1.2 Forschungsprogramm

1.2.1 Zusammenfassung

Taking as a starting point the pervasive presence of ambiguity that characterizes all levels of linguistic analysis, the task of the proposed SFB is to achieve a better understanding of the mechanisms that lead to 'ambiguity' control/disambiguation. We regard ambiguity as the result of underspecification and hence disambiguation processes as processes of specification of an underspecified input. Indeed a wide variety of linguistic processes can be captured along these lines (speech perception and recognition, morphological, syntactic and semantic disambiguation etc.) and will constitute the research agenda of the proposed SFB.

By a specification process we understand any linguistic process that can transform one linguistic representation into any one of several alternative more specific representations and which in doing so has to make a choice between these on the basis of the evidence taken from some relevant context. Specification processes are always then per definition situated in a particular context which provides constraints and triggering conditions, and they make reference to two types of representations: representations involving forms of underspecification and more specified ones.

The research program deals with two main questions: (i) what is the nature of the transformation ("incremental specification process") 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 this become relevant?

Concerning our understanding of the notion "context", we see four aspects that are relevant for the investigations in this SFB: linguistic vs. non-linguistic contexts, local vs. global contexts, dynamic vs. non-dynamic contexts, and the cross-linguistic (in)stability of contexts.

We see the SFB, with its many parallel explorations of specification in context, as providing a unique opportunity for coming to a better general understanding of how exactly specification works in contexts and in particular languages. And we think that successful transfer between theoretical linguistics and computational linguistics will not only give us a better understanding of incremental specification in language, but that it will also lead to a closer collaboration in methods and results of these two branches of linguistics.

1.2.2 Darstellung des Forschungsprogramms

At all levels of analysis, language is ambiguous. The remarkable thing about ambiguity, however, is not so much that it exists, but that it does not really create too many obstacles. It is this combination – the pervasive presence of ambiguity and yet the surprisingly few obstacles it puts in the way of effective communication – that are now seen as a major challenge to both computational and theoretical linguistics.

In recent theories of theoretical and computational linguistics the concept of underspecification has been pivotal for dealing with ambiguity. The idea then is that any process of disambiguation is actually a specification process of an underspecified input. However, in order to transform an unspecified representation to a more specified one, reference to a particular context which provides constraints and triggering conditions is necessary.

As a rule, ambiguity resolution in theoretical linguistics can only be adequately described if the interaction of different linguistic levels is taken into account. A proper account of this interaction is indispensable because often different bits of the information that is needed for disambiguation are represented at different levels. The specific domains of investigation with which the proposed SFB is concerned are:

1. the nature (and in particular the size) of the linguistic entities which undergo specification (phonemes, morphemes, words, clauses, sentences, utterances, texts)

2. the kind of properties that are to be specified

3. the kind of contextual information involved; how local or global is it? is its source/ nature linguistic or extralinguistic? is it subject to dynamic effects or not? is it crosslinguistically stable or not?

4. the means by which and the point at which in the general process of interpretation and production contextual information is made use of;

5. the interaction between the different contextual parameters that are involved in a specification process;

6. the relations between underspecification and ambiguity; and,

7. ultimately attention to the question when and where underspecification is a linguistically legitimate way of dealing with ambiguity.

These specific questions are embedded in the more general question of the proper understanding of the specification mechanisms and the domains of their application to which we come back in section 1.2.2.3. They will advance our general understanding of how language works and constitute the general and long term goals that the SFB we are proposing wishes to pursue.

These commitments imply a key role in any of our investigations for the following two parameters: i) the use of underspecified and fully specified representations and the transformations between them and ii) the nature of context and its role in the process of specification. In the next sub-sections, we specify the import of these two to the clarification of our questions.

1.2.2.1 Incremental specification

As mentioned above, the concept of underspecification has been introduced in order to deal 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. This central idea can be formally represented as in (1), from Alexiadou & Müller (2005):

(1) Linguistic expressions (LEs) may lack some property (or feature) *x* at some level of representation, and exhibit the same property *x* at some other level of representation.

The relationship between the two levels of representation is determined by various rules or principles, which can have various forms depending on the theoretical framework. The motivation behind (1) is that assuming that LEs can be underspecified permits more economical analyses that would not be otherwise available.

How is specification of an underspecified *LE*, and hence disambiguation, achieved? In principle, a specification process can be broken down into transformational steps from an initial underspecified representation of the *LE* to other representations, which differ from one another as far as their feature specification is concerned. *LE* is underspecified in *r*0 (initial representation), but more specified in *r*1 (next representation) and so on. Addition of features is achieved by making reference to a particular context and we can model this process as involving a series of specification steps σ from the initial representation (*r*0), through the intermediate representations (*r*1, *r*2) to the final representations (*r*3, *r*4, *r*5) see Figure 1. Each step in this specification process makes reference to a particular context (e.g. *c*3, *c*2), and the representations themselves are ordered with respect to specificity, e.g. *r*0 is less specified than *r*1, *r*1 is less specified than *r*2. The order of specificity formally captures the notion of information gain implicit in the term 'specification'. Intuitively, one representation is more specific than another if it provides more information or has more linguistic features or if it admits fewer readings. Several factors determine the order of specificity of the different elements in the different domains where such an ordering is required, and these factors need to be investigated.

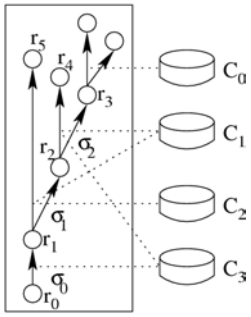

Figure 1.1: Specification process.

We come back to the discussion on the specification process in 1.2.2.4 and 1.2.2.5.

The above remarks are in line with an incremental specification process of the type used in computational linguistics, where specification is often viewed as a derivational process, in which disambiguation steps such as part-of-speech tagging, syntactic disambiguation and semantic disambiguation are applied sequentially.[1] In particular, in

[1]We are aware of the fact that the method of incremental specification is also used in psycholinguistics, and in fact this area of research has produced important results in dealing with ambiguity. We will not review this literature here, as it does not relate to our research objectives.

the classic pipeline model of computational linguistics (Figure 2), a representation undergoes a series of transformations from the lowest linguistic level to higher linguistic levels. Each step in this pipeline may involve a choice among several result representations that are formally possible.

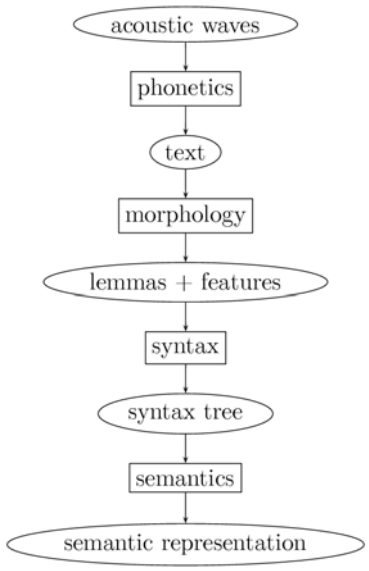

Figure 1.2: The Pipeline Model

In theoretical linguistics, the main interest lies in the close interaction between the different linguistic levels in determining a particular linguistic expression. The primary context for one level, e.g. morphology, are the other linguistic levels, i.e. – syntax, phonology etc. In this view, 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. For example, a complex word can be ambiguous because there are in principle two possible structural analyses for it, in a sense to be made precise later on in this introduction. The same applies to a particular combination of two elements, e.g. adjective-noun combinations; these can be ambiguous, as there are two possible structures in which an underspecified adjective can be inserted. The task is to investigate and develop criteria that reveal the one or the other option.

We believe that both the uni-directional "pipeline" approach and the interface approach are necessary in order to achieve a better understanding of specification and the role of the different parameters summarized under "context". However, we do take the implementation suggested in Figure 1 as a point of departure that will produce fruitful discussions which will be pivotal for our understanding of specification mechanisms.

As will become clear from the individual project proposals, most of them employ well-known representation formalisms like Underspecified Discourse Representation Structures, weighted parse forests, feature bundles and templates. We envisage the SFB as a forum for the integration of the results from the various approaches for getting a better understanding of the specification process in language and languages. We are convinced that in order for us to gain insights into the different means of disambiguation, the role of interfaces and the techniques of underspecification, it is necessary to broaden the perspective so that facts, methods, solutions, and different theoretical approaches can be compared with one another. In the first phase this discussion will be taken up primarily in area B in cooperation with projects in area D.

1.2.2.2 What is a "context" and what is "context-dependence"?

As already mentioned in the previous section, disambiguation is always situated in a particular context which provides constraints and triggering conditions. Hence the question arises: what can qualify as context for what, and thus what information that is relevant to the outcome of a given process can be treated as the "context" within which the process occurs, and on which the outcome of the process can be regarded as "dependent".

As we see it, the conceptual limitations to what can qualify as context in the study of language are a natural concomitant of the generative perspective: Linguistic entities – the objects of the study of language – can be divided into (i) basic (or "atomic") entities; and (ii) compound entities, which are built from constituent entities (being atomic or compound ones) according to certain general rules. This perspective is adopted for essentially all levels of linguistic description, with the possible exception of pragmatics. With the generative perspective comes a distinction between two types of factors that can in principle determine the linguistic properties of a compound entity: (i) the constituents from which the given compound is built and the properties that these constituents have, and (ii) the environment within which the given compound occurs, which in its turn could be a constituent of yet bigger compounds.

This schematic definition makes the notion of context dependent on an underlying notion of "composition" and "constituent structure". In practice, therefore, the criterion we have described for what kind of information is to be considered as contextual can be quite uncontroversially applied.

In this SFB there are four possible aspects of context that seem to us to play a role in specification processes. These are:

1. linguistic vs. extra-linguistic contexts
2. local vs. global contexts
3. dynamic vs. non-dynamic contexts
4. cross-linguistic stability of context

Let us briefly describe how we understand the above types of contexts and the parameters that can be used to define them. As will become clear from our discussion, these factors intersect, i.e. they combine with one another in order to give us the wealth of information required for the process of specification.

1. The distinction between linguistic and extra-linguistic contexts is rather uncontroversial and to a great extent it is theory-independent. Extra-linguistic contexts have access to extra-linguistic information, while linguistic contexts are defined clearly on the basis of the content of the individual linguistic levels.

2. The notion of context we described above is on the one hand compatible with very narrow, local contexts, e.g. with a conception according to which the context of a given expression consists of one or more other constituents with which it co-occurs in a given structure (syntactic, e.g. agreement, or phonological, e.g. assimilation rules). But it is just as possible to see contexts as extending over the entire paragraph of which the expression is part, or over the entire text, or even including various social and cultural features of the community for whose benefit the text was produced. Not all of these contextual factors are relevant to linguistics more narrowly understood, viz. as the discipline whose task is the study of grammar. In fact it will be the task of the different Projektbereiche of the SFB to articulate in more detail what kinds of contextual information are relevant within the more narrowly circumscribed research areas they represent, and the same task will arise once again for each individual project.

3. A third aspect that distinguishes between different types of contexts is their dynamics, in its two guises: as context growth or as context decay. The latter often goes together with (the temporal notion of) locality. The more local a certain type of context, the more quickly it may get replaced by another one of its kind. And for contexts which extend over somewhat longer distances we find the dynamic effects familiar from the various forms of Dynamic Semantics: the discourse context grows as the discourse unfolds and makes more and more contextual information available. (Though for discourse contexts too, context decay plays an important part, contrary to what the original proposals of Dynamic Semantics may have seemed to imply (Heim, 1982; Kamp, 1981)).

4. Finally, an aspect to which several of the projects in this SFB will pay particular attention to is the cross-linguistic stability of context, i.e. the universality of contextual parameters (primarily areas C and B). In other words, the question of whether a specification process depends on the same parameters in different languages or rather how languages differ with respect to what qualifies as a context for disambiguation. In order to determine this, contrastive analyses are necessary.

These four aspects can give only a preliminary categorization of different kinds of context. Each of these aspects is much more complex than we could account for in the previous paragraphs. This is illustrated below by a more elaborate reflection on the contrast between local and global contexts, and their interconnectedness with dynamics and language change and the methods to investigate them.

While e.g. phonetic specification occurs by considering the immediate phonetic content of a given landmark (the features of the adjacent segments), word meaning can be specified by making reference to more global and static contexts. According to certain views, word meanings can be treated as points within a large, multiply connected semantic space. Nowadays such spaces are often referred to as ‘ontologies’ (see

Bateman 1995; Casati & Varzi 1996, 1997; Bergamaschi et al. 1998; Fellbaum 1998). More recently the conception has been one contributing inspiration for network-based efforts to achieve systematic wide-coverage descriptions of (certain facets of) lexical meaning, such as WordNet and its various offshoots for languages other than English.

