New Chances for Deep Linguistic Processing

Hans USZKOREIT
Language Technology Lab
German Research Center for Artificial
Intelligence and Saarland University
D-66123 Saarbruecken
uszkoreit@dfki.de

Abstract

Recent developments in deep linguistic processing give rise to renewed optimism concerning the practical applicability of advanced grammatical analysis. For HPSG parsing a breakthrough in efficiency was achieved through a new form of international collaboration that lead to improved combinations of methods in a rather systematic way. Yet efficiency is just one of the obstacles to the utilization of deep processing in real life applications. A novel approach to the exploitation of deep parsing offers a strategy for compensating the deficiency in coverage and robustness. Through a combination of deep and shallow processing, the robustness of shallow processing is preserved for information extraction applications that exploit parsing methods of different depth. The underlying strategy and the realized architecture also provide a basis for distributed collective research on novel combinations of processing methods including applications of hyper learning.

Introduction

The grammatical analysis with linguistically designed grammars has always been a central topic of investigation in theoretical computational linguistics. However, deep grammatical analysis has so far played a negligible role in the development of language technology applications. Deep parsers have been lacking the efficiency and robustness required for real life applications. Building up coverage has been slow and costly. The few linguistic grammars that truly exhibit large coverage have caused a constant battle with extensive ambiguity.

In my talk I will present recent developments that provide new challenges and opportunities for linguistic methods. Increased efficiency of deep parsing and the embedding of a selective deep analysis into a robust shallow regime for information extraction offer ways to employ deep parsing in an environment where it can improve results without ruining robustness. Such a selective utilization was realized by our integration of HPSG parsing into a hybrid IE system, comprised of statistical and FST components. I will argue that such an approach provides a promising direction for gradual, measurable controlled progress in the development of deep grammars and lexicons.

The most serious shortcoming of today's language technology is the lack of methods that get at the real contents of text and speech, systems approximating language understanding. Therefore, the central and most ambitious problem in computational linguistics has always been the realization of deep linguistic processing involving an accurate mapping between written and spoken utterances on the one side and useful semantic representations on the other.

Modern linguistics has been able to provide theories and formalisms for the specification of grammars that express this mapping in a declarative and transparent way. Computational linguistics has contributed elaborate platforms and tools for grammar development. A few large-scale grammars have been designed exhibiting sufficient accuracy and coverage for real application tasks. However, these encouraging developments were seriously hampered by a lack of methods for language analysis that fulfill the minimal requirements in efficiency, robustness, and specificity. This simply means that all systems working with these grammars have been too slow and too brittle for real applications.

Furthermore, they have not been able to manage the vast ambiguity in natural language, i.e., they could not select among large numbers of linguistically correct analyses.

Yet the most immediate problem has been time and space efficiency. If an NLP system cannot process everyday sentences in a reasonable amount of time on a normal PC, it is not suited for most applications. Moreover, there was no chance to improve coverage and solve the issues of robustness and specificity if researchers had to wait for hours for a sentence to process. The performance problem was so severe that many promising lines of research ended without yielding any applicable results. The lack of efficiency became a major obstacle for several large endeavors involving extensive grammar development such as the IBM LILOG project with HPSG parsing on the LILOG STUF Parser in Prolog (Herzog and Rollinger, 1991), the Pargram conducted by XEROX with LFG parsing on the old Interlisp XLE platform (Butt et al., 1999), and the EU project LS-GRAM with HPSG parsing on the Prolog ALEP platform (Schmidt et al., 1996).

The situation seemed hopeless since all laboriously achieved gains in efficiency were almost immediately offset by efficiency losses due to increases in coverage or sophistication of the grammars.

1 The Battle for Efficiency

When Verbmobil (Wahlster, 2000), the largest research project ever conducted in speech technology, adopted deep linguistic processing on the basis of HPSG as one of the central methods for speech analysis in real-time translation of spoken face-to-face dialogues, this decision faced considerable criticism both from inside and outside the consortium. Why should the slowest and most complex processing method be employed in a system that strives for real-time processing? The decision could only be maintained because in the hybrid Verbmobil architecture, deep processing was just one of several processing methods and could therefore always be preempted by an analysis from a faster processing module. We will return to this point.

During the first phase of the project from 1993-1996, a team at IBM Germany in Heidelberg had been responsible for deep parsing. They tried hard to overcome the efficiency problem by combining statistical language modeling with their HPSG parsing scheme. In 1996 at the end of Phase One of Verbmobil, we were still far from getting even close to the performance requirements for the final Verbmobil prototype. When my lab was entrusted with the responsibility for deep linguistic analysis in the second halftime, it was not clear whether we would be able to deliver a component that would not always time out against the faster shallow processing modules. In the beginning we used the existing parser of our HPSG development platform PAGE that had been implemented in the Project DISCO (Uszkoreit et al., 1994). The HPSG grammar writers at Stanford University and DFKI had already worked with the PAGE development system for several years.

Interestingly, it was the immense pressure for efficiency in this speech application project that caused two members of the consortium, DFKI LT-Lab and CSLI at Stanford University, to join forces in developing completely new strategies for performance research in deep linguistic processing. Also included in the collaboration was the project PERFORM that I am conducting at Saarland University.

PERFORM contributed most of the methodology and evaluation technology. The methodological basis of the effort was the systematic, sophisticated and very detailed measurement of all relevant performance data for each version of a parser. For each parser and each parsed sentence a database record was created containing all data on numbers and sizes of successful and unsuccessful, complete and partial results, and on overall time and space consumption. The controlled utilization of the same grammars and the same test corpora was a precondition for obtaining comparable results.

The test corpora had to be annotated by the correct results and linked to previous performance data.

