

# Acquisition and Representation of lexical and grammatical data

Ulrich Heid, Christian Rohrer

Version February 3, 2005

## 1 Ziele des Programms, Fragestellungen/Objectives

### 1.1 Overview

In 2002, when we proposed the first research programme entitled “Lexical Acquisition and Representation”, in the framework of the Graduiertenkolleg “Linguistic Representations and their Interpretation”, our main focus was on the use of very large corpora, for lexical acquisition. A number of projects within the Graduiertenkolleg (e.g. Beate Dorow’s work on the British National Corpus) and outside (e.g. the DFG-funded projects TFB-32 and DLF, work on German and Dutch corpora in industry projects, etc.) have over the last few years provided basic methodological insight into the work with very large corpora, and alongside, tools have become available which support day-to-day work with these resources, such as the chunker by Kermes (2003) for German or a similar tool by Spranger (2002) and Spranger/Heid (2003) for Dutch.

Even though the analysis of very large corpora has not yet fully become a standard tool, we think that methodological aspects of such work should now be further developed:

- from corpus-based data acquisition to semi-automatic classification of data acquired from text,
- from lexical acquisition to corpus-based grammar development, and
- from work on individual areas of morphological, syntactic and semantic regularity also, in addition, to phenomena involving the interaction of these levels, and to subregularities.

Consequently, we intend to propose dissertation topics in these areas. These research topics not only pose problems at the level of acquisition, but obviously also at the level of lexical and/or grammatical representation: do we need multi-layered lexicons (the same way as multi-layered corpus annotation has come to be a favorite topic of advanced corpus linguistics)? How would those relationships in the lexicon be described and expressed formally which go beyond relations between readings? How can we adapt and port grammars and lexicons? And finally: how would different processing components interact when sentences are analyzed (e.g. a parser, a morphology system, a detailed relational lexicon, and possibly systems of lexical rules)? These questions lead us to a stronger emphasis on topics dealing with integrated components and/or with interfaces between modules implementing knowledge from different levels of linguistic description.

The considerations summarized above do not imply a complete reorientation of the programme proposed in 2002, but they are meant to further develop the more general goals stated there.

We thus propose the following topic areas (some of which allow for several dissertations to be realized):

- Acquisition of lexical data and context-based classification;
- Corpus-based grammar development;
- Interactions between grammar and lexicon – subregularities;
- Modelling relationships between phenomena from different levels of description: morphology, syntax, semantics, collocations.

## 1.2 Spezielle Fragstellungen/Specific objectives

**Acquisition of lexical data and context-based classification:** Tools for the extraction of lexical data from corpora should keep track of the context of the words they are supposed to extract.

Often, the context allows us to derive conditions for certain phenomena: for example, the verb *berechnen* ('calculate') may take a *daß*-clause. But it preferentially (and almost exclusively) does so when it is in a tense of the past. Similarly, many noun + adjective collocations have preferences for the singular or the plural; certain subcategorization phenomena have restrictions with respect to the lexical items that act as complements of a given predicate.

Such phenomena need to be analyzed and described: we would welcome work which deals both with refinements of corpus-based acquisition and with more detailed descriptive schemes, e.g. for subcategorization and for the morphosyntactic and distributional conditions under which certain subcategorization frames are observed. The resulting lexicons are useful for applications of both parsing and NL generation, and for lexicography at large. A separate study could also be devoted to the use (and usefulness) of ambiguous text material for lexical acquisition. Finally, thesis work could deal with the question of grouping contexts according to syntactic similarity. Similar contexts essentially illustrate the same phenomenon, and should be grouped, even if they differ slightly.

**Corpus-based grammar development:** Methods and tools for corpus exploration can be adapted in such a way as to allow for the extraction of structure-related, mostly (or partly, see below) lexeme-independent data. These tools would allow us to quantify the occurrence of certain grammatical constructions, and to get a clearer picture of the relationship between the coverage of a grammar and the qualitative and quantitative inventory of constructions of the targeted fragment. As a matter of fact, not much is known about the frequency of grammatical constructions. In particular, as in the lexicon, also grammatical constructions seem to be distributed according to Zipf's law: we expect a large number of rare events (cf. Baayen 2001, LNRE property). Along with the phenomenon of productivity, this raises the question whether and how it is possible to provide a complete grammar.