There is also a second notion of global context that is relevant to the aims and methods of the SFB. Many processing algorithms of current computational linguistics involve statistics. As a rule, these algorithms have to be trained on linguistic data to determine the values for their stochastic parameters. The data used for training purposes are corpora, sometimes annotated with linguistic features that the algorithm can recognise and thus it can learn faster and better. In some cases these corpora are chosen to cover as wide a spectrum of language use as possible, so that the parameter values they confer upon the algorithm that they are used to train reflect general usage in as balanced a way as possible; but in other cases, where an algorithm is designed for exclusive application to a particular type of language use – e.g. to extract information from texts belonging to a certain science or type of technology, such as biochemistry or documentation for the manufacturing and servicing of commercial aircraft (to name but two from a long open-ended list) – the training data will be chosen solely from this application domain, in order to ensure that the algorithm will reflect the biases which distinguish language use in this special area from its use in general.

Our main hypothesis is that the most global and the very local contexts are not simply separated by a large gap, but rather seem to be connected by some kind of continuum (this can be seen more clearly in the area of nominalization, a common focus of several projects in areas B and D). For one thing, global contexts are also subject to change. Their change tends to be at a slow rate, and we tend to think of it as change of the language itself, rather than as change from one set of conditions for the use of a given language to another. But the question whether certain changes should be seen as a change in current constraints on use of the same language or as change of the grammar of the language itself is often far from clear – telling these two kinds of change apart is one of the central methodological questions of diachronic linguistics.

1.2.2.3 Long term research goals

The spectrum of different types of context, with their different kinds and different degrees of locality, the different kinds of information they carry and the different mechanisms which exploit this information in determining properties of utterances and utterance parts for which they serve, is very complex and needs much further analysis. Furthermore, an understanding of the nature (and in particular the size) of the linguistic entities whose properties are to be specified, the kind of properties that are to be specified and the mechanisms used in the process of specification leading to ambiguity resolution are far from complete either. In fact, to our knowledge no systematic attempt has yet been made to chart this complex spectrum, or even a substantial part of it. We see the SFB, with its many parallel explorations of specification in context, as providing a unique opportunity for coming to a better general understanding of how exactly specification takes place in a particular context and in a particular language. And we will consider it our joint opportunity and obligation to reflect together on how the results of the individual projects can be integrated into a comprehensive picture of the processes of specification in context.

As we see it, our research will provide an answer to the following general theoretical questions in the long term:

1. What is the relationship between underspecified and more/fully specified representations? How do we reach a fully specified representation if we take an underspecified one as the starting point? Does the reverse process also hold (i.e. extracting features from a full specified representation) and what are the conditions it is subject to?
2. What is the role of the different linguistic levels in the process of specification? What is the relationship between underspecified representations on one level and fully specified representations on the next level (see the pipeline model)? If more than one specified representation exists within one level, what is the relationship between them?
3. What is the impact of contextual factors on the specification of linguistically relevant properties of linguistic expressions and their representations? How do the different contextual factors that bear on the same specification problem interact with each other: are these factors brought into play perhaps in some particular order, or are they first to be amalgamated into a set which acts as a unit? ; and when factors conflict in that they point towards two incompatible specifications, can one factor overrule the other? and what are the general principles that determine this?
4. What is the range of specification issues that one particular bit of contextual information (or one "kind of context") may be relevant for, and for how many specification issues that arise for a single utterance (and what particular combinations of them) can the same contextual information be simultaneously relevant?
5. What is the general benefit of context-dependent specification? Does it apply the same way across domains and across languages? or is it dependent on the specific domain/language? Are there cases where no specification takes place? Naturally, this question cannot receive an adequate answer before we have clarified the nature of the principles that are necessary for specification processes in the different domains.

There are two further long term goals that figure prominently in this proposal. The first one is the explicit modelling of speech context. The projects of the area A will account for the fact that linguistic representations have a rich phonetic history, reflecting speech events in various contexts, with various prosodic nuances, and coming from various voices. The correct models of language must describe the interaction of specific phonetic detail with general principles of linguistic structure. The composition of the present SFB makes it a promising framework for meeting this desirable yet elusive goal.

The second one is the use of linguistics in statistical modelling of language. The main application of statistical models is context-dependent specification. Many of these models currently do not incorporate linguistic insights (though there are some notable exceptions e.g. exemplar theory). Our long-term goal here is to advance the state of the art in computational linguistics by formalizing linguistic theories on the one hand and probabilistic models of language on the other hand in a way that allows linguistic insights to be incorporated into the "symbolic core" of statistical models. This would improve the performance of NLP applications due to better modelling as well as give new impetus to theoretical linguistic research.

Summarising, both the theoretical and the computational linguistic literature has worked out many different proposals for specification processes and many different notions of context. Individual projects of the SFB will be concerned with different selections from this general spectrum of specification processes, and as part thereof, with different choices from the general range of concept notions. For the SFB as a whole, there is the task of subjecting what its individual projects will have to say about these matters to careful comparison. Such comparisons are important, for one thing, because they may reveal how solutions from one area can be transferred to another area, where they had thus far not been applied. Within linguistics there is a tradition of successful transfers of this kind (e.g. the transfer of feature logic from phonology to syntax in the 80s or underspecification of feature values from phonology to semantics in the 90s, etc.). But the comparisons to be carried out within the SFB will touch on issues which have not been considered in such comparisons hitherto, which is why we expect that they will lead to substantial new insights and advances.

Successful transfer, however, is not the only point of such comparisons. What we also want to accomplish by their means – and this we see as a commitment of the SFB in its entirety – is to develop a more transparent and more detailed picture of the range of different forms that context-dependent specifications of linguistic properties can take.

1.2.2.4 Forschungsprogramm

The research program of the proposed SFB is split into four areas. In what follows we present the different divisions (A–D) of this program in order to articulate in more detail what kinds of contextual information for the process of specification are relevant within the more narrowly circumscribed research areas they represent. The research areas to be presented here address the role of underspecification in ambiguity resolution, the nature of the entities to be specified, the kind of properties to be specified, the nature of the contextual information involved, the way in which this information is made use of, the type of context involved and the interaction of different contextual parameters in the process of specification.

As will become clear from our exposition, the four areas assume different methods and focus on different aspects of specification processes, partly because they deal with different kinds of contextual information. We view this as an advantage that will help us clarify which principles of context dependence are important for the grammar architecture as well as for language production.

Projektbereich A: Speech, prosody and exemplar representation

In area A the underspecified representations correspond to the phonological/prosodic representations coded in underspecified distinctive features and distinctive prosodic tunes, while the fully specified representations correspond to the fully specified “exemplars” of linguistic expressions like phones, syllables and words. The fully specified representations represent/accumulate detailed phonetic knowledge that speakers have about linguistic expressions of their language, including the speaker-specific voice characteristics. Such knowledge cannot be modelled by standard categorial procedures of linguistic phonetics and phonology, because it is not purely grammatical but rather stochastic, and it is acquired by generalizing over large numbers of tokens.

The long term perspective of the projects in A is to develop a single procedure to compute underspecified and fully specified representations. This is an instance of the Incremental Specification in Context procedure (ISC), see 1.2.2.1. Incrementality of specification is central to area A, and it will partly be explored by means of methodological (analysis-by-synthesis) and theoretical (exemplar theory) approaches.

Dealing with the areas of linguistics and computational linguistics closest to speech the ISC procedure is conceptionalised in the following way: the acoustic cues are converted into distinctive features at the PIVOT/Landmark areas of the speech signal. PIVOT areas are detected by examining patterns of change in amplitude in different frequency bands. The strongest changes occur in the areas where a consonant is followed by a vowel (CV-Landmarks). Due to a type of cue in the landmark area the sounds can be classified as [sonorant], [continuant], [strident], [+back] etc. The initial estimation of features in the vicinity of syllabic nuclei is used as a first contact to the lexicon. We assume that lexical forms are represented in memory as sequences of bundles of distinctive features.

This initial ISC procedure in speech delivers not only the feature values in the vicinity of PIVOTS/ Landmarks but also a measure of confidence for the feature detection. In case the threshold of confidence (which we assume to be a statistically defined measure) is not reached, the ISC procedure proceeds incrementally to consider the immediate phonetic context of the landmark (for example, the features of adjacent segments, so called “enhancement” features). The next incremental step may consider the syllable position and the prosodic environment of the distinctive feature. In an abstract sense the ISC procedure is incrementally applied on different scales (see the temporal notion of local contexts alluded to above). In phonetics it is a time scale, in morphology, syntax and semantics the scale would be defined differently, but would project onto the time scale.

The lexicon is accessed by finding words which provide a match to bundles of (underspecified) distinctive features in prosodic structures. The context, which has been used in identifying these features and structures is a part of the ISC procedure and is also stored in memory.

The crucial part of the ISC procedure sketched here is the verification of the lexical access. We suggest that this verification is performed by an internal “analysis-by-synthesis” procedure. The procedure is applied to any hypothesized abstract word or sequence of words (or ambiguous set of words). Any such unit is internally synthesized, and the output of the synthesis is compared to the properties of the original input.

Because the phonetic parameters of the internal synthesis are determined by the vocal tract peculiarities of the hearer and as the phonetic properties of the original input signal are dependent on the vocal tract of the speaker and the acoustics of signal transmission, we suggest converting both into perceptual space parameters. That is, for the purpose of analysis-by-synthesis internal verification both signals will be normalised to fit the perceptual space of the language.

The internal analysis-by-synthesis procedure is applied in a loop and leads to the establishment of so called “exemplars”. All exemplars are represented by individual points which form reference areas (clouds) in the perceptual spaces of various linguistic units (phonemes, syllables, words, phrases, pitch accents). In our model these “exemplars” are not concrete realisations but abstract products of an internal analysis-by-synthesis process.

In our opinion, the exemplars are neither concrete acoustic patterns, nor articulatory gestural scores but are stored in the speaker's/hearer's perceptual reference space. The number of "exemplars" is dependent on the frequency of usage. This follows from the assumptions of the ISC procedure; the more often the unit is verified, the more exemplars will be generated by the analysis-by-synthesis procedure (Note: the variation in exemplars will be generated by the ever changing context of the input). The ISC procedure in speech may be illustrated as follows (Figure 3).

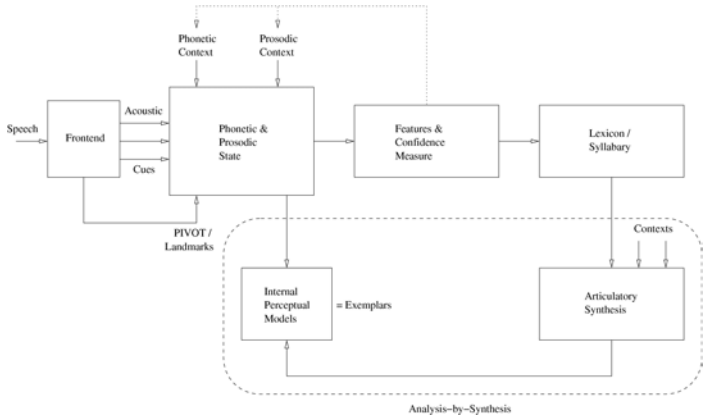

Figure 1.3: The ISC Procedure in Speech.

The individual projects in area A address the overall research program in the following way:

A1. Incremental specification of focus and givenness in a discourse context. The project will deal with the interrelation between the information structure of sentences that are part of a larger discourse or text and their phonetic form. A central assumption is (i) that the information structure of a sentence that is part of a conversation or spoken text typically relates to the sentence neighbouring it and that this information structure determines (within a certain bandwidth of tolerance) felicitous prosodic realisations; but (ii) that often a number of different contextually acceptable construals of the information structure of the sentences in its context are possible, with appropriate prosodic realisations varying in phonetic detail accordingly. Furthermore, the investigators are motivated by a shared suspicion that the prosodic realisations reflecting different ways of construing and justifying the information structure of a given sentence involve subtle prosodic distinctions (going well beyond established distinctions such as, say, that between A accent and B accent) that have thus far escaped the attention of the research on correlations between the prosodic realisation of information structure and its semantico-pragmatic significances.

Since A1 is concerned with the relations between information structure and prosody it can be seen as a bridge between area A and the other areas in the present proposal. The goal of A1 is to investigate in greater depth than has been possible up to now the ways in which the (prosodically realised) information structure of a sentence is justified by the semantic and pragmatic relations in which it can be construed to stand to the sentence or sentences that precede and follow it in a discourse or spoken text. Here the sentences surrounding the sentence in question play the part of context, in the sense of 'discourse context' as it is known in dynamic semantics.