The novel engineering platform **tsdb**, developed by Stephan Oepen (Oepen, 2002), produces detailed diagnostic reports and complex multidimensional comparisons between alternative systems.

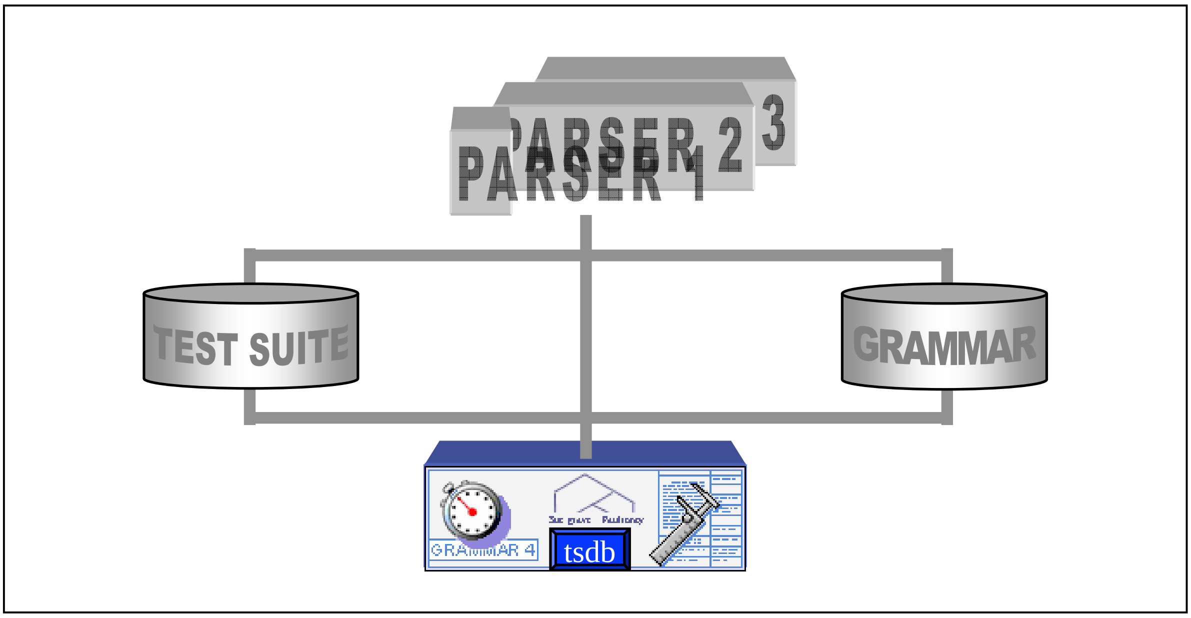

Figure 1 – The Parser Evaluation Setup

Later the Natural Language Processing Lab at Tokyo University joined the collaboration. The groups agreed on shared test suites and a reference grammar for comparative evaluation. The Stanford LinGO grammar for English developed at CSLI by Flickinger (Flickinger, 2000) and others was selected, although extensive tests were also run with large HPSG grammars for German (Müller and Kasper, 2000) and Japanese (Siegel, 2000).

At this time the LinGO grammar consisted of about 8000 types specified in 100.000 lines of source code. The lexicon contained some 6000 lexemes. An average feature structure in processing consists of more than 300 nodes.

The test suites, systematically composed of diagnostically relevant generic examples and representative examples for the Verbmobil domain, had a combined size of more than 3000 test items, in most cases sentences. They also included negative examples.

Through the tsdb platform, five systems were evaluated in this phase of the joint effort. The insights gained by the evaluation soon led to successful efforts in combining the most promising methods.

As the coordinator of the collaboration, I organized a workshop in Berlin in March 1999 where initial results and new goals were discussed and, together with Jun-Ichi Tsujii, a symposium at Schloß Dagstuhl in October of 1999 where the achievements were compared with progress in other research camps, e.g., parsing with LFG, TAG, and categorial grammar.

A number of new methods were developed by the three sites and tested in many combinations. In the end, it was a synthesis of methods reached by a true scientific and technological cross-fertilization process that brought about the breakthrough in the battle for efficiency. A first joint publication on such combination was (Kiefer et al., 1999).

All participating parsers improved in efficiency. Contrary to the predictions of our partners from the speech community, the Verbmobil HPSG parser did not end up at the tail end of the performance scale. The average distribution of total run-time among the Verbmobil processing stages demonstrates that deep linguistic processing had successfully overcome the initial efficiency problems: speech recognition 38%, prosody processing 17%, syntactic/semantic analysis by HPSG 25%, semantic interpretation and dialogue processing 14%, transfer 3% und generation 3% (Wahlster, 1997). It should be pointed out here that the short processing time for the transfer stage became possible because the results of semantic processing permit a reduction of transfer to mere lookup in most cases. The role of HPSG processing in Verbmobil and the reasons behind the choice of the grammatical framework are discussed in (Uszkoreit, Flickinger, Kasper & Sag 2000). The Verbmobil parser is described in (Kiefer et al., 2000).

The fastest parser that came out of the joint effort is the system PET developed by Ulrich Callmeier (Callmeier, 2000) in the project PERFORM. PET could not be integrated anymore in the final Verbmobil demonstrator. It is freely available under an open source for academic and commercial use. The preferred grammar engineering platform is Anne Copestake's completely reworked LKB system, which is also freely available under an open source license (Copestake, 2002).