We expect work in this field to focus on grammar development for written corpora, on grammar evaluation (e.g. via test suites), and on tools for profiling texts in terms of grammatical phenomena contained. From there, we expect contributions to a general methodology of building large-scale grammars and of increasing the coverage of a grammar. On the technical side, the extension of grammar coverage also has an impact: it may lead to more ambiguities and/or to increased processing time. An additional important task is thus to do research into methods of modifying unification grammars like LFG without increasing processing time (or only minimally), and without leading to overgeneration.

In addition, we would welcome research on the porting of grammars across closely related languages; or, diachronically, across different historical states. Taking the existing German LFG grammar, experiments could concern older stages of German. Porting of a Dutch grammar (as we have it for recursive chunking) to Afrikaans could also be considered. To carry out the porting, a morphological analyser of the other language (or language stage) would need to be prepared, and a lexicon provided. The objective of the work is then to identify rules which are shared, and rules of the source grammar which are not applicable; the latter case

will throw up phenomena which are specific to the other language, or in the historical case, which are no longer available (or not yet). Similarly, we are interested in a comparison of subcategorization lexicons, especially for older versions of German. As an LFG grammar for French is also available at IMS, similar research could also be carried out for French and Old French or 16th century French.

**Interactions between grammar and lexicon – subregularities:** Large scale grammars, such as the German grammar of IMS in the format and theory of LFG, have by now grown to a size where most lexeme-independent phenomena have been described. However, there seems to be a need, if we want to further develop Natural Language Understanding, to also cover semi-idiomatic phenomena, or smaller groups of expressions and/or grammatical structures which deviate partly from the general rules.

An example is the preference of German adjectives like “*offen*” (open), “*unklar*” (unclear), “*ungewiß*” (not sure), etc. to come with topicalized *ob*- or *wh*-clauses (*ob er teilnimmt, ist offen*). Even if this behaviour can be explained in semantic terms, it is necessary (or at least useful) to include the preference in a linguistic description, be it in the grammar or in the lexicon. Such phenomena could be related to the notion of “construction”, as it is used in Fillmore/Kay’s Construction Grammar.

More generally, large coverage grammars must include rules that are triggered lexically, i.e. that are only applicable with lexemes from a subset of the word class in question. Some such constructions deviate from the general rules: adjectives like *schade*, *klar*, *seltsam* (“a pity”, “obvious”, “strange”) are frequently used in an elliptical construction with a *daß*-clause: *schade, daß er nicht kommt*. This construction needs to be described, and it must be restricted to a lexical class; this restriction is also motivated with a view to the methodology of grammar engineering in Unification Grammar; if the rule is not restricted lexically, it will lead to overgeneration and to increased processing time.

We expect work in this field to uncover more subregularity phenomena (e.g. from the field of collocations, of multiword function words, of adverbials, etc.), to provide tools for their acquisition and – most importantly – ways to integrate their description into a formal grammar, such as LFG.

**Relationships between phenomena from different levels of description:** This topic is more concerned with the structure of lexical representation than with acquisition. If there are many relations between phenomena from different levels of linguistic description, a question is how these interrelated facts can be modeled

and how a lexical resource could be represented which kept track of as many interrelationships as necessary (in a database, in XML schemas, etc.). Examples of such interrelated facts concern, among others, the subcategorization of verbs and that of nouns, which are morphologically derived from the verbs. Or networks of collocations, or a dictionary supporting both semasiological and onomasiological access, etc. We view the dictionary here as a resource serving both NLP software and human users.

We expect work in this field to start from a relevant set of phenomena, for which a relational description would be given. Depending on the lexicographic interests of candidates, more or less work could be invested in the user-oriented aspect of an electronic dictionary which would make as many types of information available in a networked fashion as possible. An implementation of the model along with examples would be welcome.

We are particularly interested in relations between morphological and syntactic properties of words.