A2. Exemplar-based Speech Representation: the project will concentrate on the ISC procedures leading to the establishment of fully specified representations of speech. It will develop a model of the internal analysis-by-synthesis, which starts from hypothesised lexical entries, takes landmark and context information as input, and results in fully specified exemplars of the relevant linguistic unit. Challenging research questions in this project include: the establishment of category-specific exemplar regions, supposed to be frequency based, the definition of the dimensions of these regions in perceptual space and, from a procedural point of view, a computational model of the incremental specification of exemplars by enriching their underspecified featural representation with contextual information.

A3. Incremental specification in speech: The project will work out the procedures of landmark detection in phonetic and phonological contexts. It will specify the statistical procedures for confidence measures of detected landmarks. Thus it will provide the computational environment for the implementation of all ISC sensitive procedures in the A area of the SFB. The importance of this project for the research goals of Area A lies in the fact that it will introduce new methods of incremental context evaluation and specification. A general principle from mathematics and statistical signal processing dealing with incremental information is the so-called innovation approach. This approach has been successfully applied to the solution of mathematical problems (e.g. linear estimation). A challenging fundamental question is whether the innovation approach is applicable to linguistic and phonetic problems which operate on symbolic levels of representation.

Interconnections between the A projects

Beyond the general assumptions and procedures shared by all projects in Area A as described above and as incorporated in the ISC procedure, there are also connections and mutual dependencies between the individual A projects.

The new methods of incremental context evaluation and specification to be developed in project A3 will result in a novel, rich representation of speech signals. Crucial contextual ingredients of this representation comprise the traditional time-frequency dimensions as well as suitable representations of voice quality and prosody and, conceivably, paralinguistic information. Project A3 will develop a mathematical framework for computing this representation in an incremental, context-dependent manner.

The signal processing procedures developed and implemented in A3 make it possible to capture and express the detailed phonetic knowledge that speakers possess about their language. This knowledge includes not only the language specific implementation of phonetic targets, but also the voice characteristics of individual speakers applied in individual contexts. This phonetic detail will be required in capturing the prosodic correlates of information structure in A1, and the exemplar based representations of phonemes, syllables and words in A2.

A common goal of all projects in Area A for the present phase is to identify detailed prosodic parameters the role of which has not been appreciated by linguistic theory. The reason to believe that such parameters exist are different in the individual projects (coding of information structure in A1, coding of exemplar dimension space in A2, coding of emotion and other speaker specific features in A3). If it were, however, the case that such detailed phonetic knowledge is independently required in many areas of language study, it would challenge the standard models of phonology and phonetics. The consequence of this result would be that the acquisition of such rich phonetic representation would have to be modelled by linguistic models and computational procedures. Here exemplar theory and the analysis-by synthesis procedure come into play.

A central scientific goal of the Area A of the SFB in a longer-term perspective is to provide convincing evidence that speech is not (only) a "noisy vehicle of linguistic content". In developing a model of Incremental Specification in Context for the area of speech we will try to show that the natural medium of the linguistic message (viz. speech) may form an integral and coherent dimension of linguistic representation. The exemplar-based representation as we conceptualize it in Area A, is more than a collection of frequent speech events (like a syllabary, a set of pitch-accents, or an epistemic lexicon). Exemplars result from processes of specification; they are not echoes of frequent, learned, and stored events. The explicit modelling of context which we undertake will account for the possibility that exemplar representations can have a rich linguistic history, reflecting speech events in various contexts, with various prosodic nuances, and coming from various voices. The logic of exemplar representation as formulated in the Incremental Specification in Context program extends to other contextual details, including morphology, syntax and semantics.

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

All projects in B are concerned with lexical and supralexical semantics and examine different aspects thereof such as word formation, 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. The projects in B connect to the overall concerns of the SFB with aspects of context and specification in various ways: they discuss different aspects of context (local vs. global, linguistic vs. non-linguistic, language specific vs. language independent, subject to language change vs. diachronically stable) and different ways of using context in the process of specification.

Lexical and Supra-lexical Semantics

The projects which make up Bereich B focus on problems which are on or near the borderline between lexical and supra-lexical structure.

First, almost all projects are specifically concerned with questions of word formation (e.g. deverbal nouns, prefix verbs). And one feature of word formation involving nominalisation is that it often permits the condensing into a single word of information which would otherwise have to be expressed by a larger phrase or even by a full clause. The literature on nominalization has established that the parsing and interpretation of such words involves the same syntactic and semantic structures as the corresponding phrases and clauses.

Because of this, some of the work proposed by the projects in this area is motivated by the general question how word-internal structure is related to the structure of phrases

and clauses. It has often been taken for granted that lexical and supralexical linguistic structures have little if anything in common. But in recent years there has been a growing conviction that there are important similarities between word structure and the structure of complex phrases. The most radical challenge to the view that there are no significant connections between lexical structure and the structure of phrases and clauses comes from models such as Distributed Morphology (DM), which provides the background for one of the B-projects (B1). But the questions that DM asks are bound to receive non-trivial answers that require closer scrutiny of the properties of word morphology and word meaning. Therefore all the B-projects will have to deal with these questions in some form.

Finally, much of our work is more strongly focussed on the part which words play as parts of larger linguistic units – in the form and meaning of complete clauses and sentences, but also, beyond that, in the structure and interpretation of texts and conversations.

Contexts for Word Sense Determination

One of the general tasks of the SFB will be to develop a better understanding of the variety of kinds and sources of information on which the meanings of words in context can depend, of the nature of these dependencies and of the ways in which dependencies on different sources of contextual information may interact. For the projects in B aspects of these tasks are among their specific goals.

An additional question about context is raised by the projects B2 and B3 and to a somewhat lesser extent also by B5. These projects make heavy use of corpora. The diachronic project B2 is dependent on corpora of surviving texts to discover the meaning and use of words in a period when the languages under investigation had very different properties from the ones they have today. In the computational project B3 corpora are used as resources from which information about word meanings is to be extracted by automated means; and corpora will also be a major source of information in B5. Those for whom corpora are the only or primary source of linguistic information face a problem which does not confront the other B-projects: How much of the contextual information that was available to the intended addressees of texts from a corpus is recoverable for the investigator? For the computational projects – B3 and to some extent B5 – this problem is compounded with another one: How much of the contextual information that can be recovered in principle can be detected by the algorithms that are available for extracting the data that the investigator is after? These problems are perhaps most acute for B3, in which the development of extraction methods which cope with these problems is the central task.

A third type of context-related issue is raised by the project B5. The central aim of this project is to establish an ontology of event types that can serve as a semantic ‘space’ within which the meanings of arbitrary natural language verbs can be ‘located’. One of the basic assumptions which the project shares with many other attempts to develop ontological bases for lexical meaning is that the ontology can be represented in the form of a network or graph in which individual event types are connected by various kinds of edges. We argued above that such ontologies could be seen as providing contextual frames and backgrounds for the semantics of the individual words whose meanings can be located within them. B5 proposes to deal with this global aspect of context dependence explicitly.

Underspecification and Specification

If a word can mean different things in different contexts, then its lexical specification must indicate that this is so. The question is how?

The work on lexical meaning that is planned in the SFB is committed to paying this connection – between lexical specifications and the context-driven disambiguation mechanisms into which they feed – the attention it needs. This commitment is especially prominent in projects B1 and B4. One of the theoretical questions to be pursued in B1 is whether certain lexical ambiguities – more specifically, certain ambiguities of word-forming affixes – must be listed as sets of alternatives in the lexicon or can be explained as effects that are produced by a single (underspecified) lexical meaning when it is inserted in different positions of one or more structural templates.

More generally, in both these projects the problems of lexical meaning and contextual "specification" are treated as two sides of a single coin; and in this they are representative of one of the SFB's general commitments.

More on the individual projects and the connections between them

The goal of B1 (The formation and interpretation of derived nominals) is to investigate the morpho-syntax of word formation patterns as the possible source of the ambiguity found with deverbal nominals, and contrast this to the ambiguity found with non-deverbal nominals. On the one hand, this investigation is to provide a test bed for the general theses of DM in contrast to lexicalist approaches, on the other hand the project also has an important cross-linguistic orientation, insofar as it investigates similar constructions in different languages. This dimension is of special importance for a better understanding of the interactions between form and context, as languages make use of different means and structures for the formation of nominal structure. An important aspect of this project is the attempt to circumvent what is standardly viewed as lexical ambiguity as a result of structural differences.

Project B2 (Funktionsweise und diachrone Entwicklung deverbaler Nominalisierungsverfahren im Französischen und Italienischen) has a strong diachronic orientation. Its aim is comparative in both a diachronic and a synchronic sense, in that it pursues on the one hand the development of certain word formation operations within individual languages over time, while it aims on the other hand to compare the effects of these developments in the different Romance languages as these currently exist. The cross-linguistic variation of word formation mechanisms investigated in this project is thus two-dimensional – between different languages coexisting in time and between different temporal stages of one and the same language. By taking account of the diachronic aspects of word formation processes, B2 brings to the general cross-linguistic concerns of the SFB a dimension of cross-linguistic variation which has gained in importance in recent years.

The goal of project B3 (Disambiguierung von Nominalisierungen in Corpustext) is the extraction of semantic information concerning derived words (especially nominalisations) and the contexts in which they occur. The central scientific challenge of this project is a methodological one: So far, certain methods that have been used at the IMS for extracting semantic information from corpora were based on the assumption that the corpus from which information was being extracted are unambiguous – only if there is sufficient evidence that an occurrence is unambiguous would it actually be used as a site for extraction. In this way one ensures that the extracted information is reliable,

but the method is inefficient in that the number of potential extraction sites which are discarded is very large. Often the real or perceived ambiguity of a sentence is irrelevant to the information that is to be extracted and in those cases the extraction site ought to be used. One of the aims of the project is to develop ways of distinguishing between innocuous and potentially harmful ambiguities (relative to a given extraction task) so that less of the valuable information that is contained in the corpus will be thrown away.

B4 (Lexikalische Information und ihre Entfaltung im Kontext von Wortbildung, Satz und Diskurs) aims, on the one hand, to develop entries for word-forming operators, such as derived noun affixes like German *-ung* and *-er* and verbal prefixes like *nach-*, *unter-* or *mit-*. The other challenge is to tackle, in connection with ambiguous affixes, the question of disambiguation in context. The development of a formal lexical semantics for word formation processes is a central task of this project which will pursue its goals within a DRT-based framework of underspecified representations of lexical information. The project will link the semantic principles relevant for word formation to principles that are operative at the interface between word and sentence and sentence and text. The questions to be clarified in this attempt are how, when and where reference to the different notions of context becomes available.

B5 (Polysemy in a conceptual system) aims at developing and using an ontology as a frameworks for larger scale lexical specification – attuned to the specific purpose of providing syntactic and semantic specifications of French verbs, but also suitable more generally. The project will make use of a mixture of methods for determining the properties and relations that will be constitutive of the ontology. This ontology will provide the semantic anchoring space for the French verbs whose semantics will be described. Its aim is to make use on the one hand of information that can be found in existing dictionaries, using suitable (semi-)automated extraction methods to get at the information one wants. But other electronically available resources, such as corpora and semantic networks, will be accessed as well. One of the theoretically more challenging forms of lexical underspecification concerns the alternations between what are sometimes described as 'literal' and 'non-literal' meaning. For instance, the verb *soutenir* (support) can be used in a material sense in which a pedestal supports the statue that rests on it, but also in an abstract sense in which an argument can support a claim. One task of B5 is to develop ways to represent the lexical meanings of *soutenir* and similar verbs in a domain-independent form which yields their domain-specific meanings when the domain is determined by the context in which the verb occurs.

Interconnections between the B-Projects

Besides the general assumptions that are shared by all the B projects and which we have described above there are also connections between individual B projects. The projects B1 and B2 both investigate word formation patterns from a crosslinguistic perspective. To the synchronic dimension of this problem B2 adds a diachronic one, by investigating how in some languages such patterns have developed over time. The cross-linguistic generalisations that we expect from B1 and B2 will provide important input to the two other projects concerned with word formation, B3 and B4.

Computerised extraction from corpora often brings to light empirical facts that might have escaped the attention of theoretical linguists indefinitely. Thus corpus extraction provides a supplement and correction to traditional linguistic methodology that is seen increasingly as indispensable. In particular we expect that the extraction

methods developed and applied in B3 will reveal data about nominalisations and their linguistic environments that will be important input to the theoretical considerations of B1. Conversely, the insights gained in B1 will guide B3's design of extraction algorithms so that they will unearth those data that are theoretically important. B3 and B4 were originally conceived together, with the intention to optimise cooperation between the (semi-)automated methods of B3 and the slow, painstaking by-hand methods of B4. Furthermore, B4 hopes to profit from the data extracted in B3 in a similar way as B1 and also to help B3 in deciding what kinds of data its algorithms should be able to find.