The methods that were combined in the most successful systems involve all components of parsing starting with optimizations of the grammar and changes to the formalism but ranging all the way to improved algorithms for feature unification and a better representation of parse forests. The list of the techniques that contributed most profoundly to the efficiency gains include:

- the elimination of general disjunction in feature structures;
- methods for optimized structure sharing;
- improved rule filtering techniques;
- quick check computation of relevant parts of feature structures before full unification;
- efficient subsumption checking for ambiguity packing.

The overall run-time efficiency gain accomplished after three years was a factor of 2000 (Oepen et al., 2000). Space consumption was also reduced by more than an order of magnitude. Time measurements were performed on comparable hardware. This means that the gains achieved on the software side was also complemented by well-known progress in hardware efficiency. Sentences can now be analyzed in fractions of the time needed for real-time speech applications. Larger texts can be analyzed in a few seconds. The fastest parser can now be run on a standard PC. Thus HPSG parsing now meets the speed and working memory requirements for a wide range of applications. Papers on major results of the collaboration are compiled in (Flickinger et al., 2000; Oepen et al., 2002).

This breakthrough lead to increased interest in HPSG processing in several areas of research and in industry. Many theoreticians and practitioners of grammar have expressed their interest in using the software for grammar development. The first industrial applications have been developed.

2 Robustness and Coverage

2.1 The Verbmobil Approach

Now we turn to the problem of robustness and coverage. Even if efficiency had been the main stumbling block on the road to real-life natural language processing applications that work with highly accurate deep linguistic grammars, the real and potentially much more challenging problem is the design of systems able to properly deal with the rich variation of human language input in realistic application situations. We want systems to master the proper subset of a natural language that seems to be needed for a specific application. Moreover, we want a system that does not fall over if faced with some input outside this imagined subset.

It has always been tricky to distinguish issues of missing coverage from those connected to a lack of robustness. If we consider the definition of robustness from the IEEE standard glossary of software engineering terminology (IEEE, 1990) then robustness is “the degree to which a system or component can function correctly in the presence of invalid input or stressful environmental conditions.“ At first glance this definition seems to apply quite straightforwardly to language processing applications. Invalid input may be ungrammatical utterances or grammatical utterances outside the appropriate sub-language for the application. Stressful environmental conditions may result from background noise or distortions in the speech signal or from hundreds of queries simultaneously sent to a question-answering system.

However, the problem is how to define the appropriate sub-language. If we want to account in a spoken language input system for the wide variety of idiolects, dialects, accents and sociolects of speakers and if a written input system is supposed to properly deal with very large volumes of unseen texts produced by a large number of authors then the sub-language should be defined rather large. Such a demand poses a serious problem to most contemporary linguistically sophisticated competence grammars. It would require the grammar to analyze the union of many widely overlapping languages instead of relying on an additional mechanism mapping any sentence within the sub-language but outside some standard variant of the language into the most closely corresponding sentence within the standard language. The more permissive a large scale grammar becomes by increasing coverage within one variant of the language, by going beyond this variant or by even accounting for performance errors and distorted input, the more the system will be confronted with the problem of excessive local and global ambiguity. We will return to this problem later.

At this point we can state that the differentiation between problems of missing robustness and those of missing coverage with respect to some specific system depends on the specifications underlying its system design. If the implemented grammar itself is seen as the specification of the relevant sub-language, the difference becomes useless. All inputs that cannot be analyzed by this grammar automatically fall under the category of deficiencies in robustness.

In the project VERBMOBIL coverage was restricted along two dimensions. One dimension was the size of the lexicon. Since the performance of the speaker-independent recognition of continuous speech is strongly dependent on a restricted vocabulary, the size of the lexicon was first limited to 2500, later to 10000 words. The second dimension follows from the goal to realize a plausible application scenario for such a restricted vocabulary and to deeply model a domain for the integration of dialogue and knowledge processing with the analysis and generation of spoken utterances. Therefore the domain was restricted to dialogues on the scheduling of appointments. This restriction limited lexical ambiguity and excluded certain types of constructions.

In VERBMOBIL two main approaches were employed for realizing robustness. One method was based on deep grammatical processing with HPSG and utilized minimally recursive under-specified semantics. Whenever the HPSG parser could not produce an analysis covering the whole sentence, the resulting chart was sent to a processing component for robust semantic interpretation that attempted to exploit the intermediate results in the chart and knowledge about the domain and the sub-language to hypothesize at least a partial semantic representation of the input utterance (Pinkal et al., 2000).

The second approach is reflected in the fundamental design of VERBMOBIL. The architecture of the system is based on the concurrent application of several different processing methods and the final selection among alternative analyses and even translations. In case deep grammatical analysis and translation based on a semantic representation could not provide a solution, i.e. a translation in a target language (or failure to deliver a result before a preset timeout) a translation could be selected from one of the other three processing strands; translation based on dialogue analysis and dialogue acts, a statistical translation or a case-based translation. In this way translations could be produced for virtually all inputs including even extreme cases of ill-formed utterances. In about 74% of all cases the translations were correct or approximately correct. A scientifically challenging issue is the optimal selection among several results proposed by different methods or components. Within the time frame of the VERBMOBIL project very little research could be dedicated to this exciting theme. The derivation of stable confidence measures and the search for statistically reliable quality indicators could have become the objective of a follow-up project. In the realized and demonstrated VERBMOBIL system the result of deep analysis and translation was always preferred over the results of shallow processing if it could be provided.

To sum up the VERBMOBIL experience with respect to progress in deep grammatical processing, we can report breakthroughs in the areas of efficiency and robustness. Progress in run-time efficiency was achieved by the careful and systematic combination of several methods ranging from the optimization of the grammar via generic program optimization to a number of novel algorithmic solutions in parsing with unification grammars. The required robustness was achieved by combining deep processing methods with several shallow analysis and translation techniques that worked in parallel and almost independently of each other. In addition promising first results could be achieved in making deep grammatical processing more robust by exploiting the incomplete chart and in under-specified semantic representation for the construction of partial interpretations with the help of domain-dependent heuristics.