In addition, we are interested in the cooperation of morphological and syntactic processing in sentence analysis. For example, the analysis of coordinated constituents involving truncated compounds (*die Anfragen des Sozial- und des Wirtschaftsministeriums*) requires a thorough interaction between a morphology component and a syntactic grammar.

## **2 Stand der Wissenschaft und eigene Vorarbeiten**

In the following, we present a brief overview of the state of the art and of our own work in the field, for each of the topics mentioned in the previous section. The structure of the present section is thus by topics, and for each topic we first describe the state of the art in general, then point to our own work.

### **2.1 Acquisition of lexical data and corpus-based classification**

The field of lexical acquisition from corpora has considerably evolved over the last years. Obviously, the two main paradigms, symbolic and statistical, continue to be valid, however with an increasing amount of work being hybrid or otherwise combining the two aspects.

Nevertheless, not much new work has been proposed over the last years in the field of the acquisition of subcategorisation frames from text. As discussed in

Spranger/Schiehlen (2004), two approaches are still predominant, one oriented towards precision, the other towards recall. Precision-oriented work tends to focus on the correctness of the assigned subcategorisation frames, with typically quite large numbers of different subcategorisation frames distinguished. Recall-oriented work on the other hand, as it is mainly being carried out in the English-speaking world, focuses on a small number of different subcategorisation frames, trying however to exploit an existing text as much as possible. Unfortunately, not much work has been done on the use of contextual factors to further subclassify the extracted material. The only study which we are aware of, which combines collocational preferences and subcategorisation, is the dissertation by Klotz (2000).

Work at IMS: At IMS, Julia Ritz is preparing a diploma thesis on context based classification of collocations, with focus on their morpho-syntactic properties; a first outline of the motivation and tools to be used has been prepared (Heid/Ritz 2005). The general need for a detailed classification of collocations in terms of their morphosyntactic preferences has been stated in Evert/Heid/Spranger (2004) and in Heid (2005a); extraction technology keeping track of morpho-syntactic properties has been presented in Evert/Heid/Spranger (2004), along with statistical procedures for identifying the preferences, published by Evert (2004). For the field of subcategorisation extraction, less work has been done over the last years, but we expect the machinery described in Heid/Ritz (2005) to be applicable to subcategorisation extraction the same way as to collocation extraction. Work by Spranger (2004) aims at a multi-parametric analysis of subcategorisation data. Spranger's dissertation will also be partly devoted to this topic.

## **2.2 Corpus-based grammar development**

Most grammar development work proceeds either from a formal grammatical theory or, at least, from a definition of context-free grammatical rules. When a rule system is conceived, it is compared to data in corpora and then, if necessary, adapted to the corpora in question.

We are less aware of procedures which are strictly corpus-driven, in the sense that the analysis of corpus data, qualitative and quantitative, lead to a prioritisation of the grammar development steps, or even in the sense that the grammar contains rules covering phenomena which are not considered to be fully acceptable as elements of a prescriptive grammar, but which are nevertheless frequent in texts. Examples are coordinated structures where the first conjunct has a genitive and the second one a dative (“wegen des Vorschlags und dem Gegenvorschlag”). Even though in unification grammars, constraint relaxation can be used to deal with such “sloppy uses”, it is not clear how relaxed constraints can be integrated into

large formal grammars without negative impact on their computational properties (processing time, complexity, number of analyses).

Moreover, there is little work on the corpus-based preparation of test suites for grammars. Most test suite building work is part of large-scale testing of translation systems (e.g. as part of the development of commercial products).

Finally, what we think is also still lacking is a close relationship between tools based on low-level annotation or regular expressions on the one hand and formal grammars on the other hand. The first ones could however relatively easily provide qualitative and quantitative candidate data for certain types of phenomena the deep analysis of which would have to be left to a formal grammar. To some extent, work on the integration of deep and shallow processing as it has been carried out recently in the German project *Whiteboard*, goes into this direction.

Work on less frequent grammatical constructions and on subregularities in grammar and lexicon is increasingly seen to be important, because more and more evaluation work on corpus data is carried out; evaluations focusing on both precision and recall throw up phenomena where grammars fail. An important part of these are less frequent, idiomatic, or constructional phenomena.

