

1.2 Forschungsprogramm

1.2.1 Zusammenfassung

The SFB 732 *Incremental specification in context* aims at providing a formal characterization of the mechanisms that are involved in the process of disambiguation at all levels of linguistic analysis. The SFB is an interdisciplinary research center, where theoretical linguists in collaboration with computational linguists work closely together to provide a solution to the following problems: i) How do we deal with ambiguity within a language and across languages? ii) How do we choose the right interpretation for a linguistic unit among different possible ones? iii) How can we formalize this selection process via the use of statistical methods?


1.2.2 Darstellung des Forschungsprogramms

The main goal of the SFB 732 *Incremental specification in context* is to achieve a better understanding of the mechanisms that lead to ambiguity control/disambiguation at the various levels of linguistic analysis. The main hypothesis that we put forth in our first proposal is that ambiguity can best be analyzed within a framework of underspecified representations and hence disambiguation is a process of specification of an underspecified input. Specification processes are always situated in a particular context which provides constraints and triggering conditions. They relate to two types of representations: underspecified and more specified ones. The SFB 732, with its many parallel explorations of
specification in context, has as its ultimate goal to come to a better understanding of how exactly specification works in different contexts and in particular languages.

We defined underspecification as the formal method of dealing with ambiguity at all levels of linguistic analysis. Ambiguity is a property of natural language which normally involves a choice between two (or more) specific meanings that make sense in a particular context. Underspecification is a technique that allows us to represent information in a compact form, and for this reason was put forth as a useful tool in dealing with ambiguity. The basic idea of underspecification is that of a mode of representation which leaves certain features of the represented object undecided, but does this in a logically perspicuous and computationally economical way.

Given our hypothesis, our research program is thus designed to deal with the following two main questions: (i) what is the nature of the transformation from underspecified to more specified representations? and (ii) what is the role of the context in this process, what kind of information does it provide, and when does it become relevant?

Our research in the second funding period will focus on the architecture to be assumed for specification processes and on determining the size of the relevant context that plays a role in the process of specification.

I. Overview of the achieved goals of the first funding period

We have made significant progress in our understanding of these two questions and in this section we summarize our main results. We illustrate the milestones reached at the end of the first phase of the SFB 732, first for its different areas and then for the SFB as a whole.

The results achieved in each individual area suggest that the partition in four core areas of research that we laid out in our first proposal was particularly fitting for our purposes. In addition, the numerous collaborations established within each area and across areas strengthened our original conviction that successful transfer between theoretical linguistics and computational linguistics will not only give us a better understanding of the role of context in the process of specification, but will also lead to a productive exchange of methods between these two branches of linguistics. A number of characteristic examples of collaboration are given to demonstrate the importance of the SFB umbrella for achieving these substantive results.

Projektbereich A: Speech, prosody and exemplar representation

In area A rich, high-dimensional representations, characteristic of exemplar theories, have been proposed in order to circumvent shortcomings of the traditional pipeline model of language and speech processing. These were applied to subsegmental phonetic representations in A3, phonological segmental representations in A2 and prosodic representations in A1. The projects have also investigated and simulated the role that local context plays in the mental representation of language units and in language processing.

A new model of how phonetic knowledge guides speech production has been developed within area A: the Context Sequence Model (CSM). In the CSM, target acoustic patterns are determined based on selection of previously heard or produced sounds from a memory store. Since signals in the memory correspond to long stretches of continuous speech, individual speech sounds always appear in a larger context. A key property of the model is that the selection of exemplars for production is dependent on the degree of similarity between the contexts in which they originally occurred and the current produc-
tion context. In computer simulations based on realistic data extracted from a large single speaker production corpus, the projects in area A demonstrate that (1) optimal selection of context-appropriate segment-level exemplars requires consideration of about 0.5 seconds of context material preceding and following the exemplars in question, and (2) context-dependent production at this low level may be responsible for a range of frequency effects that have previously been assumed to involve the word, the syllable, and other higher levels of organization. The CSM is an alternative to pipeline sequential models of language production (cf. Levelt’s Model) as well as to Interactive Activation models (WEAVER++, Dell’s Model). Moreover, it is one of the very few models of language processing which is supported by a computer simulation.

The CSM requires robust multi-dimensional representations of phonetic and phonological space. In the computer simulations carried out in A2, amplitude envelopes but also landmark features have been used. The prosodic representations have also been represented in a multi-dimensional space (PaIntE parameters) and computed together with the properties of context (A1).

The subsegmental working packages of project A3 share the temporally local view of the CSM on the speech signal. The Multiple Description Coding (MDC) was extended to ensure reconstruction and to account for the time varying nature of speech. Furthermore, to cover the non-Gaussian distribution of many speech frames, the project tried to make the Gaussian mixture model adaptive (learnable) to the speech structure. A further task was the automatic search for the best feature set for landmark detectors, the burst detector in particular. Finally, on a segmental and suprasegmental scale, our aim has been the refined definition of parameters to describe the sound of the voice. A3 implemented ways to measure subglottal sound pressure and applied them to study the acoustic interaction between the vocal tract and the subglottal cavity.