B4 and B5 share the commitment to a language-neutral characterisation of the class of lexically relevant event types. The projects will coordinate their efforts towards such a characterisation, so that the ontology developed in B5 can be used in extending the existing (currently still quite small) DRT-based lexicon.

Projektbereich C: Noun phrases and context

All three projects in C deal with the specification of grammatical forms in the nominal domain, such as word order in adjective – noun modification patterns (C1: The syntax of nominal modification and its interaction with nominal structure), case marking (C2: Kasus und referentieller Kontext), and agreement marking (C3: Semantische Kongruenz und Determination). For all three projects, the context consists primarily of morpho-syntactic structure, e.g. the specification of an ambiguous adjective is primarily determined by the syntactic structure in which it is found.

We see three main issues involved in this research domain:

1. The focus is on dependencies between properties of noun phrases (or determiner phrases) or certain constituents thereof.

2. The properties at issue are known to depend on a complex of different contextual factors, ranging from lexical information of the constituents (hence very local) to the discourse functions of the whole DP (and hence rather global). A central goal is to identify these different contextual factors and to see how they combine or interact in determining these properties.

3. The contextual factors which are relevant for the determination of these properties are known to vary between natural languages. The C projects aim to throw light on this aspect of variation by comparing the contextual determination mechanisms of the properties they investigate for a small set of different languages that can be investigated in the necessary detail.

With respect to the first two issues, area C will first provide a fine-grained morpho-syntactic analysis of DPs in the languages that the projects are investigating. Second, area C will develop detailed linguistic criteria for disambiguation in the area of noun phrases. We expect that these fine-grained analyses can be used by the computationally oriented projects in the formulation of rules of interpretation, and in exchange we expect feedback from such projects as far as new or problematic data are concerned.

With respect to the third issue we would like to stress that this is standard practice in theoretical linguistics for the following reasons. On the one hand, languages might appear to make use of superficially identical means for the expression of certain notions; but these, when subjected to careful investigation, can be shown to reflect finer distinctions that need to make reference to different (morpho-syntactic but also semantic) representations. Our task here is to determine which of these must be treated as systematic patterns, and which not, keeping in mind that a uniform treatment is not always desirable or viable. On the other hand, as mentioned above, individual languages make use of different grammatical expressions for certain notional categories. Determining the distribution and the finer properties of these is important for the theoretical linguist, as often the correct treatment of these cases has implications for the overall organization of grammar.

Particular research objectives

The three projects in C are all concerned with noun phrases and their features. Project C1 investigates the ambiguity of adjectives modifying noun phrases as well as other constructions e.g. participles that assume a modification role, C2 investigates the parameters that determine direct object case marking, and C3 investigates the contrast between syntactic and semantic agreement.

While all three projects focus on the noun phrase they do so by looking at different linguistic levels and their interaction (syntax, morphology and semantics/pragmatics). A common shared assumption is that there are certain features inherent to the noun and the elements it combines with, but the ways these features interact with functional information varies not only from construction to construction but also from language to language. In other words, lexical items can be underspecified or minimally specified, and all sorts of additional information is provided by functional specification (standardly seen in the form of functional layers that combine with the lexical element). On this view, the flexibility associated with open class items is a result of the combination of minimally specified lexical input with closed class elements/grammatical formatives (e.g. Case, Number, Definiteness). Hence flexibility/ambiguity as well as variation is a result of different combinations of functional structure with the same underspecified lexical item. Concretely, the grammatical reflexes of this can be seen in the syntax and semantics of (adjective) modification, the morpho-syntax of case marking and the syntax and discourse semantics of agreement principles. As we will see below, in addition to this very 'local' contextual input, often one needs to make reference to more global information in order to achieve full specification.

A central issue to be investigated in C1 is what semantic effects are produced by adjectives occurring as noun phrase constituents. This issue has been known for many years to involve a variety of distinct contextual factors. Among the factors that have been identified are: (a) the syntactic position occupied by the adjective within the DP (*the visible stars* vs. *the stars visible*; *the poor boy vs. the boy too poor to buy his own ticket*). The phenomenon is not limited to English, but can be found also in Romance, e.g. French (*un grand joueur* vs. *un joueur grand*, etc.) and other languages, (b) the internal structure of the head noun. For example, the noun phrase *a beautiful dancer* is ambiguous in a way that *a beautiful cat* is not. One possible explanation could be that this relates to the different internal structures involved in the making of *dancer* and *cat*, which provide different 'placement sites' for adjectives, with the effect that different structures may arise for the combination of head noun and adjective, which may project

significantly different meanings. (The position of the adjective within the DP that is (sometimes) revealed at the surface, see (i), may limit the docking sites accessible, cf. *un grand joueur* vs. *un joueur grand*.).

The goals of C2 can be described in similar terms. The specific task of this project is to investigate manifestations in a number of different languages of what is known in the literature as Differentiated Object Marking (DOM): In certain languages direct object DPs may or must be marked in some special way (e.g. by being given overt accusative case marking) under certain conditions, while they need not or may not get such marking under other conditions. Here too, there is extensive documentation in the existing literature that the conditions in question cover a range of different parameters – animacy, definiteness, specificity, topicality – and that languages differ in the sets of parameter values for which they require or admit DOM (and not just in the way in which DOM is morphosyntactically realized).

The task of C3 is to explore a range of cases in English, German, French and Russian where there appears to be a violation of morphological congruence (e.g.: *The family are ...*; *Les enfants ...Elles ...*). Typically, this kind of freedom from morphological congruence creates the possibility of using the morphological features one chooses to express semantic information. (E.g. referring anaphorically to the set of children denoted by *les enfants* is a way to convey that this set consists exclusively of girls; and so on.)

The discussed issues are not new but well established in the linguistic tradition. What is new is that the view of the SFB program, namely the role of context in the specification of linguistic representations is taken up seriously. While traditional analyses of these issues focus on the linguistic contrasts expressed by different grammatical forms (e.g. what is the difference between a case marked direct object and an unmarked one), here these phenomena are used as a testing field for the interaction of the different linguistic components (syntax-semantics-morphology). For instance, in Turkish, a case marked direct object indicates that the object is to be understood as specific; if the direct object is not in the preverbal position the contrast mentioned above is neutralized, i.e. only in certain syntactic environments is this contrast expressed by the grammatical form. Another example concerns the influence of the position of adjectives with respect to the nouns these modify and the part this plays in determining the meaning of the noun-adjective combination (Fine, 1975; Klein, 1981). For instance, the sentence *the visible stars include Aldebaran and Sirius* is ambiguous between a reading under which it refers to the stars that are generally visible and one in which it refers to the stars that happen to be visible now. If the adjective is placed in post-nominal position the example can only refer to the stars that happen to be visible now. A further example of the interaction of different components of the context is semantic agreement, investigated in project C3. Here, the semantics of the referent competes against the syntactic features of the referring phrase. Depending on the particular language and the distance between antecedent and anaphoric term, this competition is differently resolved (the closer the distance the more likely syntactic agreement, the further the distance the more likely semantic agreement). The project investigates the construction and interaction of syntactic and semantic information in the context. Depending on locality and language particular parameters, the syntactic information overrules the semantic one or vice versa. Interestingly, often it is impossible to give clear rules for such phenomena, but only to estimate a certain probability for them – this opens up a connection between the project and the statistical approaches in the computational projects.

The work carried out in C is very much relevant for the research goals of area B, in view of the fact that there is a concentration on nominalization in this Area. B and C together will enrich our understanding of the syntax and semantics of noun phrases, by looking at different aspects of their internal and external structure, as well as contribute to a better understanding of lexical and syntactic properties and the ways in which these interact with one another at the level of interpretation. The links between projects in C and individual projects in areas A and D are mentioned in the relevant section of the respective projects.

The projects in C constitute an integral part of the SFB research in that they focus on the crosslinguistic investigation of contextual parameters, the interactions between different components of the context, and the particular specificational processes in the nominal domain. The research results will enrich the general understanding of noun phrases and contribute important material to other projects in the SFB, as mentioned above.

Projektbereich D: Disambiguation in Context

As mentioned, one of the central problems in computational linguistics is disambiguation. Theoretical linguistics has made great strides in specifying the set of possible readings of natural language phrases and sentences. In most cases, however, there is only a single reading that is valid in a given context. A complete theory of human language must explain how this reading is selected. The projects in Area D investigate how this disambiguation process can be explained and formalized as incremental specification in context.

In Area D, the specification processes in Figure 1 are interpreted as disambiguation processes that are sequentially applied. Examples are part-of-speech tagging, syntactic disambiguation and semantic disambiguation. Each process maps a pair of a less specified representation and an element of contextual information ($r1$ $c3$) into a more specified representation ($r2$). The elements of context used in individual disambiguation steps can be linguistic as well as extralinguistic. For example, the grammatical constraints embodied in part of speech tagging are linguistic whereas semantic and pragmatic disambiguation processes often use extralinguistic context.

Within D's focus on disambiguation, the goals of the SFB outlined in section 1.2 can be restated in terms of five scientific challenges.

1. Types of contextual information: The first challenge is to identify types of contextual information that give rise to explanatory and robust disambiguation. The dominant disambiguation paradigm in computational linguistics in the last decade has been machine learning in which disambiguation methods are trained on richly annotated corpora. This approach has the limitation that labelled data sets are generally small. This means that many linguistic generalizations cannot be learned because of lack of training data. Conversely, many spurious generalizations (e.g., certain lexical dependencies that happen to be frequent in a labelled corpus such as the Wall Street Journal) are introduced into models to improve disambiguation performance, even if they are hard to justify as explanatory. Projects in D will take two approaches to improve on the state of the art: linguistically motivated symbolic grammars (D1, D2 and D3) and unsupervised machine learning (D2, D3, D4 and D5). The challenge is to develop methods in D that overcome the limitations of small training sets and enable robust and explanatory disambiguation.

2. Linguistic vs. extra-linguistic context: The second scientific goal concerns the relationship between linguistic and extralinguistic context. These were identified as two major types of context in section 1.2.2.2. Historically, many linguistic theories have focussed on linguistic context. The interaction between linguistic and extralinguistic context will therefore be an important theme in Area D. In particular, we will investigate how a separation of context into linguistic and extralinguistic components can be formalized.

3. Learnability of contextual information: The third challenge is the acquisition of contextual conditioning information from corpora. Projects D2, D3, D4, and D5 use machine learning algorithms to learn this information from large corpora. Some types of contextual conditioning information are difficult to investigate with traditional linguistic methods. The computational linguistics projects in D can make an important contribution to the SFB by making hypotheses and theories about context experimentally testable on corpora.

4. Use of contextual information: The fourth scientific challenge is how to combine different types of context for disambiguation. The selection of the type of contextual information that is used for disambiguation may in turn be selected contextually. In some cases a weighted combination may be the right approach. The challenge of combining different types of context is of particular relevance if linguistic and extralinguistic contexts are separated. These questions will be mainly investigated in D2, D3, and D5, but also constitute a long-term goal of the SFB (see section 1.2.2.3).

5. Underspecification formalisms: The fifth challenge is the integration of disambiguation routines into underspecification formalisms. Disambiguation routines typically need access to contextual information that may itself be underspecified or locally not available in an underspecified representation. Hence, underspecification formalisms (such as Underspecified DRSs in D3 or Muskens Logical Description Grammar Cimiano et al. 2004 in D1) need to be reconsidered and potentially elaborated so as to meet the requirements of disambiguation. Only on the basis of such extended underspecification formalism can disambiguation routines be formalized and can the purport of different types of linguistic context and extralinguistic context be captured adequately.

In concordance with the last long-term goal, we also want to advance the state of the art in computational linguistics by formalizing linguistic theories on the one hand and probabilistic models of language on the other hand in a way that allows linguistic insights to be incorporated into the "symbolic core" of statistical models. This will improve the performance of NLP applications due to better modelling as well as give new impetus to theoretical linguistic research. The models investigated in D2, D3, D4 and D5 all contribute to this long-term goal.

A brief summary of the individual projects follows.

D1 treats ambiguities as first class citizens. Its aim is to explain why ambiguities do not multiply out during the interpretation process, but rather constrain and eliminate each other. The project will investigate lexical, morphological, syntactical and semantic

ambiguities that are being investigated by other projects in the proposed SFB. It will develop representational devices that allow for incremental specification in contexts that may themselves be underspecified. And it will formulate the principles and interactions that govern this incremental specification process during interpretation.