Research on grammar porting has been carried out, in the framework of LFG, by Xerox, where the approach has been used to create a Korean grammar from an existing Japanese one (cf. Kim et al. 2003)

Work at IMS: Over the last years, work at IMS has involved corpus-based grammar development. The chunker designed by Kermes (2003) makes use of low-level annotation and pattern matching, to construct recursive annotations along the lines of a standard grammatical description inspired by Xbar Theory. A similar philosophy underlies the work by Klatt (2004) at the Artificial Intelligence Institute of Universität Stuttgart (partly supervised by U. Heid), even though Klatt's implementation is different. The chunking approach devised by Kermes has been applied to Dutch by Spranger (2002) and is being further developed in Spranger's work towards a PhD thesis.

Since November 2004, IMS is carrying out the DFG-funded project DLFG (*Disambiguierte LFG-Grammatik des Deutschen*, Antragsteller: Christian Rohrer). Part of this project is devoted to questions of corpus-based grammar development, including the corpus-based creation of test suites and procedures to verify the coverage of a grammar. We expect a close interaction (during the period until end of October 2006) of PhD work in the Graduiertenkolleg with the DLFG project.

### **2.3 Interactions between grammar and lexicon – subregularities**

The interaction between grammatical and lexical phenomena has gained recently more interest in linguistics. For example, a section of the annual congress of the Gesellschaft für angewandte Linguistik, GAL, in 2003, was devoted to problems of the interaction between grammar and lexicon, with emphasis on corpus-based approaches (see the forthcoming Lenz/Schierholz, Eds. (2005)). Another angle from which interaction phenomena have been approached, is that of the idiomaticity of certain constructions. Work in this field has mainly been done in Construction Grammar (for example Goldberg (1995), or more recently Fried/Östmann (2003)). Klotz (2000), which falls outside the paradigm of Construction Grammar, also is evidence of the interest in the interrelationship between grammar and lexicon.

Work at IMS: At IMS, Heid/Kermes (2002), have concentrated on preferences in the subcategorisation of adjectives (e.g. the distributional preferences of “offen”, “unklar”, “ungewiss” discussed above, in section 1.2). Interestingly, these adjectives not only display the distributional preference of having their subject clause topicalised, but they also have a marked tendency to co-occur with the verb “bleiben” (remain), the frequency of which is considerably higher than with other adjectives. The impression, as far as these data are concerned, is that the semantic specificity of the class of adjectives considered leads to distributional as well as collocational specificities. In this sense, these adjectives display a subregularity. Similar work has been done by Zinsmeister/Heid (2003) on adverbs which can be used predicatively (of the type “er ist unterwegs”, “alles ist anders”). Similarly, certain adverbs from a rather closed class can act as an oblique complement of manner, which in certain cases is even obligatory: *er verfährt so/ebenso/anders*. Subclasses of such adverbs need to be identified and described, and the description has to be integrated with the existing LFG grammar.

Work on such subregularities is an explicit element of the ongoing DLFG project at IMS.

### **2.4 Relationships between phenomena from different levels of description**

Within recent electronic dictionaries, some attempt is being made to model interrelationships between linguistic phenomena from different levels of description. A prominent example of such a dictionary is FrameNet, where syntactic,

morpho-syntactic and semantic descriptions are deliberately parallelised and interrelated. Furthermore, there are implicit interrelationships, by the fact that words are grouped in frames. However, morphologically related words (e.g. verbs and their nominalisations) are explicitly related in FrameNet. Such relationships within word families are however systematically included in the NomLex dictionary, a subcategorisation lexicon for nominalisations of verbs. This dictionary also relates individual subcategorised complements of nouns with the complements of the verbs derivationally related with the noun in question. A combination of the relationships present in FrameNet with those given in NomLex would be ideal.

In pedagogical lexicography, less detailed descriptive work has been done, however the intention is the same: to explicitly relate semantically close items, items belonging to the same morphological family, as well as collocations around base items. Both the electronic dictionary DAFLES and the electronic dictionary ELDIT are typical examples of this attempt.

