 5, 97-142.

Stefanie Simon / Yannick Versley / Erhard Hinrichs
Sonderforschungsbereich 833 „Bedeutungskonstitution“, Universität Tübingen

(A-)Symmetrie als Unterscheidungskriterium bei der Annotation von Diskursrelationen

Bei der Annotation von Diskursrelationen ist die Rolle des verwendeten Inventars entscheidend: Während Mann und Thompsons (1988) *Rhethorical Structure Theory* von einem offenen Inventar ausgeht, beschränken sich neuere Annotationsvorhaben auf kleinere Mengen von Relationen, etwa 16 bei der Penn Discourse Treebank (Prasad et al. 2008). Für eine reliable Annotation ist es jedoch wünschenswert, das Relationsinventar nicht nur überschaubar zu halten sondern die getroffenen Unterscheidungen darüber hinaus mittels linguistischer Tests zu motivieren und abzusichern.

In der linguistischen Literatur bereits untersucht (Umbach/Stede 1999; Lang 2000; Hinrichs/Lau 2008) ist das symmetrische Verhältnis bei Kontrastrelationen. Ein solches Kriterium der Symmetrie – unabhängig von der syntaktischen Realisierung durch eine hypotaktische oder parataktische Konstruktion - ist nicht nur für die Unterscheidung kontrastiver und konzessiver Lesarten von *aber* hilfreich, sondern kann auch der Trennung temporaler und kontrastiver Lesarten von *während* dienen.

Für das laufende Vorhaben, die Tübinger Baumbank des Deutschen/Schriftsprache (TüBa-D/Z) durch eine Annotation von Diskursrelationen zu ergänzen, erlaubt dieses Kriterium eine stärkere Systematisierung der Diskursrelationen und eine größere Trennschärfe innerhalb der gewählten Unterscheidungen.

Hinrichs, Erhard/Lau, Monica (2008). “In Contrast - A Complex Discourse Connective.” In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08).*

Lang, Ewald (2000). “Adversative Connectors on Distinct Levels of Discourse: A Re-examination of Eve Sweetser's Three-level Approach.” In: Couper-Kuhlen, Elizabeth / Kortmann, Bernd (Hrsg.), *Cause, condition, concession, contrast: cognitive and discourse perspectives.* Berlin: de Gruyter.

Mann, William / Thompson, Sandra (1988). “Rhetorical Structure Theory: Toward a Functional Theory of Text Organization.” *Text* 8(3), 243–281.

Prasad, Rashmi / Dinesh, Nikhil / Lee, Alan / Miltsakaki, Eleni / Robaldo, Livio / Joshi, Aravind / Webber, Bonnie (2008). “The Penn Discourse Treebank 2.0.” In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08).*

Umbach, Carla / Stede, Manfred, (1999). *Kohärenzrelationen: Ein Vergleich von Kontrast und Konzession*. KIT Report 148, FB Informatik, TU Berlin. Available at: http://flp.cs.tu-berlin.de/publikationen/kit/r148/Kohaerenzrelationen.pdf.

Jan Strunk
Sprachwissenschaftliches Institut, Ruhr-Universität Bochum

Investigating Relative Clause Extraposition in German Using an Enriched Treebank

I describe the construction of a corpus for research on relative clause extraposition in German based on the treebank Tübinger Baumbank des Deutschen / Schriftsprache (TüBa-D/Z) (Telljohann et al. 2005). I also define an annotation scheme for the relations between relative clauses and their antecedents which is added as a second annotation level to the syntactic trees. This additional annotation level allows for a direct representation of the relevant parts of the relative construction and also serves as a locus for the annotation of additional features which are partly automatically derived from the underlying treebank and partly added manually using the tool SALTO (Burchardt et al. 2006).

The corpus is intended for research on relative clause extraposition both from a production perspective (building a model of relative clause extraposition as a syntactic alternation) and a comprehension perspective (attachment disambiguation of relative clauses). Moreover, it can be used to test claims made in the theoretical literature on extraposition.

I will report on the results of some pilot studies testing such claims with regard to the influence of syntactic locality, definiteness, and restrictiveness on relative clause extraposition. These pilot studies show that although the theoretical claims often go in the right direction, they go too far by positing categorical constraints that are not supported by the corpus data and thus underestimate the complexity of the data.

Burchardt, Aljoscha / Erk, Katrin / Frank, Anette / Kowalski, Andrea / Padó, Sebastian (2006). "SALTO – A versatile multi-level annotation tool." In: *Proceedings of LREC 2006*, May 22-28, 2006, Genoa, Italy.