D2 will improve the disambiguation methodology developed by Johnson, Geman, Canon, Chi, & Riezler (1999) and Riezler, Holloway King, Kaplan, Crouch, Maxwell III, & Johnson (2002) and adapted to German in the DFG-financed project (DLFG) and extend this methodology to choice in generation. Furthermore, D2 will integrate both categorical and statistical information from external resources, such as frequencies of (bi)lexical dependencies computed on very large corpora, collocations, or conceptual hierarchies into the process. Finally, D2 plans to enrich the LFG representations with information on anaphora and, to a smaller extent, information structure and to test to what extent this additional information can be fruitfully exploited for disambiguation.

D3 investigates information extraction (IE) as an application of computational linguistics, in particular its impact on the more general problem of automatic text understanding. To this end, a medium-sized corpus of police reports is converted into underspecified semantic representations. This is achieved in a processing pipeline of consecutive (or incremental) specification steps. On the level of underspecified semantic interpretations, the IE questions can by answered through inferences. The project has two goals: a deeper understanding of the types of inferences required in practical applications like IE, and a constructive proof of the usefulness of logic-based representations in tasks that involve large-scale interpretation of NL text.

D4 aims to develop and implement a statistical disambiguation method for syntactic analyses (parse trees) based on lexicalized probabilistic context-free grammars (PCFGs). Unlike previous approaches, the statistical model employed in this project is modular and comprises two components, a standard unlexicalized PCFG which is trained on a treebank (a manually created collection of parse trees), and a separate set of parameters encoding context knowledge in the form of lexical dependencies, which is extracted from unannotated corpora. In order to overcome sparse data problems, the lexical parameters will be learned from large unparsed corpora using a bootstrapping approach. The model will be applied to English and German.

D5 investigates syntactic disambiguation as a particular type of specification. An ambiguous sentence is processed by a treebank-trained parser. A subset of the readings are extracted from the analysis. The most likely parse is then identified based on contextual knowledge acquired by way of biased learning from (monolingual and bilingual) unannotated corpora. We use the framework of Exemplar Theory for relating the ambiguous sentence to similar contexts in the unannotated corpus. Two similarity measures are considered: a language model as a baseline and a measure based on grammatical dependencies.

1.2.2.5 Stellung des geplanten Sonderforschungsbereichs in seinem weiteren Fachgebiet

The research program described above aims at contributing to the general theoretical and methodological discussion on the need for underspecification and the ways of achieving full specification.

We have already mentioned that the introduction of the concept of underspecification is an important step in dealing with ambiguity for both theoretical and computational linguistics. Underspecification was introduced for two reasons. First, in computational linguistics, the requisite contextual knowledge for deciding between different options is often not available. Nevertheless, if only fully specified representations were available, an informed choice could and would have to be made; and the choice will have to be made at some later processing stage if the information has become available at that point. Second, the presence of underspecification formalisms triggered a change of view also in theoretical linguistics: Underspecified representations are often regarded as cognitively more plausible than collections of fully specified ones. However, this is a perspective that remains incomplete until a story has been told about how fully specified representations are actually derived from underspecified representations, which is the focus of the research in this proposed SFB. This is where the concept of specification described here enters linguistics as the science of a human cognitive capacity.

In particular in phonetics and phonology, underspecification is primarily needed to circumvent overspecified representations, i.e. representations where different choices for the value of a certain feature do not have observable consequences. In morphology, underspecification may lead to a more economical and more systematic organization of grammatical information. In syntax and semantics, the required depth of interpretation may vary between different tasks: Often, the representation formalism (e.g. predicate logic) imposes decisions which are difficult to make and furthermore have no observable consequences for the task at hand (e.g. machine translation). But since in certain cases those decisions do have consequences, they cannot simply be thrown out of the overall formalism.

Underspecification may also serve different purposes, as will become clear in the following subsections on underspecification at the individual linguistic levels. The proposed SFB with its four areas will be an important addition to this picture. As we will see, while the notion of underspecification has been mainly worked out in phonetics/phonology and morphology, its status in syntax is much less clear. Within semantics the concepts of underspecified representations and incremental specification have gained much in clarity and precision through work over the past 15 years (much of which was done at the Universities of Stuttgart and Saarbrücken).

I. Underspecification and Specification in Phonetics and Phonology

Sounds in the languages of the world are represented in terms of acoustic, articulatory and perceptual features, which have a capability of distinguishing meaning – the distinctive features. From the very moment when distinctive features were introduced into linguistic theory, the degree of specification, or rather underspecification, has been subject to rather violent swings of fashion.

In the original Prague School conception of representation, nondistinctive features were universally absent even if their phonetic attributes were present on the surface. Within the system of distinctive sound properties Trubetzkoy (1939) formulated criteria to determine which of the universally possible ‘correlations’ were irrelevant in a system. Such irrelevant properties were PERMANENTLY UNDERSPECIFIED, or simply speaking, nonspecified within the system.

The view of non-specification of certain features was taken over by early generative phonology. Unlike the mono-stratal Prague School phonology, generative phonology in the early stages was strongly derivational in nature. The underlyingly unspecified features were allowed to be filled in by phonological rules during phonological derivation, yielding completely specified surface representations. More pertinent reasons for temporary underspecification arose due to the study of tonal and harmonic systems (Ringen, 1975; Goldsmith, 1976). It was argued that features from the surrounding environment spread to underlyingly toneless syllables and featureless vowels. The question of which features are universally available for insertion into unspecified lexical entries has been explicitly raised within the framework of Lexical Phonology (Kiparsky, 1985). This line of research, taken up most prominently by Diana Archangeli, led to the discovery of features and feature values which show a high degree of inactivity and invisibility in phonological systems, even if they are contrastive within these systems. Kiparsky (1985), and Archangeli (1988) argued that this phonological inertness should be coded by (universal) nonspecification of such features in underlying representations. The amount of internal and external evidence that was collected for what became to be called RADICAL UNDERSPECIFICATION is impressive (cf. Archangeli 1988; Stemberger 1991 1992; Paradis & Prunet 1991).

However, no general principle for determining which features can be universally underspecified and under which conditions these underspecified features (and segments) are inserted into the phonological derivations could be provided. A critical reaction to radical underspecification has been to restrict it to a small core of reliable cases, particularly those cases where an underspecified feature is truly predictable (Clements, 1987; Steriade, 1987; McCarthy & Taub, 1992). Close scrutiny of these solid cases has shown that underspecification can be restricted to nondistinctive feature values. In other words, surface contrast excludes underspecification. This model, called CONTRASTIVE or RESTRICTED UNDERSPECIFICATION, has been used to form a theoretical basis for the type of underspecification required in widely accepted non-linear (autosegmental) phonological and phonetic analyses.

In a seminal paper, Keating (1988) examined phonological theories of underspecification in view of the persistence of nonspecification straight down to the motor planning level. Keating argued that ‘when phonetic rules build trajectories between segments, an unspecified segment will contribute nothing of its own to the trajectory’ (Keating, 1988). Hence, the theory of phonetic underspecification predicts a correlation between phonological nonspecification and phonetic transparency. Work on glottal transparency by Keating and others (cf. Stemberger 1993; Vollmer 1997) and publications on other aspects of transitional coarticulation (cf. Pierrehumbert & Beckman 1988; Cohn 1990 1993; Vollmer 1997; Zsiga 1997) point out that underspecification may persevere into phonetics. This research shows that phonetic data can be used to make inferences about the lack of specification at higher levels of representation.

Phonetic underspecification is assumed to be an important part of the task dynamic model. Byrd (1996) argues that for some types of segments timing has to be lexically specified. When temporal relations are crucial in making phonological contrasts they have to be fully specified. Candidates for such fully specified segments are multiply articulated stops, ejectives, implosives, clicks and other ‘multigesture’ segments. According to current phonetic theory these segments are special in the sense that they are in a way ‘overspecified’.

In the ACOUSTIC models of speech (cf. Stevens 1989 1998; Stevens et al. 1986) it is argued that the articulatory-acoustic-auditory relations are quantal in the sense that the acoustic pattern shows a change from one state to another as the articulatory pattern is varied through a range of values. For example, a fricative consonant radically changes its acoustic (spectral) correlate for the feature [strident] when the position and degree of articulatory stricture gradually moves through a set of values (cf. Stevens 1989). The cases where the acoustic parameter undergoes large changes for relatively small manipulations of the articulatory parameter form the basis of distinctive features and the basis for underspecification. Stevens suggests that a phonetic theory should specify only those areas of the signal when a feature CHANGE is implemented. These areas, to which Stevens refers as LANDMARKS in the acoustic steam, have to be fully specified, the rest of the acoustic characteristics of the signal can be interpolated from the landmark areas. Thus in the speech signal there will be an alternation between narrow 'landmark' regions marked by acoustic events where there are rapid changes, and temporal regions where the acoustic parameters remain relatively steady. The idea behind 'landmark specification theory' would be, that the temporally broad areas of stability can be predicted from the temporally narrow areas of change.

A similar discontinuity, which Stevens has proven for the articulatory-acoustic relationships, has been shown for the acoustic-auditory relationships as well. It has been observed that the measure of the auditory response, obtained through a psychological procedure like a perception experiment, shows a non-monotonic change as an acoustic parameter is manipulated. Libermann, and the group of researchers at the Haskins Laboratories have convincingly argued for the special 'categorical' mode of perception of certain speech like sounds. Their explanation for this perception mode hinges on the relation between perception and articulation (the Motor Theory) and has been refined in the gestural-score task model discussed above. Another explanation for this non-monotonicity has been provided by the perceptual DIVA model of Guenther and Perkell (cf. Guenther et al. 1998; Dogil & Möbius 2001).

The DIVA (for Directions Into Velocities of Articulators) assumes that speech is represented by a set of regions in perceptual space. This follows from the assumption of the model that goals of communication are predominantly perceptual, and thus that the invariant representations of these goals should be specified as regions in the human perceptual space. Interestingly, Guenther has shown that only the directions in the perceptual space have to be invariantly specified. Once the direction that the perceptual parameter is moving towards is learned, its acoustic and articulatory targets can be predicted by a general modelling algorithm (in case of DIVA an unsupervised neural network). The model is reminiscent of the acoustic 'landmark specification theory' in which only the areas of parameter change had to be specified. It is more general, however, because it models a more abstract mode of language representation and because it includes a realistic learning strategy. The relevance of the model for the research programme proposed here lies in the fact that it not only provides an abstract theory of underspecification (directions in perceptual space) but also provides computational means of specifying acoustic and articulatory implementation of this representation.

As mentioned in the previous section, in this SFB the focus will be on the methods of specification in phonetics and prosody. We will extend new theoretical approaches, such as "exemplar theory", to provide a basis for procedures of full specification in phonetics and linguistics. These procedures will facilitate the integration of numerical and symbolic information for speech interpretation.

Phonological representations have traditionally been highly underspecified, while phonetic representations are considered to include all details relevant for speech. As discussed above, in this SFB we propose to formalise the relation between these two representations by a procedure dubbed "Incremental Specification of Speech" (see the description of project area A for details). The core of this procedure is the analysis-by-synthesis loop in which the underspecified lexical representations of phonemes and syllables are matched with the set of stored "exemplars" which include all the language and speaker specific details of the analysed lexical form, (see Figure 3). The matching procedure has a precise computational formulation. One of its primary characteristics is the inclusion of prosodic context in the analysis-by-synthesis phonetic specification procedure. Since prosody is strongly predetermined by syntactic and semantic representation ("information structure") the specification procedure relates to these linguistic levels of representation as well.

II. Underspecification and Specification in Morphology

Underspecification in morphology was introduced to systematically account for instances of syncretism, i.e. ambiguity primarily in inflectional but also in derivational forms. As is well-known, natural languages show a great deal of syncretism that is cases in which the same morphological form is found in different syntactic-semantic contexts. Distinguishing between (systematic) syncretisms and cases of accidental homophony is not always easy. The general question about how such distinctions are to be made and relatedly, how one can be certain of which is correct, is of great importance.