The introduction of different views on lexical items, most prominently semasiological and onomasiological, is an objective of many electronic dictionaries, but it has so far mainly been realised in separate portions of electronic dictionaries (see e.g. ELDIT). DeSchryver (2005) claims that the XML-based representations underlying the dictionary production tool TshwaneLex would in principle allow to annotate lexical material onomasiologically, along with its regular semasiological description, so as to make the dictionary queriable according to both views. He also points towards possibilities of relating a systematic grammatical view of lexical items with a particularised, semasiological view.

Work at IMS: At IMS, the work on IMSLex, as described in detail by Fitschen (2004) has led to the inclusion of a relational component in the XML DTD of IMSLex. Fitschen uses this device so far only for cross references in a morphology lexicon, but it could also be used more widely. Heid will be doing experiments on more sophisticated approaches to lexical modelling in the framework of a project on architectures for electronic dictionaries, coordinated by Rufus H. Gouws, of Stellenbosch University, scheduled to start in March 2005 and to last for two years.

In the field of the interaction between morphology and syntax, as well as morphology and collocations, some work has already been done at IMS. Aldinger's ongoing dissertation, in the framework of the Graduiertenkolleg, is devoted to a morphosyntactic, syntactic and semantic analysis of nominalisations in *-ung* leading to the formulation of rules relating syntactic and semantic properties of verbs with those of their morphologically related nouns. Aldinger (2002) concentrated on the subcategorization behaviour of German particle verbs, as compared with that of their base verbs. In the field of collocations, Zinsmeister/Heid (2004) have analysed the collocational behaviour of noun compounds, in relation to that of

their heads. More work in this field needs to be done, but the study seems to throw up collocation preferences as an indicator of the degree to which compounds are lexicalised. Even though Zinsmeister/Heid (2004) more concentrates on the acquisition of such data from corpora, the material provided by this study could serve as a starting point for lexical modeling work.

### 3 Topics/Themen

In the following, we list tentative topics for dissertations.

- Area 1: Acquisition of lexical data and context-based classification
  - Syntactic subcategorization and morphosyntactic preferences (verbs, adjectives)
  - Morphosyntactic preferences in collocations: e.g.
    - \* active/passive, tense, number etc. in noun+verb-collocations;
    - \* attributive/predicative use of adjectives in noun+adjective collocations
  - Extracting linguistic evidence from ambiguous data.
  - Grouping syntactically similar contexts, to provide structured evidence of linguistic data.
- Area 2: Corpus-based grammar development
  - Measuring the coverage of a grammar (corpus tests, test suites, qualitative and quantitative aspects);
  - Corpus-based grammar writing and tuning (methodology; adaptation of grammars to texts, etc.).
  - Porting of a German grammar to older versions of German; porting of a French grammar to 16th century French or to Old French; identification of shared as opposed to individual phenomena in grammar and subcategorization lexicon.
- Area 3: Interactions between grammar and lexicon – subregularities
  - Finding and describing (semantically motivated) subregularities; modelling within a formal grammar framework (preferably LFG).

- Area 4: Relationships between phenomena from different levels of description
  - Data Models for network-like electronic dictionaries: implementing and using a relational dictionary (types of relations, dictionary user groups and usage situations, formal modelling, user interfaces, implementation). Example: The syntax of derivationally related words (verbs and their nominalizations; verbs and adjectives in *-bar, lich*, etc.). A model for an electronic dictionary indicating the subcategorization of derivationally related words.
  - Collocations of compounds: evidence for lexicalization? (collocations of compounds vs. collocations of the heads of these compounds: shared vs. non-shared collocations: Rules for collocation inheritance). Similarly: collocations with derivationally related words (e.g. *drink heavily, heavy drinking, heavy drinker*).
  - The cooperation of morphological analysis and syntactic analysis in a large coverage NLP grammar (phenomena involving both kinds of tools, ways of combining the tools, new approaches).

## 4 Verzahnung innerhalb des Kollegs/Links to other parts of the Graduiertenkolleg

The topics proposed here are closely related with the syntax oriented proposals in the Graduiertenkolleg. The focus is here on corpus based methods, which is complementary with respect to the other, more theoretical proposals. We would encourage close links between the theoretical aspects followed in the Linguistics department and our applied and NLP-oriented proposals.