2.2 The WHITEBOARD Approach

In this section a new approach for utilizing deep grammatical processing in real-life applications is described. The approach is based on an integrated architecture for a variety of shallow and deep processing components.

The combination of deep and shallow methods for improving the performance of analysis is not a new idea. Several approaches have been proposed and demonstrated for augmenting a deep parser with shallow methods for preprocessing such as POS tagging and the analysis of complex names or other fixed multiword expressions. Another way of combining the virtues of deep and shallow methods is the employment of shallow analysis tools such as statistical phrase or chunk parsers for selecting among alternative readings.

The strategy I proposed as the basis of our WHITEBOARD project, differs from these solutions. Instead of letting shallow processing assist the deep parser, we let deep processing assist a shallow processing IE system.[1] This decision is in parts based on our experience with information extraction.

Let me start with the demands of the application. In IE recall and precision are constantly improved through the design of more sophisticated rule systems and through the application of statistical or symbolic machine learning techniques. However, in the detection of relations, IE systems are often confronted with texts in which the exact assignment of detected entities to the appropriate argument slots of hypothesized relations seems to require a deep grammatical analysis – frequently even with some semantic filtering. The combinatorics behind such constructions makes each type rather rare in data collections, too rare at least for contemporary learning approaches.

Assume IE in medical and pharmaceutical texts by which we try to find relevant relationships between medicines and medical conditions. A medicine may be indicated to treat a condition, however a condition may also be a side effect or a counter indication. Since both conditions and

[1] Actually this master-slave relation is complemented by several ways exploiting shallow processing for the deep parser.

side effects can occur very close to the reference of the medicine,

A similar problem arises in personnel news such as management succession reports where references to several persons can occur in close proximity to the reference of a management function. Consider the following pair of examples from German:

(i) Peter Mischke zufolge wurde Dietmar Hopp
Peter Mischke according was Dietmar Hopp

gebeten den Vertrieb zu übernehmen
asked the sales to take over.

According to Peter Mischke, Dietmar Hopp was asked to take over the sales department.

(ii) Peter Mischke wurde von Dietmar Hopp
Peter Mischke was by Dietmar Hopp

gebeten den Vertrieb zu übernehmen
asked the sales to take over.

Peter Mischke was asked by Dietmar Hopp to take over the sales department.

In order to correctly fill the slots, the system needs to recognize which of the two persons mentioned in these sentences is the successor in the position. To this end, the subjects of gebeten and übernehmen need to be determined. This cannot be done without an analysis taking into account the interaction of word order, control and passivization. I would claim that a pseudo-grammar which can assign the correct slot fillers in these cases will come so close to the functionality of a deep grammar that we might be better off if we employ an already existing deep grammar. Here we seem to back at the performance arguments because in typical IE applications thousands of documents need to be analyzed within seconds or minutes.

However, if a reasonably efficient deep parser is consulted by the IE system in just the cases where the IE system cannot decide among several slot assignments and if the IE system can accept or disregard the results of the deep parser, neither efficiency nor coverage or robustness will be severed by the deep parser. This is the avenue we proceed.

The next logical step is the consequent specialization of deep grammars with respect to certain IE domains and tasks. Our Verbmobil grammar has already been a mix of generic and specialized components, so we are now specializing the HPSG grammar for IE tasks.

The WHITEBOARD[2] architecture that has been implemented is very general. It supports the combinations of processing methods currently under investigation in the WHITEBOARD project but also countless others. The underlying idea is as simple as it is powerful: several shallow and deep processing components annotate an input text with XML markup. In theory the components could perform their annotation in any order and even be called several times. They can look at and exploit previously annotated markup. Many types of linguistic information cannot be merged into a single XML markup structure. Among them are identifiers for multi-part objects such as discontinuous constituents or co-references and annotations involving conflicting bracketings such as the markup of ambiguous readings or alternative structuring hypotheses. Therefore we employ a multitude of annotation layers. Hyperlinks and information on covered spans in the text connect the different layers. Some layers may introduce (additional) auxiliary layers that may contain representations that are not coded as XML markup. In this way the layer encoding the labeled bracketing produced by the HPSG parser can connect to an auxiliary layer containing feature structures that do not have to be merged into the annotated text. These auxiliary layers may also be used for storing analog data such as sounds or pictures. The core part of the architecture is the WHITEBOARD annotation machine (WHAM). The WHAM can be called by an application that needs to know about the

[2] We discovered that the term „Whiteboard architecture“ has been used before. In computational linguistics it was proposed at the 1994 Coling by Christian Boitet and Mark Seligman (Boitet and Seligmann, 1994) to refer to a powerful extension of blackboard architectures.