Projects A1 and A2 enjoyed a significant collaboration during phase one. A number of key questions were addressed, including the richness and granularity of prosodic representations. In particular, this collaboration established: a) how different tokens of a pitch accent are realized and to what extent the frequency of occurrence of the pitch accent type has a bearing on these realizations; b) what effect information status has on realizations of a pitch accent type and how the frequency of occurrence influences the information status type on these realizations, and c) whether frequency of occurrence in pitch accents and in information status play a combined role. Cosine similarity analyses of the parametric signatures of pitch accent realizations demonstrated a correlation between an increase in frequency of occurrence of pitch accent type and an increase in variability. With respect to the information status categories, sensitivity to frequency of occurrence is also present: higher frequency H*L accented new tokens exhibit greater variability than given tokens. Finally, a combined effect of frequency of a certain information status category and frequency of a certain pitch accent type was observed. This work yielded a number of publications listed in the respective reports of the individual projects and would not have been carried out without the SFB.

A further collaboration which would have not been possible without the SFB is the one between A1 and D2, which resulted in a successful interaction between formal semantics and computational linguistics. D2 applied the insights on Information Status gained in A1 to the task of Realization Ranking. This led to a novel approach to implementing formal semantic theory computationally, and the experimental results showed a significant
improvement in the performance of the model. The SFB was the ideal setting for this collaboration, stimulating the inter-disciplinary research across two fields which traditionally scarcely interact.

A second example of the collaboration between areas A and D is the development of the Multi-Level Exemplar Model (MLM) and the concept of half-word representations. The MLM has been successfully applied to modeling syllable duration effects, grammaticalization, and the acquisition of local grammatical knowledge.

**Projektbereich B: Meaning and disambiguation at the interface between words and phrases**

Research in area B is devoted to the investigation of systematic ambiguities of certain verbs and deverbal nouns. The projects achieved a number of important results at the empirical, methodological and theoretical level. A first important result for area B is the development of sophisticated analyses of suffixes available crosslinguistically to build eventive nominalizations (B1, B2, B4, B5). These were found to differ semantically and aspectually in a way that makes it possible to explain their distribution with different verbal roots. In one project, B2, these distinctions were used to explain the diachronic development of the various suffixes.

A second important result concerns the investigation of the semantic and pragmatic context for the interpretation of nominalizations (B1, B3, B4). This brought one problem to the attention of the SFB: in the absence of argument structure, certain nominalizations can be simultaneously interpreted as events and results, which suggests that their syntax remains unspecified. Some work in area B addressed this puzzle and proposed a new way of distributing the processes of metonymic change or shift to different parts of the context (B1). We intend to pursue the investigation of modes of disambiguation along these lines during phase two of the SFB. A close cooperation between B3, B4 and B5 has been planned to this end (see the relevant work packages in the project proposals for the three projects). Moreover, collaboration among these projects revealed that one can reliably distinguish between cases of ambiguity that can be resolved on the basis of morpho-syntactic information and cases of ambiguity that rely on coercion. The investigation of this division of labor will continue in the next phase of the SFB.

Projects B4 and B5 also dealt with verbal ambiguity. In this work, and in line with research in this area outside of Stuttgart, a consensus emerged that this ambiguity can be attributed to a super-ordinate classification of roots that feed verb formation into manner vs. results (mono-eventive vs. bi-eventive). The two projects established a close cooperation in order to determine the source of the factors that can influence verbal interpretation, such as structure, lexical semantics, as well as world knowledge and ontological classification, which will feed work on argument structure in the next phase of the SFB (see the research goals for the second phase).

B2, B3 and B5 have been concerned with corpus-based methodology which in the case of B2 had a diachronic dimension as well. In B3, the disambiguation tool is a dependency grammar with grammatical functions which turned out to be more adequate than constituent-based grammars. B3, as other projects in the SFB, noted several problems with the sequential pipeline model and developed an integrative system which combines word type identification with a syntactic analysis. B5 created a technical infrastructure which required the use of Description Logic (OWL-DL) and reasoning tools. This made it
possible to successfully extend the reasoning process from mere argument checking to disambiguating meanings with a complex presuppositional structure in the psych verb domain (e.g., *encourager*), which constitutes valuable input for the second phase of the SFB.

On the theoretical level, research within area B led to the development of a model of word formation that combines ingredients from two schools of thought: Distributed Morphology (DM) and Discourse Representation Theory (DRT). Projects B1 and B4, in collaboration with D1, provided word-structures that reflect the core properties of deverbal nouns. These projects also developed a battery of tests that help disambiguate the different interpretations available for derived nominals. While the two projects explored different paths concerning root classification and the interaction between functional structure and root meaning, they share the commitment to the idea that the same principles that guide the formation of sentences also guide the formation of words. While B4 focused on the semantic formalization, B1 concentrated on the morpho-syntax and the contextual factors that influence interpretation, the result being a frame of syntax-semantics of -ung nominalizations which B1 extended to a number of languages other than German such as English, Romanian, Greek and Spanish. In the next phase of the SFB, B1 and B4 (in cooperation with B6) will attempt to develop a theory of argument structure that relies on these basic principles. We think that such an enterprise would have been impossible without the SFB. It also raises significant issues for the division of labor among the different levels of grammar which we will address in the second phase.

### Projektbereich C: Noun phrases and context

The two projects in area C focused on noun phrases and their features. The projects investigated the interaction of these features with syntax, morphology, semantics and pragmatics in a crosslinguistic perspective from an empirical and theoretical point of view. The projects view the pieces of linguistic information from these different levels as the relevant context by means of which certain syntactic or morphological forms can be understood. In this way, the C-projects clearly showed that fine-grained interaction between different levels is necessary to capture the interpretation of noun phrases, a result whose generality will be tested in the next phase of the SFB. Crosslinguistic investigations showed that similar interactions are active in different languages, although they may differ in their actual syntactic and morphological realization. The integration into the SFB and the close interaction with other projects led to developing a broader perspective on the phenomena we investigated (adjective position, case markers, double definiteness, differential object marking).