There is also a close connection with work proposed by A. Stein, on collocations. Theses on this topic could be co-supervised, in particular for German, French, Italian or other Romance languages. This includes work proposed under the German/French cooperation in the Graduiertenkolleg, where also the interaction between morphology and syntax is an important aspect, including work on nominalisations and on the syntax of noun phrases.

### Zitierte Publikationen des IMS

- [HEID/RITZ 2005] Ulrich Heid, Julia Ritz: “Extracting collocations and their contexts from corpora”, ms., (Stuttgart: IMS), submitted for *COMPLEX-2005*, Budapest, June 2005

- [HEID 2005A] Ulrich Heid: “Specificités morpho-syntaxiques des constructions à verbes support: analyse sur base de corpus”, erscheint in: *Linguisticae Investigationes*, 2004, (Sonderband über Constructions à verbes support)
- [EVERT 2004] Stefan Evert: “The statistics of word cooccurrences – word pairs and collocations”, PhD Diss., Stuttgart, 2004, à paraître 2005
- [EVERT ET AL. 2004] Stefan Evert, Ulrich Heid and Kristina Spranger: “Identifying Morphosyntactic Preferences in Collocations”, in: *Proceedings of LREC-2004*, Lisboa, 2004, SS. 907 – 911.
- Klatt (2004)
- Fitschen (2004)
- Spranger (2004)
- [SCHIELEN/SPRANGER 2004] Schiehlen, Michael; Spranger, Kristina (2004) *Automatic Methods to Supplement Broad-Coverage Subcategorization Lexicons*. in Proceedings of the 4th International Conference of the Language Resources and Evaluation (LREC '04) pp. 29-32 Lissabon.
- [ZINSMEISTER/HEID 2004] Heike Zinsmeister, Ulrich Heid: “Collocations of complex nouns: Evidence for Lexicalization”, in: *Proceedings of KONVENS 2004*, 2004
- [KERMES 2003] Kermes, Hannah (2003) *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis IMS, University of Stuttgart Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.
- [ZINSMEISTER/HEID 2003A] Heike Zinsmeister, Ulrich Heid: “Identifying predicatively used adverbs by means of a statistical grammar model”, in: *Proceedings of Corpus Linguistics 2003*, (Lancaster), 2003
- [ZINSMEISTER/HEID 2003B] Heike Zinsmeister, Ulrich Heid: “Significant Triples: Adjective+Noun+Verb Combinations”, in: *Proceedings of Complex 2003*, Budapest, 2003.
- [HEID/KERMES 2002] Ulrich Heid, Hannah Kermes: “Providing Lexicographers with Corpus Evidence for fine-grained syntactic description: Adjectives taking subject and complement clauses”, erscheint in: *Proceedings of the Xth EURALEX International Congress*, (København: CST/KU) 2002

- [SPRANGER 2002] Kristina Spranger: *A lexically informed chunking analysis as a starting point for the extraction of linguistic information and terminology from Dutch text*, (Stuttgart: Univ. Stuttgart, IMS), 2002 [= Diploma Thesis], 115pp.
- [SPRANGER/HEID 2003] Kristina Spranger, Ulrich Heid: “A Dutch Chunker as a Basis for the Extraction of Linguistic Knowledge”, to appear in: *Proceedings of CLiN 2002 (Groningen, November 2002)*, (Amsterdam: Rodopi), 2003