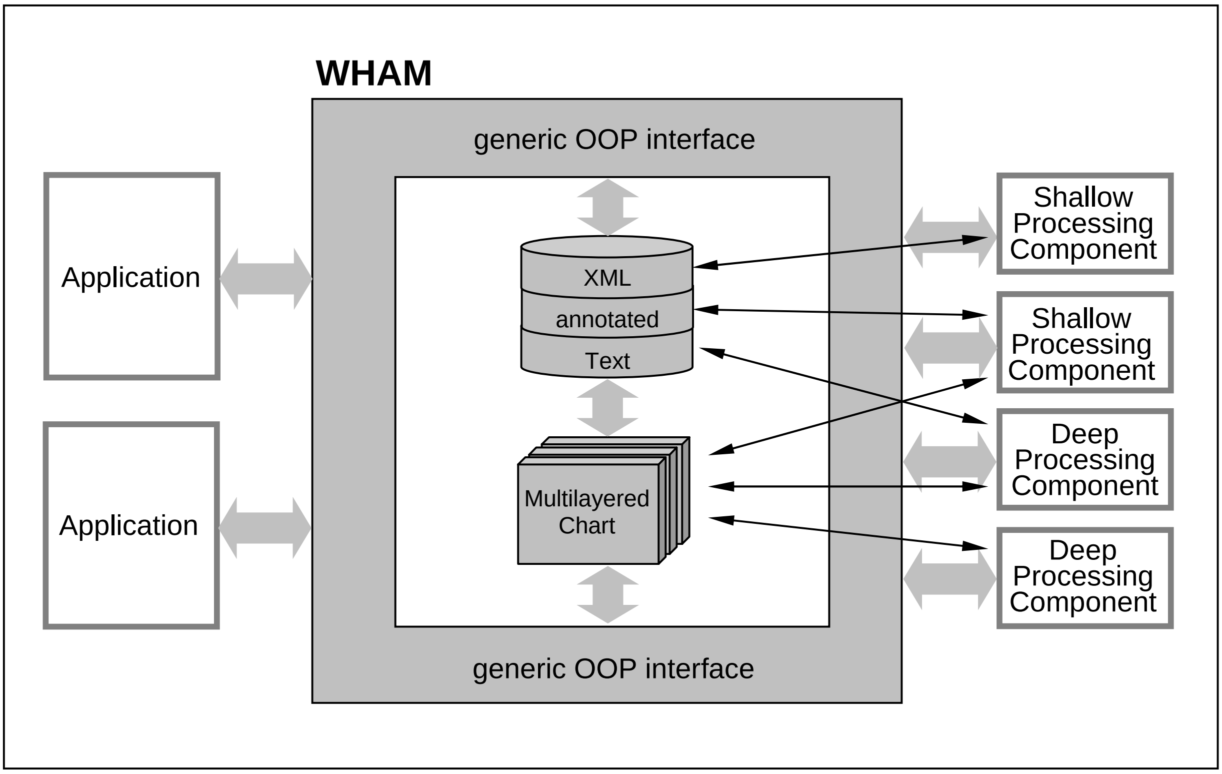
WHAM
generic OOP interface
Application
Application
XML
annotated
Text
Multilayered
Chart
Shallow Processing Component
Shallow Processing Component
Deep Processing Component
Deep Processing Component
generic OOP interface

Figure 2 – The WHITEBOARD Annotation Machine

WHAM's generic OOP interface and about the components that may be requested. The architecture of the WHAM is presented in Figure 2.

The application passes an input text to the WHAM together with specifications about the requested components, the sequence of their activation and the nature of the requested result. The WHAM interface can now call the specified components in the requested order. It consists of iterators that walk through the different annotation levels and reference and seek operators that jump to corresponding annotations on different levels. These can return, for instance, all part-of-speech label tokens of the current sentence or the next named entity of a certain type starting from the current token position or the next temporal prepositional phrase. Other components of the OOP are accessor methods that return the information encoded in the chart as well as general methods that enable navigation through the type system and the feature structures of the deep processing components. The WHAM works with two mechanisms for the representation of results, an external offline representation in which the input text is enriched with XML markup in an additive monotonic fashion and an internal online annotation chart with multiple levels in which the more abstract structural information (bracketing hypotheses even involving discontinuous constituents) and multi-dimensional characterization of complex objects such as syntactic and semantic feature structures are encoded in a uniform linguistically motivated form.

Each interface to a specific processing component is realized as a subclass of the generic WHAM interface. If a new component has to be integrated a new subclass has to be defined and its component-specific representations have to be specified by a new DTD for the XML markup or a set of transformation rules for the chart or both. Shallow processing components that have been integrated include our statistical part-of-speech tagger and phrase parser TNT by Thorsten Brants (Brants, 1999) that is based on cascaded hidden Markov models, the chunk parser CHUNKIE by

Wojciech Skut (Skut, 1999) that extends the functionality of TNT, and our shallow processing platform SPPC by Jakub Piskorski and Günter Neumann (Piskorski and Neumann, 2000), a rule-based system that employs cascades of weighted finite-state machines for tokenization, morphological analysis, part-of-speech filtering, named entity recognition and the detection of chunk, sub-clause and sentence boundaries. In the current setup, deep processing is represented by the HPSG parser PET. Ulrich Callmeier and Ulrich Schäfer extended the PET system to allow – instead of single words – multiple word input items that even may be overlapping and ambiguous forming word hypothesis graphs. Another extension permits the dynamic creation of atomic type symbols so that arbitrary symbols can be added to feature structures for flexible interfaces to external components such as morphology tokenization or named entity recognition.

In addition to the central shallow and deep processing components, a further knowledge source was added exploiting the same interface mechanisms to the WHAM. In order to obtain more fine-grained semantic information on the lexical level, especially sortal information, the German version of WordNet, GermaNet graphs was utilized. Compared to GermaNet the current HPSG lexicon is quite small. Its semantic classification is rather sophisticated in the Verbmobil domain but lacking fine-grained sortal information in most other subject areas. In order to draw benefits from the integration of the rich information source GermaNet, a mapping from the sortal categories of GermaNet synsets to the HPSG sort system had to be obtained. To this end a supervised learning algorithm was employed that was trained on those words annotated with semantic sorts in the HPSG lexicon and with synsets in GermaNet. The learning algorithm then computes such mappings for words that do not yet have corresponding sort information assigned in the HPSG lexicon. It does this by calculating a relevance measure for each possible association. The hypothesized mapping was evaluated in the domain of business news. From a corpus of business news 4664 nouns were extracted that were not represented in the HPSG lexicon. 2312 of these unknown words were contained in GermaNet. GermaNet assigned 2811 senses to these words. These word senses were then automatically mapped to HPSG sorts. An evaluation of these sort assignments by human judges yielded a rather promising result. In 76.52% of all assignments the mapping suggested by the algorithm i.e. with the highest relevance measure turned out to be correct. In 27% of all cases the correct sort had received the 2nd or 3rd highest relevance assignment.