Project C1 developed a typology of the ambiguity of adjectival modification in Romance languages and made important contributions to the understanding of the internal structure of the DP, and the internal structure of adjectives. A major puzzling finding was the unexpected misbehavior of French adjectives. The ambiguity found in other Romance languages can be accounted for in terms of two distinct syntactic structures, but the French data contradict that. In order to account for these data, a post-syntactic explanation for the availability of both adjectival positions was developed within the framework of DM.

C2 investigated the relation between case assignment, the semantic-pragmatic (nominal) features of arguments, and the referential context. The different parameters of referential contexts are often represented on scales, like the Animacy Scale, and the Referentiality Scale. The project investigated the interaction between these scales and were able to show
that very often there was interaction for certain values but not for others. One of the main contributions of C2 is that case marking does not only depend on the referential properties of the argument, but also on the lexical semantics of the verb assigning case to it.

The two projects interacted with each other in determining the semantic features of determiners, discussing relevant data and typological generalizations. They further established collaborations with other projects within the SFB. With A1 they collaborated on a) the discussion of referential categories such as definiteness and specificity, b) the information status of referents (different kinds of givenness) and the phonological representation of such categories, and c) the investigation of the role of Focus in NP Ellipsis and other linguistic environments. This collaboration was particularly fruitful, as it enabled them to formulate detailed semantic descriptions of givenness and partitivity and it also initiated some research questions for the new project C4. With B1 they collaborated on two levels, theoretical (nominal categories) and methodical (testing semantic hypotheses). On the theoretical level the projects discussed the incremental specification of nominal features in context, and on the methodical level they shared experience in the development of questionnaires to test semantic hypotheses. This was a new methodology for both projects, since questionnaires are more commonly used for testing syntactic hypotheses. There was an intense discussion with B4 on nominal categories, which led to a joint seminar (see the list in section 1.6.2).

Projektbereich D: Disambiguation in context

Area D considered incremental specification in context and disambiguation from the perspective of parsing and generation. The projects viewed disambiguation as a process of specifying which of several readings of a linguistic form are intended – a process that is incremental in many computational approaches, in particular in pipeline models.

D1 developed a general architectural framework for the construction of underspecified semantic representations. The construction algorithm is modular in the sense that it allows each meaning-bearing sentence constituent to make its own, separate, contribution to the semantic representation of any sentence containing it. These constituents are then combined in a series of steps that model the process of specification. D1 worked out this procedure for a wide range of linguistic phenomena.

D2 reached a mature understanding of the balance of grammatically fixed constraints delimiting the candidate set and contextual properties for the disambiguation decision in parse ranking and investigated which types of lexical properties are most helpful for disambiguation. For realization ranking, the project performed in-depth explorations of contextually relevant factors that can be read off sentence-internal properties. D2 also extended the methodological basis for evaluating realization ranking with web experiments that acknowledge a gradedness of acceptability of alternative realizations. D2 and A1 demonstrated that information status can be used successfully in natural language generation of German sentences. This research would not have been possible without the SFB, bringing together different disciplines like theoretical and computational linguistics and computer scientists.

D4 developed a wide range of linguistic features to improve the performance of statistical parsing. The resulting parser is among the best unlexicalized parsers of English and is the parser with the highest accuracy in predicting empty categories, a linguistic phenomenon that had received little attention in statistical parsing before. Another important
achievement of the project is the clustering model for predicate-argument tuples of varying lengths, which represents the selectional preferences of the predicate with WordNet concepts. The model was successfully applied to word sense disambiguation, one of the most important problems in statistical NLP.

D4’s work on parsing was strongly influenced by results obtained by D2 and D5 – e.g., the fact that unsupervised training is seldom beneficial for parsing and that re-ranking is a more promising direction of research. Conversely, a number of projects within the SFB (in particular D5) based part of their research on the infrastructure developed in D4. The main research results of D5 in phase one were achieved in collaboration with A2 and D4.

In its collaboration with D4, D5 addressed parsing, based on the BitPar parser developed in D4. For the monolingual setting, D5 showed that linguistic features extracted from unlabeled corpora can enhance statistical parsing for some attachment ambiguities, but not for others. In the multilingual setting, D5 developed an automatic process based on parse reranking for improving the parses of bitext (a text and its translation). This is the first work (together with work done in parallel at UC Berkeley) that shows the utility of bitext in parse reranking. In its collaboration with A2, D5 developed two new exemplar-theoretic models, the CSM, mentioned above, and the Multi-Level Model (MLM). The relevance of these models for phonetics was discussed above. From a computational point of view, these models are a new formalization of context-dependent processing, with applications to both phonetics and syntax. The CSM explains how units are formed in a context-dependent way; the MLM models the interaction of these units with each other and with context. D5 also investigated the important distinction between horizontal and vertical axes of context and showed that the MLM can model them in one unified model. This distinction will be one of the new central topics of phase two of the SFB. The work on CSM and MLM has received international recognition and grew out of the close collaboration of phoneticians and computational linguists that was made possible by the SFB.

II. Summary and overall assessment

During phase one of the SFB, we examined the specification processes that arise in each of our four areas, and we investigated the question how disambiguation takes place on the basis of an underspecified representation. Given our results, we think that we now have a much better understanding of the mechanisms involved in disambiguation at all linguistic levels represented by the SFB projects. We are confident that we developed significant tools and models to tackle ambiguity and made substantial contributions to the general development of each area, as can be seen by the international response to our work (presentations at well-known conferences in all areas world-wide and major peer-reviewed publications).