To this end, underspecification is understood as the property that allows syncretisms to be stated systematically. The null hypothesis here is that the morpho-syntactic features with respect to which underspecification applies in morphology are exactly those that are independently motivated in the syntax. Underspecification has been discussed in great detail in the context of inflectional morphology and in connection with the status of paradigms, as well as feature decomposition. Since such questions will not occupy us here we do not review this literature. For further discussion, see Bierwisch (1967); Bobaljik (2002); Stump (2001); Plank (1991); Müller, Gunkel, & Zifonun (2004), the contributions to Müller et al. (2004); Embick & Noyer (2004).[2]

The consequence of underspecification is that there is a competition between different markers for one and the same context. This competition can be resolved by appealing to an ordering of the rules that introduce markers, see e.g. Wurzel (1998); Halle (1994). A more interesting concept, however, is the notion of specificity, of which a number of variants exist e.g. Elsewhere Principle, Blocking Principle, Panini's Principle etc. A version of this is given below:

[2]We note here that while in most frameworks crucial reference to underspecification is being made, its role in Optimality Theory is not so clear. In fact given the expressive power of OT it is not clear whether underspecification is a concept that could be available. Nevertheless, Wunderlich (2004); Müller (2002) and others crucially presuppose underspecification in their treatment of paradigms.

(2) Subset principle

> 'The phonological exponent of a Vocabulary item is inserted into a morpheme...if the item matches all or a subset of the grammatical features specified in the terminal morpheme. Insertion does not take place if the Vocabulary item contains features not present in the morpheme. Where several Vocabulary items meet the conditions for insertion, the item matching the greatest number of features specified in the terminal morpheme must be chosen.'

What is more important for our research goals is the notion of underspecification as it has been used in derivational morphology. As we will see, similar concerns arise there, which raise important questions as far as the distinction between derivation and inflection is concerned. In what follows, we limit ourselves to the debate between (a particular type of) syntactic vs. lexicalist approaches to derivational morphology.

As noted by Beard (1987) among others, in Indo-European there are many affixes that have remarkably unrelated meanings. The English suffix *-er*, for instance, is associated with the agentive nominalization. However, it also marks the comparative form of adjectives as well as instruments. Similarly, *-ing* can appear in a number of environments:

(3) a. smoker, singer

b. fast, faster

c. grinder, opener

(4)	a. The boy is cutting flowers	progressive
	b. He ruined the knife cutting flowers	deverbal adverb
	c. His cutting the flowers dismayed us	gerund
	d. Cutting is for the birds	imperfective nominalization
	e. He brought his cuttings in	resultitative nominalization
	f. The boy cutting the flowers is cute	present participle
	g. a very cutting remark	adjective

Patterns such as the above are a problem for the linguist and the language learner, since there is no single form/meaning correspondence. The question then is: why and how can a given affix create different types of meanings? For Beard, to define this situation as homonymy would allow the number of affixes to explode on purely theoretical grounds. In order to treat this ambiguity he suggested that what we call suffixes are associated with conditions on insertion.

Researchers working within the framework of DM further developed this idea. In this framework, what we see in (3) and (4) are two sides of the same coin: on the syntactic side, the structure as well as the meaning of the constructions are distinct. On the morphological side, however, the same items, *-ing* and *-er*, appear in different contexts. This is only possible if we assume that the items *-ing* and *-er* are underspecified so that they can 'fit' into several different semantic/syntactic contexts. Thus what we need to specify are the conditions on insertion of the particular item,

which would then allow a particular form to appear in a number of different contexts. The obvious question here is whether one would want to assume underspecification for all instances of *-ing* and *-er*.

Presumably the comparative *-er* and the progressive *-ing* cannot be subsumed under this proposal, as they spell out rather different categories than agentive *-er* and nominal *-ing* (assuming that adverbial/adjectival uses of *-ing* are derived from other basic forms). On the other hand, the instrumental *-er* could be seen as being just one item that spells out a nominal category in a particular syntactic context, which differs from that of the agentive nominal. As Levin & Rappaport Hovav (1988) have stressed, the relevant difference is absence vs. presence of event readings and argument structure respectively. Clearly, while for Beard all cases are treated uniformly as instantiations of disjointness of form and meaning, for DM some cases are simply instances of accidental homophony.

There is, however, an alternative take on this issue that is advanced by lexical approaches to derivational morphology (e.g. Lieber 2004). While for DM, the items in themselves have no semantic value, they function simply as the overt realisation of an abstract morpheme signalling nominal, comparative, aspect etc, Lieber holds that semantic features such as [material], [dynamic], [location], [Inferable Eventual Position or State], [Bounded], and [Composed of Individuals], make up, in different combinations and specifications, the distinct semantic 'skeleton' of an affix (and in fact any lexeme) in a compositional fashion. To deal with affixal polysemy, Lieber proposes that polysemy chiefly arises through the underspecification of affixal semantic structures. Hence the ambiguity relates to underspecified representations of the lexical semantics of the individual affixes similar to the underspecified meanings of ambiguous words.

A second important question in this context concerns the existence and distribution of affixes that appear to produce words with closely related meanings, e.g. *-er* and *-ee* or *-ing* and *-ation* in English nominalizations.

For DM, a detailed study of such affixes suggests that most likely the syntactic as well as semantic contexts in which they appear are not identical, hence the conditions for insertion differ. Alternatively, and for the cases of complete identity in meaning, the individual affixes make reference to different sets of roots (i.e. bases). Lieber, on the other hand, argues that such affixes make exactly the same fundamental semantic contribution to their bases. Specifically, *-er* and *-ee* form dynamic nouns: the skeletal contribution of these affixes will be nothing more than the features [+material, dynamic] and an associated "R" argument, that is, the highest reference argument of the semantic features, with additional semantic requirements on the co-indexed arguments, such as "sentient" and "nonvolitional" for the *-ee* affix. See also Plag (2004) and Booij & Lieber (2004).

The issues discussed here pertain to a great extent to questions of the syntax-morphology interface. Clearly, in the domain of word formation the description of syntax-morphology interaction very much depends on the perspective adopted with respect to the ordering of these two components, i.e. the question of whether morphology comes before or after syntax, and whether it should be recognized as an individual level or not (see also Ackema & Neeleman 2004, and Borer 2005 forthcoming).[3]

Chomsky (1970), and much subsequent work recognized the need to distinguish

[3]There are of course a number of non-trivial interactions between morphology and phonology that we are glossing over here, as these do not constitute an integral part of our investigations in the area of word formation.

between formations that are created before syntax (in the lexicon), and those created in syntax. According to the view that word formation is distributed over the lexicon and the syntax, the division of labour is as follows. Idiosyncratic patterns (and possibly non-productive ones) are relegated to the lexicon. Much of the work done on word formation during the early 80's was based on the assumption that there is an independent word formation component, the lexicon, but that its interaction with syntax is severely restricted. It was assumed that the word formation component is ordered prior to D-structure, that is prior to the availability of any syntactic operations. The word formation component and the syntax interact only in one fixed point: the output of the former is the input to the latter.

Under Lieber's view, discussed in this section, it is possible to assume that affixes have double representations in the lexicon, differing only in the presence vs. absence of the features [+/- material] [+/- dynamic] or representations that are unspecified with respect to these features. Depending on how these representations map to or are compatible with the feature constitution of the verbal base we can predict and explain the behavior of underspecified affixes. Syntactic structure does not play a role.

Within DM, word formation takes place in syntax. It is assumed that syntax manipulates (bundles of) abstract features; these bundles are the *morphemes*. These features are provided with phonological content at PF, for the members of the functional vocabulary (**late insertion**; insertion takes place under competition where the most specified form wins, following (2) above). The terminal nodes that are the sites for insertion are fully specified; that is to say, they contain a full complement of syntactico-semantic features. However, the vocabulary items that determine insertion into these positions need not be fully specified, with the result that a single phonological exponent may potentially appear in more than one syntactico-semantic context. In particular, as far as the examples with *-ing*, discussed above, are concerned: we could derive the differences between the participial *-ing* and the gerund nominal *-ing* by arguing that in the former case, *-ing* spells out Aspect, while in the latter case *-ing* spells out a nominal category. On this view, *-ing* appears in two distinct structures, which are fully specified. In the nominal domain, if we want to capture the difference between the gerund and the result nominalization, we would need to argue that the gerund *-ing* spells out a noun that actually embeds a verb, while in the result nominal *-ing* nominalizes a bare root thus capitalizing on the notion of root vs. outer-cycle attachment put forth in Marantz (2001). The intuition here is that the suffix attaches to projections of different levels, essentially an insight developed by Abney (1987), cf. Ackema & Neeleman (2004).[4]

We are of course aware of the fact that the above summarizes just a subset of the debate on this issue, which finds itself in the middle of the discussions on the contribution of bases and of affixes, as well as the status of affixes themselves and the structures in which these occur. Some further semantic aspects of this area will be discussed in the sub-section on semantics. For further discussion the reader is referred to Plag (2004); Ackema & Neeleman (2004); Embick & Noyer (2004); Borer (forthcoming) among others.

In this SFB, we will contrast the two approaches of dealing with ambiguity in morphology. We view this as one major challenge that will contribute to the dialogue between frameworks and to the general debate between lexicalist and syntactic approaches. By this we aim at a clearer understanding of (the limits of) underspecification and most importantly at a more balanced distribution of the causes of

[4]DM shares a number of assumptions with other syntactically driven approaches, but important differences among them of course exist, which will not be reviewed here (see Borer forthcoming, and Alexiadou et al. forthcoming for a comparison).

ambiguity, i.e. structural as opposed to or combined with lexical (in the sense of root/base related), as well as of the division of labor between syntax and morphology and their interaction with meaning.

III. Underspecification and Specification in Syntax

As indicated in the previous sections, underspecification is an essential feature of the way in which information is represented in phonetics-phonology and morphology. In syntax the question whether underspecification plays a comparable role is more controversial. Much current work in theoretical syntax is driven by the persuasion that there is no underspecification in syntax of the kind there is in morphology: underspecified representations at the level of morphology correspond to fully specified representations at the level of syntax, and it is the task of the theoretical linguist to describe the morphology-syntax interface in a way which shows how this is possible and how it works.

This is not to say that syntactic underspecification plays no role whatsoever. For one thing there are cases where morphological underspecification manifests itself at the level of syntax. One example is the double case constraints that obtain for the wh-words of free relatives in German (Groos & van Riemsdijk 1981; Sauerland 1996 and others). These constraints should be in the syntactic representation. This condition is satisfied when the wh-word plays the same case role in relative and main clause, e.g. when it acts in both as nominative, as in (5.a), or as accusative, as in (5.b).

(5) a. Was heute bekannt gegeben wurde, stimmt nicht.

b. Was sie gesagt hat, habe ich nicht verstanden.

c. Was sie gesagt hat, stimmt.

d. *Wem/wer sie das gesagt hat, war verärgert.

However, it is not the actual cases that are assigned by the main and the relative clause which must coincide, but only the forms which overtly realize these cases. Thus (5.c) is fine. But both versions of (5.d), where neither *wer* nor *wem* cover both case requirements, are ungrammatical. Apparently speakers judge the grammaticality of such sentences on the basis of syntactic representations that exploit the case syncretisms that can be found in forms like *was*.

Nevertheless, underspecification in this sense appears to be quite limited in syntax when compared with phonetics, phonology or morphology. It should be stressed, however, that this does not mean that the role of underspecification in syntax is marginal. Underspecification and incremental specification are increasingly seen as crucial aspects of the *computation* of linguistic representations. It is an essential ingredient of many of the language processing systems (see e.g. Collins 1996). In computational linguistics underspecification has become an essential design feature, which today is found, in one form or another, in any competitive NL parser (see Billot & Lang 1989).

Abstractly, the use of underspecified representations and their (possibly stepwise) transformation into more specified representations can be described as follows (see also 1.2.2.1). Both in syntax and semantics underspecification and specification are widely assumed to take the following form: there is a given formalism *R* in which both

underspecified and more specified representations are formulated (cf. the box in Figure 1). One and the same linguistic expression *LE* will in general have many different representations in *R*. The set *R*(*LE*) of representations of *LE* is partially ordered by the relation $>_{LE}$. The fully specified representations of *R*(*LE*) are maximal in *R*(*LE*) with respect to $>_{LE}$. It is usually assumed that each *R*(*LE*) contains at least one fully specified representation and, more specifically, that for each *r* in *R*(*LE*) there is at least one fully specified *r*' in *R*(*LE*) such that $r' \geq_{LE} r$.

We can associate with each *r* in *R*(*LE*) the set *FS*(*r*) of all fully specified representations *r*' of *R*(*LE*) such that $r' \geq_{LE} r$. *FS*(*r*) is the set of full representations into which *r* 'can be expanded'. A further assumption is that *FS*(*r*) fully characterises *r* within *R*(*LE*), in the sense that $r' >_{LE} r$ iff *FS*(*r*') is properly included in *FS*(*r*): The less underspecified a representation is, the smaller the set of fully specified representations into which it can be expanded.