3 Conclusion and Future Research

Although the number of centers conducting projects in deep grammatical processing has decreased and therefore also the visibility of this line of research, considerable progress has been achieved.

For time and space limitations and for other obvious reasons, I have concentrated here on developments in processing with HPSG. Other breakthroughs have recently been accomplished in parsing with Lexical Functional Grammar, Tree-Adjoining-Grammar, Categorial Grammar and Dependency Grammar. Many of those improvements are based on the combination of statistical methods and grammatical processing, on better and larger grammars as well as on clever and careful engineering. The current XLE system for LFG, for instance, exhibits an impressive efficiency and integrates a powerful machinery for dealing with massive ambiguity. (King et al., 2000)

One of the goals should be the systematic comparison of deep parsers across linguistic, formal and technological frameworks. For this purpose the evaluation against the Penn Treebank (Marcus et al., 1993) is the only game in town. Such evaluation in the spirit of Parseval that has revolutionized the field of parsing research is much more suited for shallow or medium depth parsers (such as PCFG system) than for parsers producing a sophisticated semantic representation. Also for evaluating the value of partial results with respect to applicability, the current evaluation setup is of limited value.

Nevertheless it is a great achievement of our colleagues from the LFG and TAG communities that they made their parsers ready for the evaluation against the WSJ corpus and were able to achieve impressive and encouraging results, e.g. (Riezler et al., 2002).

In a workshop at LREC 2002 called "Beyond Parseval" (Carroll et al., 2002), the participants representing several distinct schools in deep linguistic parsing almost unanimously acknowledged the shortcomings of the Penn Treebank annotation. In the final discussion of the worshop they agreed with a proposal by (Briscoe et al., 2002) showing an annotation scheme based on grammatical relations that is better suited for comparing results of different grammar frameworks as well as for robust partial parsing and multilingual annotation.

Although we can point at considerable progress in deep linguistic progressing, hard problems remain to be solved. Our international collaboration has entered into a new phase. The partnership has been broadened by including additional groups, among them Cambridge University, University of Sussex and University of Edinburgh. In the current phase of collaboration, methods for improving robustness, coverage and disambiguation constitute the main objectives of theoretical research. A focus lies on statistical methods for extending grammar and lexicon and for learning disambiguation preferences. The group at Saarland University has achieved progress in exploiting the NEGRA/TIGER treebank of German, the Stanford group is building the Redwoods Treebank (Oepen et al., 2002), a dedicated HPSG treebank.

In addition, the combination of deep and shallow processing for IE and other applications will be further pursued by the consortium. New results on robust parsing and partial interpretation with robust minimal recursion semantics open new application domains.

Some partners (DFKI, Cambridge University, University of Sussex, University of Trondheim and Stanford University) have obtained funding through a recently approved EU-sponsored project named Deep Thought, in which they will work together with three innovative industrial companies (CELI, Torino; Edify, Edinburgh, and XtraMind, Saarbruecken) on the practical exploitation of combined deep and shallow processing starting from the WHITEBOARD architecture.

The third area of collaboration is multilingual grammar development. Continuing a theme of our Verbmobil research, linguistic, formal and practical issues in sharing information among different grammars are investigated (Müller and Kasper, 2000; Siegel, 2000). Among the practical goals of this research are shorter development times for new grammars, reusability of semantic resources and uniform multilingual applications. Emily Bender (Bender et al., 2002) and others have already developed a so-called matrix grammar containing the shared components of the English, German and Japanese grammars that is used to get a warm-start on grammars for additional languages,

The work plan for the current phase of the collaboration can be viewed at:

http://www.coli.uni-sb.de/~hansu/Collaboration.html

4 Outlook

Only very few of us still expect to see one day the automatic understanding of natural language texts as a result of a sudden breakthrough in research. Certainly, I do not foresee

Instead, I would predict that information extraction will become the leading research paradigm in computational linguistics. In our discipline, this paradigm subsumes all applications that can recognize relevant types of information in human language texts. These can be topics, the most important sentences of a text, named entities, binary relations, event templates, or complex relational concepts. On one side of the scale we have text filtering, as the extraction of a category in a binary classification, on the other extreme, we have the extraction of complex relational objects, representing the meaning of an entire discourse.

In this way, information extraction spans the continuum between the most modest language technology applications and true language understanding.

If we want to convince the research community funding agencies and industrial clients of the value of sophisticated grammatical analysis, it will not suffice anymore to insist on the necessity of deep processing for reaching the ambitious long term goal. We will have to demonstrate that deep processing can contribute to progress on the long way of small and controlled steps that lies ahead of us. Thus the advancement of linguistic processing needs to become subject to similar measurement and evaluation as progress in shallow processing. This requires improved methods for combining and comparing alternative approaches and techniques. Common tasks in today's technology evaluations and shared data collections mark the beginning of a new era of massively collaborative research.