Turning to the contribution of the SFB as a whole, consider once again the following general theoretical questions that we stated in our first proposal, and we expect our research will answer in the long term:

1. What is the relationship between underspecified and more/fully specified representations?

2. What is the role of the different linguistic levels in the process of specification?

3. What is the impact of contextual factors on the specification of linguistically relevant properties of linguistic expressions and their representations?
What is the range of specification issues that one particular bit of contextual information (or one “kind of context”) may be relevant for?

What is the general benefit of context-dependent specification?

The work carried out within the first phase of the SFB highlighted the importance of context in the process of specification, verifying our initial hypothesis that without context there is no specification (question 5). Although, as we have been able to show, in some instances, a specification process can be delayed till further information becomes available, ultimately specification takes place by considering a wider range of contextual factors, which is a key to successful communication. In our first proposal, we presented two models of how a specification process may take place in context and thus how the remaining questions may be answered:

a) a pipeline model which involves constructing representations at a number of different levels (belonging to phonetics, morphology, syntax, semantics). At each level the initially constructed representation can be an underspecified one, but the final result ought to be (ideally) a coherent sequence of fully specified representations. According to the pipeline model construction proceeds “form bottom to top” throughout: the representation levels are linearly ordered and representations are constructed one after the other, in a fixed order. Moreover, specification also propagates along this pipeline: specification at level 1 may trigger specification at level 2 and so on, but always in this direction.

b) an interface-based model, where the focus is placed on the interaction between two adjacent linguistic levels in determining a particular linguistic expression. Ambiguity resolution does not involve a series of transformations from one level to the other, but rather the availability of several ‘templates’/feature combinations in which underspecified elements can occur.

Some of the work that has been done during phase one of the SFB indicates, not surprisingly perhaps, that often specification takes the form of more complex patterns than the pipeline or the interface-based model and requires the development of new/additional models. We note here that one of the formal models developed in phase one of the SFB employing computational learning techniques and applying them to a corpus of natural speech in context (CSM) defies the original classification as either a pipeline or an interface model.

Specification processes of an underspecified representation often operate in parallel rather than in sequence and thus question both step-by-step filtering as in (a) or the interaction between just two adjacent modules as in (b). Within a single project or in a small group of related projects, a natural move was to assume a modified module architecture that will allow an integrated specification process for all those sub-processes which have the potential to operate in parallel, using a multi-dimensional representation for all relevant internal and external distinctions. This move was in essence successfully made by many of the SFB projects (in some cases following more or less established considerations from the respective scientific subfield). This enabled us to approach question 4 in a novel way by considering a variety of factors and triggering conditions that lead to disambiguation.

Zooming in on question 3 in our first proposal, we named four possible aspects of context that seemed to us to play a role in specification processes:
1. linguistic vs. extra-linguistic contexts
2. local vs. global contexts
3. dynamic vs. non-dynamic contexts
4. crosslinguistic stability of context

1, 3 and 4 will continue to play a role within the SFB. However, in the different areas of the SFB, different dimensions play a role in disambiguation processes, which cannot all be subsumed under the same principles of *locality*. For instance, *locality* is understood as linear in area D, temporal in area A, structural in some projects of area B, and discourse-related in some projects in areas B and C. In other words, what kinds of contexts are used differs between areas (and even within single areas).

*Locality* thus may not be the right term to cover all these possibilities. But since all these kinds of contexts serve as information sources for incremental specification, there are some traits that they must all share. This we will address in phase two of the SFB.

III. **Internal organization of the SFB during the first phase**

The communication between the principal investigators and the members of the SFB created a highly interactive atmosphere. There have been two SFB colloquia which take place on a (bi-)weekly basis: one on Mondays, which is an internal colloquium, and one on Thursdays that primarily figures external speakers. Besides these general meetings, there were several additional meetings in smaller groups, where two or three projects worked together on a particular problem. Especially fruitful was the integration of younger researchers in these meetings, which led us to include a graduate school project in this proposal. In general, the integration of young researchers into the ongoing research of the SFB offered them a number of significant opportunities: on the one hand, they were able to present their work and receive feedback from the senior staff of the SFB; on the other hand, they were able to establish contacts with several external visitors and discuss their research projects with them. This undoubtedly makes them more visible in their respective fields and hence increases their chances of tenure hiring in the future.

We created *SinSpeC*, the Working Papers of the SFB 732. *SinSpeC* aims at publishing ongoing work within the SFB in a fast and uncomplicated way in order to make our results known to the scientific community and strengthen our international relationships. It publishes papers by the staff of the SFB as well as papers by visiting scholars or invited scholars. The working papers are published by the *Online Publikationsverbund* of the University of Stuttgart (OPUS) and are available online.

During phase one of the SFB, a number of workshops and colloquia were organized and numerous guests came to Stuttgart who had an impact on the development of our research (for a complete list see the Z report). Here we would like to mention the following key meetings: in 2006 our inaugural meeting in November, to which Prof. Borer, Prof. Hirschberg, Prof. Picallo and Prof. Johnson contributed as the invited speakers. In 2007, the conference on *Nominalizations across languages and frameworks* in December, as well as the international SFB conference on underspecification in theoretical and computational linguistics. Both events brought to Stuttgart a number of researchers working on issues of disambiguation from a variety of perspectives. In 2008, the conference on derivational
morphology in May that aimed at triggering a discussion among different morphology frameworks and our Klausurtagung in November with Prof. Bybee, Prof. Crouch, Prof. Danlos and Prof. Hay as the invited speakers. In 2009, the Roots workshop in June and the workshop on Nominal systems across languages in September (the latter in Barcelona). The international interest in all these events has been significant.