Specification processes are functions that transform underspecified representations *r* from any set *R*(*LE*) into less underspecified representations from *R*(*LE*). In general these processes are non-deterministic – they could lead from a given input representation *r* to any one of a set of output representations {*r*1,...,*rn*}. Moreover, these processes are typically guided by context in that they will select a proper subset – or, ideally, a single element – from the set of potential outputs, if suitable contextual information is available. In an optimal specification theory an unambiguous utterance of an expression *LE* should always be processable in such a way that it yields a fully specified representation, reached from an initial underspecified representation via one or more specification steps.

Some challenges for underspecification accounts of this kind are the following:

1. To define a suitable representation formalism for both specified and underspecified representations.

2. To develop ways of determining, for each linguistic expression *LE*, the set *R*(*LE*) of possible representations of *LE*. In practice, this is usually accomplished by (i) specifying an algorithm which converts expressions *LE* into initial representations and (ii) specifying transformation processes which turn initial representations into less underspecified representations of the same input.

3. To define the partial order for each set *R*(*LE*). This can be done globally, e.g. by defining the relation on the set of all representations of *R*. Alternatively, the transformation processes may provide such a relationship provided they qualify on independent grounds as always reducing underspecification of the representations which they transform.

4. Of particular importance for the SFB is the task of developing context-dependent specification processes. This requires not only their formal characterization – how are the output representations of the process formally related to the input representations? – but also the identification of the contextual information which the given process uses and the way it affects the output. Once contextual parameters are identified, partial specification of the underspecified representation may be required to actually gain access to the relevant context if that context is not locally available.

5. Specification procedures cannot be expected to turn every underspecified representation eventually into a fully specified one; there will always be cases in which the context doesn't provide the information that would be needed for this. Among such cases there are on the one hand those where further processing of the underspecified syntactic representation that has been computed converges on a single semantic interpretation. But there are also cases where a more specific syntactic representation is needed in order that further processing can proceed successfully. In such cases there may be no other option than guessing at the right representation among those compatible with the underspecified representation which the system has managed to compute.

From this last point of view statistical processors (e.g. stochastic parsers, which compute probability distributions over possible syntactic analysis trees) occupy a special position. On the one hand, the output of such a system is seldom if ever truly unambiguous. (It is extremely rare for such systems to assign 1 to one possible analysis tree and 0 to all others.) But on the other hand stochastic parsers provide a better basis for educated guesses at the correct analysis tree because the trees they present as possible analyses (those that receive a probability >0) are ranked according to certain probability weights. Thus, if one representation has a significantly higher weight than all others, then that may be a good reason to guess it to be the intended analysis.

In this SFB, we will apply the scheme of syntactic underspecification just outlined to specific processes. Syntactically underspecified representations are constructed for interpretation by a template matching process in information extraction, a module for lexical acquisition, and specialized machine learning algorithms.

IV. Underspecification and Specification in Semantics

The past two decades have seen a considerable quantity of work in the area of semantic underspecification. A number of different semantic underspecification formalisms have been developed, generally, if not in all details, with an eye to solving the same problems. We would like to mention in particular the work on Quasi Logical Form (QLF) which began to emanate from SRI International at Cambridge in the late eighties (Alshawi & Crouch, 1992) as well as research at the Universities of Saarbrücken (Radical Underspecification, Pinkal 1996) and Stuttgart (Underspecified Discourse Representation Theory, Reyle 1993) that is still continuing today. Common to these approaches is the commitment to a model-theoretic semantics for the underspecified as well as the full representations. An impression of the influence of the latter two approaches can be gleaned from (van Deemter, 1996), still one of the representative surveys of trends in semantic underspecification today. (Pinkal, 1996) was the first effort to treat semantic and syntactic underspecification within one and the same formalism. (More on this issue in Section V below).

Besides a proper model-theoretic semantics there are other general features that are also desirable for underspecification formalisms. One of these is *closure under specification mechanisms* (Ebert, 2005). What makes underspecification formalisms useful in applications are algorithms for turning underspecified representations into full representations (or representations of a higher degree of specification than the input) and in the approaches mentioned special efforts have been made to provide such algorithms

as part of the overall package. It is a natural (in fact, a quasi-inevitable) constraint on formalisms that come with such algorithms that the representations which these algorithms derive belong to the formalism as well.

A last, but very important desideratum for semantic underspecification formalisms is that they not only provide formal definitions of inference relations between underspecified representations – such relations are a concomitant, and to some they are the very point of a model-theoretic semantics –, but also algorithmic implementations of these relations. (In UDRT providing such 'inference engines' was a central part of the motivation from the beginning.) Inference engines for underspecified representations are especially important because semantic disambiguation often involves logical deduction from the semantic information that has been obtained so far – that is, in underspecification formalisms, the information represented by the underspecified representation that has so far been computed. The obvious way in which one would want to proceed in such cases is to draw inferences directly from this underspecified representation.

The method sketched in the previous paragraph may suggest a paradox: By drawing inferences from the underspecified representation at hand one is to arrive somehow at a modification of this representation into a less underspecified and thus more informative one. How could this be? The answer to this question is complex, and what we have to say about it here is presumably only one part of a complete answer. In many situations semantic specification of a given underspecified representation *r* takes the following form. One assumes one of the formally possible specifications of *r* and derives from this specification *r*', together with additional information C provided by the context of the sentence (or other expression) for which the representation *r* has been computed, a logical contradiction, or perhaps some other kind of inconsistency. On the strength of this derivation, *r*' can be discarded as a possible specification of *r* in the given context and *r* may therefore be transformed into a more specific representation *r*'' which excludes *r*' as an option. It is often possible in such cases to express the difference between *r* and *r*' by a simple condition *D*, which in conjunction with *r* yields the stronger *r*'. Under those conditions the simplest and most efficient way to arrive at the logical or circumstantial inconsistency would be to derive it from the conjunction of *r*, *C* and *D*. But of course this can be done only by an inference mechanism that can make use of underspecified representations as premises.

This method for resolving underspecification appears to be especially effective for the elimination of lexical ambiguities. For this reason it is of special interest for the SFB, in which lexical ambiguity and its resolution are a central theme to which several projects are devoted. We note in this connection that we perceive a discrepancy – at least this is true in relation to our own work in this area – between the enormous importance of lexical ambiguity for the interpretation of language (irrespective of how it is used) and the comparatively little attention it has had in the development of underspecification formalisms and their applications. This is one point where we see the SFB as presenting us with a challenge which will give to our ongoing work new impulses and a new direction.

As with all types of non-lexical ambiguity, an underspecification treatment of lexical ambiguities requires two things: (a) we need a suitable form of underspecified representation for them (to be precise: we need a suitable form for each case of lexical ambiguity, with no certainty *a priori* that the same forms will do for all cases); (b) we need rules or algorithms for resolving or reducing these underspecified representational forms. The first of these requirements has a direct bearing on lexical semantics in the

narrower sense of ‘formal semantic lexicography’ (i.e. on that part of formal semantics which is concerned specifically with the design of formats for lexical entries and the formulation of entries in those formats).

So far we have, in our few explorations of ways to use underspecification in the treatment of lexical ambiguity, made use of the very simple assumption that the lexical entry of an ambiguous word merely lists its different meanings as a kind of disjunction. Even when we operate on this simplifying assumption about the semantic representation of lexical ambiguity there is a range of non-trivial questions about ambiguity resolution that we have to confront. Many of these questions arise in connection with the numerous ways in which the ambiguous words in a sentence can interact with other contributors to the meaning of the sentence, and which may bring their disambiguation about. Of special interest are interactions where the other contributor – another word or some aspect of sentence structure – is ambiguous as well.

Even the probably most common source of lexical disambiguation, violation of selection restrictions, the logical form of which seems elementary, poses problems of considerable magnitude in practice. Let us focus on verbs, which are the category of words most often treated as the paradigm in discussions of selection restrictions. A verb typically comes with selectional restrictions that are associated with each of its different argument positions, and phrases filling these positions must have lexical heads that are compatible with these restrictions. The compatibility requirement can either lead to disambiguation of the noun or of the verb or of both. To mention just a couple of examples: in *die Absperrung abreißen*, *Absperrung* must be understood as referring to a physical object (a fence or similar), while in *die Absperrung durchführen*, *Absperrung* must refer to an event; *einstellen* has the meaning of ‘abrogate’ in the combination *ein Verfahren einstellen*, and that of ‘appoint’ in *eine Sekretärin einstellen*; and in *die Wurzel ziehen* or *Der Hahn läuft* we see how combinations of an ambiguous verb and an ambiguous noun may involve a kind of meaning ‘coordination’ – one meaning of the noun goes together with one use of the verb – with the result of two possible meanings for the combination rather than four.

An algorithm capable of detecting violations of selection restrictions must be able to verify whether a given noun, or particular reading of an ambiguous noun, is consistent with the selection restrictions for the argument position that is occupied by the phrase it heads. It is widely assumed that this can and should be done on the basis of concept ‘hierarchies’ whose nodes provide on the one hand concepts in terms of which the selectional restrictions can be formulated and on the other hand the means for identifying the meanings of nouns. The links between the nodes of such a hierarchy capture the logical relations between them and can – so the assumption goes – be used in a direct way to determine whether a noun whose meaning is identified within the hierarchy is compatible with selectional restrictions that have been given a hierarchy-based definition. But in actual practice implementation of this idea has proved hard, even when vocabularies are artificially restricted.

But violations of selection restrictions is just one source of disambiguation among countless others. Indeterminacy of lexical meaning and its contextual resolution comes in many different guises. The differences manifest themselves both at the level of lexical representation – i.e. in form or content of the lexical entry – and at that of the principles of disambiguation in context. We will try to give a glimpse of the variety and complexity of these issues by means of a few further examples, chosen with the intent that they serve to illustrate several points at once.

First consider once more the noun *Absperrung*. As noted above, it is ambiguous between an object reading and an event reading, and moreover it can be used also to denote the state of affairs of something being closed off. One question relating to this and other deverbal nouns is how the lexical entry should represent their ambiguity, and another, related question is where their ambiguity comes from (see also the section on *Underspecification and Specification in Morphology*). As to the second question, the answer is clearly: from the *-ung* operation which turns *absperren* into *Absperrung*. For its range of possible denotations is something that *Absperrung* shares with other *-ung* nouns, but not with the verb *absperren* from which it derives. What is needed here, therefore, is a characterization of the *-ung* formation as an ambiguous or underspecified operation from verb meanings to noun meanings. (We believe that this operation can be defined in such a way that the semantic descriptions for *-ung* nouns, which result when it is applied to particular verbs capture their ambiguity by means of a natural form of underspecification, but cannot go into details here; see the discussion in the B projects).

Next consider the combination of *Absperrung* with a prenominal adjective, as in *heutige Absperrung*, *unterbrochene Absperrung*, *hölzerne Absperrung*, *abgeschlossene Absperrung*, etc. In the presence of *heutig* and *unterbrochen*, *Absperrung* is still ambiguous, in the presence of *hölzern* or *abgeschlossen* it is no longer. But in each of these four cases two quite different processes of meaning determination are involved. In those cases in which *Absperrung* is interpreted as an event, the process involves the syntax and semantics of the underlying verb *absperren* and in that connection, morphosyntactic as well as semantic considerations; if *Absperrung* is interpreted as object, then it suffices to apply the adjective directly to its referent, just as in other cases of determining the meaning of combinations of pronominal adjectives and nouns, in which the noun is not derived from a verb.

It should also be noted that here as with other adjective-noun combinations the contribution of the adjective may vary with the interpretation of the noun, and such combinations will be looked at closely in the proposed SFB. For instance, in the event reading of *heutige Absperrung* the event must take place today, but when *Absperrung* is interpreted as referring to an object the existence of this object must overlap today, but may extend over a longer period. Similarly, in *unterbrochene Absperrung*, *unterbrochen* will either refer to a temporal interruption of the closing off event, or to a spatial or physical interruption (a break) in the fence; and so on. In these processes the polysemy of the adjective interacts with the ambiguity of the noun; and as we just saw, in some of them this interaction is mediated by a morpho-syntactic process. We finally note that in those cases in which the adjective disambiguates the noun, as in *hölzerne Absperrung* or *abgeschlossene Absperrung*, ambiguity resolution is a very ‘local’ process in which the adjective which serves as ‘context’ for the disambiguation of the noun is its phrase mate in the compound.

In the next examples we focus on a couple of sentences containing *heutige Absperrung* as a constituent.

(6) a. Um halb acht haben sie die heutige Absperrung erledigt.

b. Um halb acht haben sie die heutige Absperrung fertiggestellt.

We note the following points. First, in both (6a) and (6b) *Absperrung* is now disambiguated by the verb, in (6a) to the event reading and in (6b) to the object reading. In both cases disambiguation rests on violation of selection and thus is one of the ‘easy cases’ discussed above.