From all we learned during the past decades of research, human language is not less complex than the subjects of chemistry, genetics, or geology. Yet the organization of contemporary research in language and language processing stands in stark contradiction to this insight. Each individual center builds its own software systems for the processing of one or more languages. Sometimes systems and tools from other centers are adopted or re-implemented, but there is no infrastructure or common tasks that permits the comparison, exchange and combination of methods and systems. An international multi-site collaboration within one school of research is already very difficult to organize and maintain. In the future we need ways of comparing and combining results across theoretical frameworks and research communities.

Heterogeneous architectures such as the WHITEBOARD can provide means for combining methods and components. A new R&D paradigm of collective research can only be realized, however, if the necessary infrastructure is available. Without common corpora, annotation schemes and tasks, the envisaged combination of methods cannot be achieved. We need large multilingual corpora automatically annotated by a variety of processing tools. These corpora also constitute common tasks. For each participating language, several shallow and deep processing systems should enrich the text by layers of annotation :

- categorizers
- segmentizers/tokenizers
- statistical and rule-based POS taggers,
- morphological analysers
- full text indexing
- shallow chunk and phrase parsers
- wordnets and other thesauri
- information extraction systems NEE, relation extraction, template filling
- statistical parsers
- deep linguistic parsers such as LFG, HPSG, DG and CG parsers
- systems that determine temporal or discourse relations
- summarization systems

Some portions of the texts should be hand-corrected for obtaining measures of reliability and gold standards for evaluation. The multi-layered annotations can be encoded in XML. For maintaining a uniform correspondence between annotation and text spans and for allowing the extension to speech documents, an approach of annotation graphs (Broeder et al., 2000) will have to be exploited.

A number of the components utilized for annotation will be based on machine-learning techniques. The envisaged corpora could thus serve as an extremely valuable resource for higher-order learning, so-called hyper-learning.

The annotation of complete and partial semantic analyses poses a special challenge since compatibility between the results of deep processing and information extraction needs to be established. Robust minimal recursion semantics or other semantic formalisms accomodating underspecified and partial representations will be needed for encoding semantic information. For annotating the corpora by meta-information the OIL/DAML (Connolly et al., 2001) format may be utilized in order to exploit existing connections with the Semantic Web community and to improve the value of the corpus.

In my opinion, the question is not whether such resources will be created but rather when the work is going to start. A consortium of research centers from several European countries has already decided to initiate actions into the outlined direction.

Acknowledgements

The project Verbmobil was funded by the Bundesministerium für Bildung und Forschung under grant 01 IV 701. The project WHITEBOARD is funded by the Bundesministerium für Bildung und Forschung under grant 01 IW 002. The project Perform is funded by the Deutsche Forschungsgemeinschaft as part of the SFB 387 "Resource-Adaptive Cognitive Processes".

My special thanks go to all participants of our international collaboration, mainly to Tania Avgustinova (Saarland U.), Liviu Ciortuz (U. of York), Ann Copestake (Cambridge U.), Dan Flickinger (Stanford U.), Emily M. Bender (Stanford U.), Ulrich Callmeier (Saarland U.), John Carroll (U. of Essex), Ann Copestake (Cambridge University), Andreas Eisele (DFKI), Frederik Fouvry (Saarland U.), Anette Frank (DFKI), Bernd Kiefer (DFKI), Valia Kordoni (Saarland U.), Hans-Ulrich Krieger (DFKI), Chris Manning (Stanford U.), Takaki Makino (Tokyo U.), Rob Malouf (San Diego State U.), Mitsoishi Minoru (Tokyo U.), Takashi Ninomiya (Tokyo U.), Stephan Oepen (Stanford University), Detlef Prescher (DFKI), Ivan A. Sag (Stanford U.), Melanie Siegel (DFKI), Kenjiro Taura (Tokyo U.), Kentaro Torisawa (Tokyo U.), and Jun-Ichi Tsujii (Tokyo U.) – in alphabetical order.

I am especially grateful to Stephan Oepen, who created the engineering platform tsdb and served as the main driver of the collaboration as well as to the members of the DFKI project WHITEBOARD Berthold Crysmann, Anette Frank, Bernd Kiefer, Stephan Müller, Günter Neumann, Jakub Piskorski, Ulrich Schäfer, Melanie Siegel, Feiyu Xu, Markus Becker and Hans-Ulrich Krieger for their realization of the hybrid architecture.

References

E. Bender, D. Flickinger, and S. Oepen. 2002. The Grammar Matrix. An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In *Procedings of the 19th International Conference on Computational Linguistics*. Taipei .

S. Bird and M. Libermann. 1999. Annotation Graphs as a Framework for Multidimensional Linguistic Data Analysis. In *Proceedings of the Workshop Towards Standards and Tools for Discourse Tagging Association for Computational Linguistics*.

C. Boitet and M. Seligman. 1994. The "Whiteboard Architecture: A Way to Integrate Heterogeneous Components of NLP Systems. In *Proceedings of the 15th International Conference on Computational Linguistics*. (Vol.1). Kyoto.

T. Brants. 1999. Tagging and Parsing with Cascaded Markov Models - Automation of Corpus Annotation. Doctoral dissertation. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Vol. 6. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken.

T. Briscoe, J. Carroll, J. Graham, and A. Copestake. 2002. Relational Evaluation Schemes. In J. Carroll, A. Frank, D. Lin, D. Prescher, H. Uszkoreit, (eds.). *Beyond PARSEVAL* LREC-02, Las Palmas.

D. Broeder, H. Brugman, A. Russel, R. Skiba, and P. Wittenburg. 2000. Towards a Standard for Meta-Descriptions of Language Ressources. In *Proceedings of LREC 2000.*

M. Butt, S. Dipper, A. Frank, and T. King. 1999. Writing Large-scale Parallel Grammars for English, French, and German. In M. Butt, and T. King (eds.). *Proceedings of the LFG99 Conference*. CSLI On-Line Publications.