During our Klausurtagung in Blaubeuren in November 2008 we identified the new projects to be included in the SFB in phase two. To this end, we established a procedure starting with the writing up of a short abstract of the key concept, followed by extensive discussion and two presentations, one during our weekly SFB meetings and one in Blaubeuren. The decision was then made by the SFB board after a lengthy discussion in January 2009. Due to this process, a number of ideas were turned down and modifications were suggested to some of the accepted proposals.

The SFB board met every two months and sometimes on a monthly basis to discuss current developments and strategic issues.

The centrally allocated funds such as the so-called Pauschale Mittel were of considerable importance for providing infrastructure measures and research assistant support to the two Juniorprofessors in the SFB (Prof. Marzo and Prof. Padó) and Prof. Kuhn as well as contributing to the publication of dissertations completed within the SFB.

To our knowledge, there is no comparable institution that has the same research goal as the SFB 732 in or outside Germany. While a number of researchers in various places work on some of the topics on the research agenda of specific projects in the SFB, there is no comparable complex enterprise which involves transfer and close cooperation of theoretical linguists (of different specializations) and computational linguistics. We think that this collaboration will be of major importance for the future profile of the University of Stuttgart, which orientates itself towards the creation of interdisciplinary Research Centers. The two institutes in the SFB have a unique opportunity, also in collaboration with the Forschungsverbund Sprachwissenschaft und Kognition, to play an active role in the organization of these centers.

IV. Research goals and structure of the SFB 732 in the second funding period

We stated above that one major result of phase one of the SFB is that a modified module architecture must be assumed that will allow an integrated specification process for all those sub-processes which have the potential to operate in parallel.

One central question we intend to pursue in the next phase is whether there are ways to combine the various emerging sub-architectures, each of which is locally motivated by a complex of interacting phenomena, into a coherent large-scale architecture. The most radical alternative is to assume a single, fully integrated process for all conceivable linguistic specification sub-processes. This may however be problematic from a theoretical, empirical as well as a practical/computational perspective. It may eliminate presentational anchoring points for important theoretical notions and thus call established explanatory devices into question. Computationally, a limited degree of step-by-step filtering of a set of candidate analyses is often the most effective way of managing the immense search space of all possible full specifications for a given initial partial specification.

Concerning the architecture, the guiding objective for the second funding period will be to investigate the combination of “module complexes” that have been established to derive important contextually controlled specification sub-processes, and to situate the module
combinations in the continuum ranging from a relatively sequential pipeline architecture to what we label a integrated process architecture.

While the development of a unified theory spanning all linguistic levels, theoretical paradigms and methodological approaches is to be expected at a further point in time, we do expect that the juxtaposition of neighboring approaches (along various dimensions) under the above common research questions will help establish results with higher implications for the value of our architectural assumptions for theoretical and computational purposes.

We mentioned above that the notion of locality of contexts may not be the right term to cover all possibilities (within the SFB), as the kinds of contexts that are used differ between areas. But since all these kinds of contexts serve as information sources for incremental specification, there are some traits that they must all share. Incremental specification involves contexts that vary in how rich they are, i.e. how much information they carry. For some specifications modest contexts suffice, other specifications require richer contexts. Common between the different areas is that contextual richness is determined along two orthogonal dimensions: a horizontal and a vertical one, a distinction reminiscent of the familiar one between paradigmatic and syntagmatic relations in the linguistic tradition. On the horizontal axis, we measure the size of the context, that is, the quantity of material that surrounds an underspecified expression. On the vertical axis, we measure the context’s informational “depth”, that is, what aspects or features that are relevant for disambiguation are to be extracted from the context, and how much general information (such as world knowledge) that adds to the specification-relevant significance of those aspects/features is to be taken into account.

By defining what is measured along the two axes in more concrete terms we obtain a more determinate structure of the size of the relevant context. This structure will vary depending on the kind of specification for which the contexts are needed in the different areas of the SFB. How these context sizes are structured and how similar or different their structures are is a question we want to explore more closely in phase two.

For instance, in area A, the MLM captures the emergence of syllabic durational autonomy from the underlying segmental level. The horizontal co-occurrence of segments leads to the emergence of a syllable-level paradigm.

In area B, a coarse classification of roots into manner vs. result provides the informational depth and seems to have significant consequences for their behavior in processes of word-formation and in the realization of the verb’s arguments. On the other hand, relations that emerge in a particular syntactic environment yield disambiguation effects as soon as new material enters the composition: for instance in the case of Absperrung, two main interpretations are in principle possible: an object and an event reading, but in the string die eiserne Absperrung only the object reading survives.

In area C, verb classes build paradigms when it comes to classification in terms of Affectedness scales. However, the specific relations between verbs and the distinct realizations of their arguments yield the horizontal axis. Similarly, discourse particles that are able to occupy the Nacherstposition build a vertical, paradigmatic opposition to the ones which are not able to occupy this position, whereas the horizontal axis is represented by the aspects of the sentence and its immediate discourse context that lead to the observed discourse effects of the discourse particles in this particular position.