### Andere zitierte Publikationen

- Baayen 200?
- DeSchryver (2005)
- Lenz/Schierholz, Eds. (2005)
- Fried/Östmann (2003)
- Klotz (2000)
- Goldberg (1995)
- Multilingual Grammar Development via Grammar Porting, [http://ling.uib.no/bscw/bscw.cgi/d3172/Multilingual Grammar Development via Grammar Porting](http://ling.uib.no/bscw/bscw.cgi/d3172/Multilingual_Grammar_Development_via_Grammar_Porting), Roger Kim, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, Hiroshi Masuichi, and Tomoko Ohkuma. 2003. Multilingual Grammar Development via Grammar Porting. In *ESSLI 2003 Workshop on Ideas and Strategies for Multilingual Grammar Development*. pp. 49-56. <http://www2.parc.com/istl/groups/nltt/papers/essli03kor.pdf>
- Porting Grammars between Typologically Similar Languages: Japanese to Korean, [http://ling.uib.no/bscw/bscw.cgi/d3176/Porting Grammars between Typologically Similar Languages a Japanese to Korean](http://ling.uib.no/bscw/bscw.cgi/d3176/Porting_Grammars_between_Typologically_Similar_Languages_a_Japanese_to_Korean), Roger Kim, Mary Dalrymple, Ronald M. Kaplan, and Tracy Holloway King. 2003. Porting Grammars between Typologically Similar Languages: Japanese to Korean. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC-17)*. <http://cslp.comp.nus.edu.sg/colips/conference/PACLIC17/index.htm>

## URLs von elektronischen Wörterbüchern

- <http://www.icsi.berkeley.edu/framenet/>
- <http://nlp.cs.nyu.edu/nomlex/>
- <http://www.kuleuven.ac.be/dafles/acces.php?id=null>
- <http://dev.eurac.edu:8081/MakeEldit1/Eldit.html>
- <http://www.ims.uni-stuttgart.de/projekte/TFB/papers/lrec04-lv.pdf>

## Andere relevante Publikationen der Antragsteller

- [HEID 2005B] Ulrich Heid: “Corpusbasierte Gewinnung von Daten zur Interaktion von Lexik und Grammatik: Kollokation – Distribution – Valenz”, erscheint in: Friedrich Lenz, Stefan Schierholz (Eds.): *Corpuslinguistik in Lexik und Grammatik*, (Tübingen: Stauffenburg) 2005
- [HEID 2005C] Ulrich Heid: “Corpus based lexicography”, soll erscheinen in: Anke Lüdeling et al. (Hg.): *Corpus Linguistics. An international handbook* (Berlin: Mouton de Gruyter) 2005
- [HEID 2005D] Ulrich Heid: “Computergestützte Phraseologie II”, soll erscheinen in: Harald Burger et al. (Hg.): *Phraseologie. Ein internationales Handbuch*, (Berlin: Mouton de Gruyter) 2005
- [HEID 2004] Ulrich Heid: “On the presentation of collocations in monolingual dictionaries”, in: *Proceedings of the eleventh EURALEX International Congress*, Vol. II, SS. 729 – 738, (Lorient: UBS) 2004
- [HEID 2002] Ulrich Heid: “La mise à jour semi-automatique de dictionnaires: une application de l’acquisition lexicale et de la méta-lexicographie”, in: *Revue Française de Linguistique Appliquée* (Amsterdam: De Werelt) 2002, SS. 53 – 66
- [HEID ET AL. 2004B] Ulrich Heid, Bettina Säuberlich, Esther Debus-Gregor, Werner Scholze-Stubenrecht: “Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition”, in: *Proceedings of LREC-2004*, Lisboa, 2004, SS. 419 – 423.

- [HEID ET AL. 2004C] Ulrich Heid, Stefan Evert and Bettina S"auberlich, Esther Debus-Gregor, Werner Scholze-Stubenrecht: "Supporting corpus-based dictionary updating", to appear in: *Proceedings of the eleventh EURALEX International Congress*, Vol. I, SS. 255 – 264, (Lorient: UBS) 2004
- [HEID/SPRANGER 2003] Ulrich Heid/Kristina Spranger: "Extracting terminologically relevant contexts from chunked corpora" Knowledge", in: *Conférence TIA-2003*, (Strasbourg), 2003
- [KERMES/HEID 2003] Hannah Kermes/Ulrich Heid: "Using chunked corpora for the acquisition of collocations and idiomatic expressions", in: *Proceedings of COMPLEX-2003*, (Budapest) 2003
- [DE SCHRYVER 2004] Gilles-Maurice de Schryver: *Concepts and Tools for lexicography in the electronic age, a case study of dictionary compilation in South Africa*, (Gent: RUG, African Languages and Cultures), 2004, ms. [=PhD. diss.], 449 pp.