Telljohann, Heike / Hinrichs, Erhard W. / Kübler, Sandra / Zinsmeister, Heike (2005). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Tübingen: Seminar für Sprachwissenschaft, Universität Tübingen.

Thorsten Trippel / Sami Awad / Marc Bohnes / Patrick Dunkhorst / Carolin Kirchhof
Universität Bielefeld

INFORMER: A Rich Interface for Accessing Highly Structured Linguistic Data

INFORMER, the INterface FOr Rich MEtadata Resources, is a new approach in accessing highly annotated multimodal or otherwise deeply structured linguistic data to satisfy the information needs of users without requiring knowledge of operators, a querying language, or data structures. This characteristic feature of INFORMER is achieved by composing the search interface in a purely data-driven manner, that is, search fields and hierarchies in the user interface are filled based on the data categories and bits of information available in the underlying data. The advantage of this is that a user of such an interface will never end up with an empty result set while being guided to an answer. The linguistic approach is reflected in several ways. Unlike the NXT query language or AGTK, a user does not need to learn a query language, though still has the option of creating rather complex, hierarchy using searches. For modeling the combination of search options in the interface a context-free grammar was used. Presenting relevant options to the user was achieved by lexicographic techniques. An expansion of searches by synonyms, translations, or other related words is accomplished by integrating a termbase. The system is currently implemented for a corpus of written texts, the next step is the integration of multimodal-multi-tier annotation. The implementation is based on standards such as XQuery, XML, the Termbase eXchange language TBX, the interface is web-based and portable to other systems. The underlying database is the XML server Tamino.

Heike Zinsmeister / Margit Breckle
Konstanz University / Vilnius Pedagogical University

ALeSKo – An annotated learner corpus

Learner corpora collect texts of second language learners in order to provide an empirical basis for qualitative and quantitative studies in second language acquisition (Granger 2002).

In this poster we present a learner corpus of written texts of Chinese L2 learners of German. Currently it comprises 43 argumentative texts with an average length of 316 tokens and an overall corpus size of 13,587 tokens. The data is complemented by German L1 texts from the FALKO-Vergleichskorpus (FALKO, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko).

The goal for creating the ALeSKo corpus is to investigate coherence in learner texts. Coherence manifests itself on various levels of a text: among others, this includes appropriate reference handling and a smooth flow from one sentence to the next. To this end, the data is annotated with referential and syntactic information.

The creation process and the corpus design are strongly influenced by the criteria of FALKO (Lüdeling et al. 2008). We adopt its multi-layered approach and annotate layers of part-of-speech, topological fields (cf. Doolitle 2008), coreference and we also mark deviations. In addition, Vorfeld elements are classified for their pragmatic function (Breckle/Zinsmeister, this conference). Furthermore, we integrate the results of an experiment on the use of referential expressions. In the experiment, referential contexts for referring expressions were identified with respect to the semantic dimensions of definiteness and specificity (cf. Ionin 2003). Against this background, the L2 learner use of articles (definite, indefinite and null) was investigated (Breckle/Zinsmeister, to appear).

In this poster, we present the different layers of annotation in ALeSKo. We will discuss the problems that arise in the data annotation, the quality of annotation – i.e. the inter-annotator agreement – and what conclusions can be drawn from this.

Breckle, Margit / Zinsmeister, Heike (to appear). „Zur lernersprachlichen Generierung referierender Ausdrücke in argumentativen Texten.“ In: Skiba, Dirk / Deming, Kong (eds) *Textmuster – schulisch, universitär, kontrastiv*. Frankfurt/Main: Peter Lang.

Breckle, Margit / Zinsmeister, Heike (this conference). “The Vorfeld in Second Language Acquisition of Chinese Learners of German.” Talk presented at the Workshop Information Structure in Language Acquisition (32nd DGfS).

Doolittle, Seanna (2008). *Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen-Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner*. Masters Thesis, Humboldt Universität Berlin.

Granger, Sylviane (2002). “A Bird’s-eye View of Learner Corpus Research.” In: Granger, Sylviane / Hung, Joseph / Petch-Tyson, Stephanie (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Ionin, Tania (2003). *Article semantics in Second Language Acquisition*. PhD Thesis, Massachusetts Institute of Technology.

Lüdeling, Anke / Doolittle, Seanna / Hirschmann, Hagen / Schmidt, Karin / Walter, Maik (2008). „Das Lernerkorpus Falko.” In: *Deutsch als Fremdsprache* 2(2008), 67-73.