In area D, similar distinctions as for instance in area B play a role as regards the lin-
guistic phenomena that are being modeled computationally. Beyond this, however, the work in this area is in a position to compare different architectural decisions in the design of probabilistic models, and to investigate how they relate to contextually constrained (under-)specification at the classical levels of linguistic description (phonetics, phonology, morphology, syntax, semantics, pragmatics). Besides accurately capturing the linguistic structure in the representations, a major concern of probabilistic modeling is the reliable corpus-based estimation of the model parameters. Here as well, a distinction of two dimensions of modeling can be helpful for systematic comparisons.

In collaborations across projects, described in the relevant sections of the individual projects in our proposal, we can address the question of how the different hierarchies of context sets relate to one another (either by comparing their theoretical characterizations, or, perhaps more reliably for more remote theoretical approaches, by comparing how concrete corpus instances are classified as belonging to this or that set). By using samples of language usage (i.e., corpus data) in the characterization of context sets, frequency effects (due to linguistic and/or extra-linguistic contextual factors) can be taken into account in characterizing similarity and can be modeled by assuming a probability distribution of alternative specification options. Probabilistic modeling of context set membership may in some cases provide an alternative derivation of an effect that would otherwise require a particular representational assumption. The SFB as a whole is in a position to compare such alternative models of specification in context and to relate them to each other, taking into account theoretical as well as algorithmic concerns and the reliability of statistical estimation.

A welcome addition to the SFB are psycholinguistic investigations (and at least in one project, A2, issues of language acquisition). These have not figured in our research during phase one of the SFB. But in phase two, they will play a role in all areas. We see this as an important and promising extension of our research program, as it complements our corpus and introspective data and introduces a procedural aspect in the process of context-specification. We will make this more specific in our description of the research program for the individual areas below.

There were two further long term goals that figured prominently in our first proposal: the explicit modeling of speech context and the use of linguistics in statistical modeling of language. As can be seen from our report, significant progress has been made in both areas. In the phase two we will continue to pursue the goal of advancing the state-of-the-art in computational linguistics by formalizing linguistic theories on the one hand, and probabilistic models of language on the other, in a way that allows linguistic insights to be incorporated into the “symbolic core” of statistical models. In our report we have presented clear examples of how this improves the performance of NLP applications and gives new impetus to theoretical linguistic research.

**Projektbereich A: Speech, prosody and exemplar representation**

The development of exemplar-theoretic models in phase one of the SFB, in particular the CSM, mainly addressed static phenomena. In phase two, the exemplar models will be extended to account for dynamic properties of speech, such as the change of exemplar representations over time, the emergence of categories, the dynamics of exemplar transfer from L1 to L2, and in particular, the role of context in specification and disambiguation when exemplars change over time. Both the CSM and the MLM are formal models of
perception and production in context. The MLM has some properties in common with the pipeline model as originally envisaged for phase one of the SFB in that features of units of speech on one level of linguistic description percolate to units on the next higher (or lower) level. By contrast, the CSM, which presupposes the concept of an integrated context, defies the original classification as either a pipeline or an interface model, a finding that has contributed to the general discussion within the SFB on converging sub-architectures proposed for phase two. The exemplar models developed in phase one will be put to crucial tests in phase two. For instance, the concept of phonetic convergence investigated in the new project A4 will require the application of the exemplar models in a natural setting of incremental specification in a communicative context, viz. phonetic convergence. It will also help augment the exemplar models with a function accounting for recency of occurrence.

The continued collaboration between A1 and A2 aims to establish a better understanding of the interaction between segmental and prosodic specifications and of the role of prosodic features in exemplar representations. A2 and A3 will collaborate on tackling the fundamental problem of speech segmentation—as a prerequisite of speech acquisition—using unsupervised and supervised methods, respectively. Beyond area A, there will be continued strong collaboration especially with projects in area D. The exemplar-based language models developed in A2 will be evaluated in D2, which will use them for realization ranking, and in D7, which will use them for modeling topical global context. The half-word models developed in phase one will be applied to finding covert readings of ambiguous sequences in D6.

In area A, a new project has been added, namely A4. The SFB is a natural environment for A4 and the project complements the research goals of this area in a significant manner. First, as mentioned above, it will contribute to the augmentation of exemplar models. Second, since incremental specification naturally occurs in spontaneous conversational speech, the project will advance our understanding of the dynamic aspect of context-specification, as the influence of the conversation context grows dynamically with the duration of interlocutors’ exposure to it.

A further change in this area concerns A3. Prof. Dogil no longer figures as a co-principal investigator in this project. He will be responsible for project A4 and as a co-principal investigator for project A1 and the integrated graduate school project. The research program of A3 can be easily carried out by the other two principal investigators (Prof. Yang and Prof. Wokurek).

Projektbereich B: Meaning and disambiguation at the interface between words and phrases

As in phase one, all projects of area B will deal with lexical and supralexical semantics and examine different aspects thereof such as word formation (nouns and verbs), the similarity between word-internal structure and phrase structure, and the type of lexical information that provides the building blocks for the meaning of phrases and clauses.

In the next phase, we envisage a common direction of research in three domains. First, several B projects try to determine the factors that license argument structure in the area of verbs and nouns. We hypothesized that a lexical entry registers only a core meaning, and this core meaning combines with certain syntactic constructions. But the core meaning influences the syntactic behavior to a considerable degree. The conclusion is that a better
understanding of the issue of argument alternation requires a better understanding of the interaction between roots, i.e. the idiosyncratic part of meanings, and syntactic structures. The projects in area B, as well as C2, tackle this question from different perspectives: e.g. some are more concerned with the behavior of the root, others with the nature of the syntactic structure. This multi-dimensional approach will definitely contribute to the general discussion on these topics.