U. Callmeier. 2000. A Platform for Experimentation with Efficient (HPSG) Processing Techniques. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit (eds.). *Natural Language Engineering, Vol. 6 (1)* Special Issue on Efficient Processing with HPSG.

J. Carroll and S. Oepen. 2000. Performance Profi- ling for Parser Engineering. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit (eds.). *Natural Language Engineering, Vol. 6 (1)* Special Issue on Efficient Processing with HPSG.

J. Carroll, A. Frank, D. Lin, D. Prescher, and H. Uszkoreit (eds.). 2002. Beyond PARSEVAL - Towards Improved Evaluation Measures for Parsing Systems. Workshop at the 3rd International Conference on Language Resources and Evaluation LREC-02., Las Palmas.

D. Conolly, F. Van Harmelen, I. Horrocks, D. McGuiness, P. Patel-Schneider, and L. Stein. 2001. DAML + OIL Reference Description. In the web: http://www.w3.org/TR/daml+oil-reference

A. Copestake. 2002. Implementing Typed Feature Structure Grammars. CSLI Publications, Stanford.

B. Crysmann, A. Frank, B. Kiefer, S. Müller, G. Neumann, J. Piskorski, U. Schäfer, M. Siegel, H. Uszkoreit, F. Xu, M. Becker, and H. Krieger . 2002. An Integrated Architecture for Shallow and Deep Processing. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*. Philadelphia.

D. Flickinger. 2000. On Building a More Efficient Grammar by Exploiting Types. In Flickinger, D. and Oepen, S. and Tsujii, J. and Uszkoreit, H., editors, *Natural Language Engineering, Vol. 6 (1)* Special Issue on Efficient Processing with (HPSG), pp. 15-28.

D. Flickinger, S. Oepen, H. Uszkoreit, and J. Tsujii (eds.). 2000. Journal of Natural Language Engineering 6 (2000) 1. Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation. Cambridge University Press. Cambridge.

O. Herzog and C.-R. Rollinger (eds.). 1991. Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence. Springer. Berlin.

IEEE. 1990. IEEE Standard Glossary of Software Engineering Terminology, IEEE Std 610.12-1990, IEEE, Computer Soc., Dec. 10, 1990.

B. Kiefer, H. Krieger, J. Carroll, and R. Malouf. 1999. A Bag of Useful Techniques for Efficient and Robust Parsing. In *Proceedings of the 37th Meeting of the Association for Computational Linguistics*. College Park, MD.

B. Kiefer, H. Krieger, and M. Nederhof. 2000. Efficient and Robust Parsing of Word Hypotheses Graphs. In W. Wahlster (ed.). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.

T. King, S. Dipper, A. Frank, J. Kuhn, and J. Maxwell. 2000. Ambiguity Management in Grammar Writing. In *Proceedings of ESSLLI 2000.*

M. Marcus, B. Santorini, M. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics 19(2).*

S. Müller, and W. Kasper. 2000. HPSG Analysis of German. In W. Wahlster (ed.). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Heidelberg.

S. Oepen, D. Flickinger, H. Uszkoreit, J. Tsujii. 2000. Introduction. In D. Flickinger, S. Oepen, H. Uszkoreit, J. Tsujii (2000).

S. Oepen. 2002. Competence and Performance Profiling for Constraint-Based Processing. Doctoral dissertation. *to appear.*

S. Oepen, D. Flickinger, J. Tsujii, and H. Uszkoreit. 2002. Collaborative Language Engineering. A Case Study in Efficient Grammar-based Processing. *CSLI Publications*, Stanford. *to appear.*

S. Oepen, D. Flickinger, C. Manning, and K. Toutanova. 2002. LinGO Red Woods - A Rich and Dynamic Treebank for HPSG. In the web: http://lingo.stanford.edu/redwoods/

M. Pinkal, C. Rupp, and K. Worm. 2000. Robust Semantic Processing of Spoken Language. In W. Wahlster (ed.). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.

J. Piskorski and G. Neumann. 2000. An Intelligent Text Extraction and Navigation System. In *Proceedings of the RIAO-2000*. Paris.

C. Pollard, and I. Sag. 1994. Head-driven Phrase Structure Grammar. Chicago, IL & Stanford, CA: The University of Chicago Press and CSLI Publications.

S. Riezler, T. King, R. Kaplan, R. Crouch, J. Maxwell, and M. Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*. Philadephia.

P. Schmidt, S. Rieder, A. Theofilidis, and T. Declerck. 1996. Lean Formalisms, Linguistic The- ory, and Applications. Grammar Development in ALEP. *Proceedings of the 16th International Conference on Computational Linguistics.* Kopenhagen.

M. Siegel. 2000. HPSG Analysis of Japanese. In W. Wahlster (ed.). *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer, Heidelberg.

W. Skut. 1999. Partial Parsing for Corpus Annotation and Text Processing. Doctoral dissertation.

H. Uszkoreit, R. Backofen, S. Busemann, K. Diagne, E. Hinkelmann, W. Kasper, B. Kiefer, H. Krieger, K. Netter, G. Neumann, S. Oepen, and S. Spackman. 1994. DISCO – an HPSG-based NLP System and its Application for Appointment Scheduling. In *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto.

H. Uszkoreit, D. Flickinger, W. Kasper, and I. Sag. 2000. Deep Linguistic Analysis with HPSG. In W. Wahlster (ed.). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Heidelberg.

W. Wahlster. 1997. Erkennung Analyse Transfer Generierung und Synthese von Spontansprache. In *Spektrum der Wissenschaft 4/97.*

W. Wahlster (ed.). 2000. Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Heidelberg.