Secondly, several B projects (B1, B4, B5 and B6) in cooperation with project C2 aim to define the role of Affectedness in grammar. Their goal is to provide a semantic characterization of Affectedness and discuss the constructions that languages use to express Affectedness.

Thirdly, projects B3, B4 and B5 are going to jointly explore the different forms that semantic ambiguity can take and the disambiguation/specification mechanisms that can be used to disambiguate these different forms. The starting point for this work are detailed investigations concerning the disambiguation of sortal ambiguity (as found in nouns like \textit{Behauptung} or \textit{Absperrung}) that have been carried out during phase one in B3 and B4, as well as work on incremental specification of lexical meaning that has been done in B5. The goal of the cooperation is to come up with a detailed, theoretically informed description of the different types of ambiguity and disambiguation mechanisms, with close attention to their logical and algorithmic properties. We also want to see which of the described disambiguation mechanisms can be implemented or approximated by modules that could be incorporated into NLP systems.

In area B, project B1 figures only Prof. Alexiadou as its principal investigator in the next phase of the SFB. There are two reasons for this. First, the research questions that fell under Prof. von Heusinger’s area of expertise have been successfully answered during phase one (and will be part of the investigations in B3, B4, and B5) and second, the orientation of the project in phase two is mainly morpho-syntactic.

One project, B2, will not continue into the next phase, and two new ones have been added, B6 and B7. Due to the departure of the principal investigator of B2 (Prof. Becker), the project terminates its investigation at the end of the first phase having produced a number of significant results on the aspectual properties of nominalizations mentioned in our exposition here and in the report of the project.

B6 investigates a core problem of the morphology-syntax-lexicon interface, namely the complex relationship between Voice morphology and argument structure. The project is an important addition to area B, and shares with projects B4, B5 and C2 a common interest in the behavior of verbs and the realization of arguments. In cooperation with the other projects in this area, as well as with C2 and D2, it will provide significant input towards the development of a theory of argument structure. B7 is a further welcome addition to our work on word-formation as it enriches our research agenda with the process of conversion, a topic that has not figured prominently in the first phase of the SFB. This topic is directly relevant for the research in B1, B4 and B5 concerning the structure of the lexicon and its relationship to syntax and morphology. Additionally, B7 will carry out a series of experiments which will contribute to our understanding of the structure of the mental lexicon, and will complement the other research methods within the SFB.

\textbf{Projektbereich C: Noun phrases and context}

The C-projects of the first phase examined the specification of grammatical forms in the
nominal domain. We will continue this direction of research, but we will extend the context according to which we evaluate the grammatical forms. C2 will focus on the interaction between the lexical semantics of verbs and the semantic properties of arguments with respect to case alternations. C4 will investigate the interaction of discourse structure and the semantics of discourse particles. In this way, both projects will provide detailed analyses of the interaction of different linguistic levels (contexts) that lead to one grammatical form, and contribute to the overall research program of the SFB.

In area C, one project, C1, will not continue into the next phase, and a new project, C4, has been added. With respect to C1 we note, as stated also in the report of the project, that the project concluded the core of its investigation as requested by the reviewers of our first proposal, which was to provide a typology of adjectival modification in Romance. A number of further results on the structure of the noun phrase, the structure of adjectives, and multiple determiners across languages will be presented in two forthcoming dissertations that have emerged from this project.

C2 brings two new perspectives into the discussion of argument alternations: (a) sensitivity of argument alternations to universal semantic scales modeled as feature architectures and (b) the combination of argument alternations with Differential Object Marking. C2 will elaborate the concept of Affectedness, which also plays a crucial role in the explanation of restrictions on voice formation addressed in project B6 (as well as in B1, B4 and B5). It will contribute to the ongoing crosslinguistic exploration of the principled relation between verbal semantics and other semantic properties.

C4 will provide a model for the specification of underspecified discourse particles in one particular and well defined position, the “Nacherstposition”. There is a clearly defined set of discourse particles that are allowed in this position. Besides their original meaning as indicators of discourse relations, once they are in this designated position, they are also involved in marking additional discourse phenomena, such as topic change. The project plans to investigate the semantic properties of these discourse particles and their interaction with discourse structure. The research goals of the project correspond to the main area of interest of the SFB in two ways: (i) by investigating the process of specification of underspecified input in context, and (ii) by focusing on discourse context as a context type in which complex phenomena like information structure, discourse particles and use of underspecified lexical items have their share. The research questions concerning the interaction between discourse structure and lexical information will be investigated in collaboration with B4 and A1. Moreover, C4 will partially adopt an experimental methodology to approaching its research questions, which will enable the formulation of generalizations in a still poorly understood area.

Projektbereich D: Disambiguation in context

The projects of area D address the issue of specification in context from a computational perspective. Given a partially specified linguistic representation how can a contextually appropriate fully specified representation be computed? Typically, the initial partial representation reflects surface information of an utterance, such as the string, and the specification process models the comprehension process, i.e., disambiguation among possible readings. In other cases, the initial representation may be an abstract content representation and the specification process amounts to choosing among realization options, modeling the production process. Several D projects investigate aspects of parse disambiguation,
while some others are concerned with vagueness, adopting a variety of formalisms for (under-)specified representation and paradigms for modeling choice. The generalization of a particular paradigm, a discriminative ranking approach, to the reverse process is also addressed in the D area. Furthermore, the production view is underlying in generative statistical models that are employed in several projects. Parameters for such production models are estimated from annotated data or bootstrapped from un-annotated data with unsupervised learning techniques, such as EM algorithms. The models are then applied to obtain underlying representations for a given surface representation by statistical inference. The specification process is classically viewed as a sequence of modular filtering steps operating on certain levels of representation, the classical pipeline model (e.g., part-of-speech tagging, syntactic parsing, semantic role labeling). With the means of underspecification, it is, however, possible to delay specification decisions at each level. This technique is used extensively in all D projects; in particular with machine learning techniques for inducing disambiguation models from corpus data, i.e., from samples of natural language use in context, it is common to observe indirect filtering effects that can skip levels. Hence, an integrated, non-pipeline model is conceptually more appropriate. At the same time, a limited amount of sequential filtering is often used for algorithmic considerations. The balance between step-wise filtering and integrated interaction of specification options will be a central question for the D projects in the next phase.

D2 will continue its successful research, focusing on the effect context has on generation. Larger discourse contexts will be taken into account and the importance of contextual clues in generating from underspecified semantic representations will be investigated.

D4’s main topic will remain statistical parsing, but here, too, the envisioned research concerns an extended array of contextual factors. Data sparseness is the main problem when trying to estimate parameters for larger contexts. D4 will investigate a novel estimation method, based on decision trees, for dealing with this problem. In addition, D4’s model of argument structure will be extended to include adjuncts.

In area D, two projects, D1 and D5, will not continue into the next phase and two new projects, D6 and D7 have been added. The main reason for the completion of D1 is that the current and future research interests of the principal investigator diverge from the research goals of the SFB. Some aspects of the research of D1 will continue in B4. In D5, the next obvious research questions are mainly in the areas of machine learning and statistical estimation. But a project solely focusing on the machine learning and statistical estimation aspects of biased learning would not be a good fit for the SFB as a whole. Some aspects of the research in D5 will be pursued in A2, which will use the methods A2 and D5 developed for syntax to model phonological well-formedness.

The two new projects in this area, D6 and D7, both constitute important extensions of our research goals. D6 will address the interpretation process of verb-object pairs where an implied event is not overtly realized. The interpretation process will be viewed as an instance of incremental specification in context, looking at factors like lexical semantics, aspect and plausibility. A central goal of D6 will be to investigate the question of whether it is possible to integrate all the different contextual factors into one multi-dimensional model. In addition to computational models, D6 will use psycholinguistic methods as much of the evidence on covert events stems from psycholinguistics.

D7 will deal with an important NLP application, sentiment analysis. Work on sentiment analysis has been dominated by machine learning approaches that have not systematically
looked at context, linguistic factors, and languages other than English. D7 proposes a linguistically motivated framework in which different contextual factors can be systematically investigated. An additional focus will be sentiment analysis in German.

A further change in this area concerns project D2. Prof. Kuhn, who is about to move from Potsdam to Stuttgart, will join the SFB as co-principal investigator in this project. Prof. Kuhn has long been involved in the LFG ParGram project. The three new directions adopted in the next phase of D2 (see the relevant part in the project) reflect the research interests of Prof. Kuhn.

1.2.3 Umgang mit den im Verbund erzielten Forschungsdaten

Introduction

The subdisciplines represented in the SFB are all fields that heavily rely on linguistic data. For this reason, management, publication, reusability and preservation of these linguistic data is a high priority for us.

Linguistic data present a number of specific challenges for data management: the complexity of linguistic annotation, copyright and related rights, privacy issues and, frequently, the large size of the resources. Many of these questions require national and international collaboration and are the subject of ongoing research, development and standardization. The IMS is an active participant in these activities. In particular, the IMS (represented by Apl. Prof. Heid, Dr. Schmid and Prof. Schütze) is a partner in D-SPIN, a research consortium led by Prof. Hinrichs, Tübingen, that aims to provide the basis for a stable and sustainable infrastructure for language resources and technologies. As part of his infrastructure, D-SPIN designs web services and establishes resource registries for sustainable accessibility of German language resources. D-SPIN is the German national complement to CLARIN, the pan-European collaboration for creating and coordinating language resources and technology and for making them available and readily usable. D-SPIN is currently funded until 2011; a continuation is envisaged (see also section 1.7).

The IMS (represented by Apl. Prof. Heid) is also a member of NA 105, Normenausschuss Terminologie, des Deutschen Instituts für Normung (DIN). NA 105 is the technical standards body of DIN responsible for mirroring work of the ISO groups on linguistic resource standards (ISO TC37/SC-4) and defining instances of such standards for German.

The premier international conference for issues relating to the management of linguistic data is LREC, the Language Resources and Evaluation Conference, organized by the European Language Resources Association (ELRA). The SFB has been an active participant in this conference with more than half a dozen submissions to the 2010 conference.

Data management in phase one of the SFB

The main types of data resources created during phase one are annotated speech and text corpora. The linguistic annotations of these corpora are encoded in standard formats such as German SAMPA for segmental information, GToBI(S) for prosodic information, GIZA format for translation alignments, Penn Treebank format and PAULA for parse trees, and formats like MAF and SynAF being developed by ISO TC37/SC-4 for (morpho-)syntactic information.

In addition to annotated speech and text corpora, several of the theoretical linguistic projects in the SFB have collected data via introspection and questionnaires. In theoretical