Second, (6a,b) introduce some new semantic indeterminacies. These are of a different sort, all relating to the phrase *halb acht*. Without further context this phrase is underdetermined in two respects: (i) Is the time it denotes in the morning or the evening? (ii) Which day does this time belong to? With these indeterminacies comes a third one: (iii) Is the referent of *halb acht* situated before or after the utterance time? In order to resolve these 'ambiguities' we have to ascend to the wider context in which the sentences in (6a) and (6b) are used. Suppose for instance that either (6a) or (6b) occurs as part of a discourse in which it is preceded by one of the following two sentence combinations:

(7) a. Seit drei Stunden kann keiner mehr in das Gebäude rein. Die Handwerker waren vor dem Frühstück wieder fort.

b. Die Handwerker sind noch nicht fertig, kommen aber nach dem Tee wieder.

It seems intuitively clear that when either sentence in (6) is used as follow-up to (7a), the referent of *halb acht* is 7.30 a.m. of the day on which the discourse is uttered and that this time is before the utterance time; and further that when the sentences occur as follow-up to (7b) *halb acht* denotes 7.30 p.m. and follows the utterance time. The inferences needed to make such identifications, when they are made fully explicit, they usually reveal the need for additional premises that reflect intuitively obvious aspects of lexical knowledge or world knowledge, which human interpreters tend to apply in the course of making sense of discourse without even being aware of them. Consider the combination of (7a) and (6a). To mention just some aspects involved in arriving at the conclusion that *halb acht* denotes 7.30 a.m. in this case: (i) knowledge about clock times and their names, (ii) knowledge about what time of the day people (normally) have breakfast, (iii) linking *sie* in (6a) with *die Handwerker* in (7a) (ii) linking *die heutige Absperrung* in (6a) to *das Gebäude* in (7a). But the most difficult part for any formal treatment of an inference-based interpretation like that of *halb acht* is a way to handle the default inference that the job the workers finished was completed before the time when they left. It is in particular for such reasons that the formal reconstruction of inference-based interpretation in a discourse context is the very difficult and complex task it is.

Thirdly and finally, there is one further ambiguity in the sentences in (6) which gets resolved in the contexts provided by (7a) and (7b). This is the interpretation of the present perfect. For reasons which are now quite well understood, but which we do not present here (see Alexiadou 2003; Reyle et al.), prospective uses of the present perfect in German – like those in (6a,b) in the context of (7b), which entail that 7.30 p.m. follows the utterance time – can only be interpreted as describing result states. For instance, the meaning of (6a) in this context is that by 7.30 p.m. the fencing off will have been completed, though the actual time of completion could have been earlier. This is not so in the context of (7a) which implies that the referent of *halb acht* precedes the utterance time. In this case the time of the adverb, i.e. 7.30 a.m. is the time *at* which the fencing job was completed. These last observations indicate that the semantic contributions of tense forms like the German present perfect are semantically underspecified by themselves and that their specification typically comes from the sentence or discourse context in which they are instantiated; but that the specification of the underspecified information they contribute is based on complex principles that are particular to just one part of grammar and lexicon.

V. Underspecification and Specification in Syntax + Semantics

According to the most prominent current assumptions in both syntax and semantics, underspecification at either level makes heavy use of underspecification of constituent structure. This formal similarity suggests that underspecification-based systems for integrated processing of syntax and semantics could be designed in such a way that syntactic and semantic underspecification are 'aligned' as much as possible: underspecified syntactic representations should be transducible directly into underspecified semantic representations. Clearly this feature is dictated by considerations of efficiency.[5] Alignment also allows for the direct computation of underspecified semantic representations in those cases where no resolution of the remaining ambiguities is ever called for, or where it is desirable to leave the decision whether they should be reduced to a late state of processing – e.g. because the validity of a certain inference from the representation crucially depends on how its remaining ambiguities (or some of them) are resolved.

This last consideration – postponing disambiguation decisions until there is a good reason for seeing disambiguation as necessary – presents an additional complication for systems with a 'pipeline architecture', which first compute, perhaps in several stages, a syntactic representation and then, on the basis of the syntactic representation thus obtained, a semantic representation, where the second computation too may involve a number of successive stages. Typically processing of this sort involves not only the computation of new representations, but also the discarding of old ones, and thus of information which the old representations encoded but the new ones do not. This is the way in which many systems handle syntactic 'surface information' – once it has done its work in the computation of content it is no longer needed and is thrown away accordingly. Notoriously, this is a feature of theoretical or practical design that is supported by a wealth of psycholinguistic evidence; postponing disambiguation decisions, however, can come into conflict with such a general modus operandi. For it may well be that the disambiguation which proves to be desirable only at a point far down the chain of successive processing stages requires information that while present and accessible at some early stage, has been eliminated in the meantime.

In order to be able to cope with this problem the processing system has to be designed in such a way that it retains (in an economical and yet readily accessible form) just that information which might still be required later on, when the need arises for a disambiguation that a system operating without underspecification would have carried through at an earlier point; for if the information is still accessible at this later point, then the disambiguation operation can still be executed then and there. The design of such systems is a subtle matter, as it requires making predictions about precisely which parts of the information that is up for discarding might still be needed and should therefore be kept. An example of such a system is the one that was developed in the context of Verbmobil (Schiehlen et al., 2000) in which the representations (the Verbmobil Interface Terms) while in many respects like UDRSs, are defined for the very purpose of retaining information that allows for late disambiguation without backtracking.

[5]The alternative would be to first compute the complete set of fully specified syntactic representations compatible with a given syntactic representation, then compute the corresponding semantic representation for each of these, and finally to try and find an underspecified semantic representation which captures just the range of semantic representations that have been computed in this way.

Deutsche Zusammenfassung.

Ausgehend von der Beobachtung, dass Ambiguitäten auf jeder Ebene der linguistischen Analyse gegenwärtig sind, setzt sich der angestrebte SFB das Ziel, ein besseres Verständnis der Mechanismen zu erlangen, die es ermöglichen, Ambiguitäten zu kontrollieren bzw. aufzulösen. Wir nehmen an, dass Ambiguität die Folge von Unterspezifikation ist, und dass demnach Disambiguierung einen Prozess der Ausspezifizierung eines unterspezifizierten Inputs darstellt. Eine Vielzahl linguistischer Prozesse lassen sich als solche Prozesse erfassen (Sprachwahrnehmung und -erkennung, morphologische, syntaktische und semantische Disambiguierung, etc.) und können daher unter diesem Gesichtspunkt im SFB erforscht werden.

Wir verstehen unter einem Spezifikationsprozess jeden linguistischen Prozess, der eine linguistische Repräsentation in eine von mehreren stärker spezifizierten Repräsentationen transformieren kann, wobei der Kontext die notwendige Information liefert, eine dieser stärker spezifizierten Repräsentationen auszuwählen. Es ist also entscheidend, dass Spezifikationsprozesse immer in einem bestimmten Kontext lokalisiert sind, der Beschränkungen und auslösende Bedingungen liefert. Spezifikationsprozesse verbinden zwei Repräsentationstypen: jene, die in irgendeiner Weise unterspezifiziert sind und solche, die stärker spezifiziert sind.

Das Forschungsprogramm lässt sich durch die folgenden zwei Fragen charakterisieren: (i) was ist die Natur der Transformation (inkrementeller Spezifikationsprozess) unterspezifizierter in stärker spezifizierte Repräsentationen? Und (ii) welche Rolle spielt der Kontext in diesem Prozess, welche Art von Information hält er bereit und zu welchem Zeitpunkt wird er relevant?

Was unser Verständnis des Begriffs Kontext anbelangt, so sehen wir vier Aspekte, die für die Untersuchungen in diesem SFB relevant sind: linguistische vs. nicht-linguistische Kontexte, lokale vs. globale Kontexte, dynamische vs. nicht-dynamische Kontexte sowie die sprachübergreifende (In)stabilität von Kontexten.

Wir sehen in dem SFB mit seinen zahlreichen parallelen Untersuchungen zur Spezifikation im Kontext eine einmalige Gelegenheit, zu einem besseren allgemeinen Verständnis davon zu gelangen, wie genau Spezifikation innerhalb von Kontexten und in einzelnen Sprachen von statten geht. Wir denken außerdem, dass ein erfolgreicher Austausch zwischen theoretischen Linguisten und Computerlinguisten nicht nur zu einem besseren Verständnis inkrementeller Spezifikation in Sprache führen wird, sondern dass dadurch auch eine engere Zusammenarbeit bei den Methoden und Ergebnissen dieser beiden linguistischen Teilbereiche erreicht werden wird.

Langfristige Forschungsziele

Langfristig wird unsere Forschung auf die folgenden, allgemeinen theoretischen Fragen Antworten geben:

1. Was ist die Beziehung zwischen unterspezifizierten und stärker/voll spezifizierten Repräsentationen? Wie gelangen wir von einer unterspezifizierten Repräsentation als Ausgangspunkt zu einer voll spezifizierten Repräsentation? Gibt es auch den umgekehrten Prozess (d.h. dass Merkmale aus voll spezifizierten Formen extrahiert werden) und welchen Bedingungen unterliegt er?

2. Welche Rolle spielen die verschiedenen linguistischen Ebenen im Spezifikationsprozess? Was ist die Beziehung zwischen den unterspezifizierten Repräsentationen einer Ebene und den voll spezifizierten Repräsentationen auf der nächsten Ebene (siehe das Pipeline Modell)? Falls auf einer Ebene mehr als eine spezifizierte Repräsentation existiert, was ist die Beziehung zwischen diesen Repräsentationen?

3. Welchen Einfluss haben kontextuelle Faktoren auf die Spezifikation linguistisch relevanter Eigenschaften linguistischer Ausdrücke und deren Repräsentationen? Wie interagieren die verschiedenen kontextuellen Faktoren, die am selben Spezifikationsproblem teilhaben, miteinander: greifen diese Faktoren in einer bestimmten Reihenfolge, oder müssen sie erst miteinander verschmolzen werden, um dann als Einheit zu agieren? Wenn die Faktoren miteinander in Konflikt stehen, weil sie auf zwei miteinander inkompatible Spezifiktionen hinauslaufen, kann dann ein Faktor den anderen außer Kraft setzen? Was sind die zugrundeliegenden Prinzipien, die dies regeln?

4. Was ist die Bandbreite von Spezifikationsproblemen, für die ein spezifisches Stück Kontextinformation (oder ein Kontexttyp) relevant sein kann, und für wie viele Spezifikationsprobleme, die in einer Äußerung auftreten, (und für welche ihrer Kombinationen) kann dieselbe Kontextinformation gleichzeitig relevant sein?

5. Was ist der allgemeine Nutzen kontextabhängiger Spezifikation? Ist er domänenübergreifend und sprachübergreifend derselbe? Oder ist er von der spezifischen Domäne / Sprache abhängig? Gibt es Fälle, in denen keinerlei Spezifikation stattfindet? Naturgemäß können diese Fragen erst dann eine adäquate Antwort finden, wenn wir die Natur der Prinzipien, die für die Spezifikationsprozesse in den verschiedenen Domänen notwendig sind, geklärt haben.

Die explizite Modellierung des Sprachkontexts stellt ein weiteres herausragendes Merkmal des hier vorgeschlagenen Vorhabens dar. Die Projekte im Bereich A werden der Tatsache Rechnung tragen, dass linguistische Repräsentationen eine reiche phonetische Geschichte haben, die Sprachereignisse in verschiedenen Kontexten – mit unterschiedlichen prosodischen Nuancen und produziert von unterschiedlichen Stimmen – widerspiegelt. Ein korrektes Sprachmodell muss in der Lage sein, die Interaktion spezifischer phonetischer Details mit generellen linguistischen Strukturen zu beschreiben. So wie der vorliegende SFB aufgestellt ist, bildet er einen viel versprechenden Rahmen für dieses erstrebenswerte, jedoch schwer fassbare Ziel.

Abschließend ist es auf lange Sicht unser Ziel, linguistische Information in die statistische Modellierung von Sprache zu integrieren. Die Hauptanwendung statistischer Modelle ist die kontextabhängige Spezifikation. Momentan verwenden viele dieser Modelle keinerlei linguistische Erkenntnisse (mit der erwähnenswerten Ausnahme der Exemplartheorie). Unser langfristiges Ziel ist es, den Stand der Technik in der Computerlinguistik durch die Formalisierung linguistischer Theorien einerseits und probabilistischer Modelle von Sprache andererseits, dahingehend voranzutreiben, dass es möglich wird, linguistische Erkenntnisse in das symbolic core statistischer Modelle zu integrieren. Die dadurch verbesserte Fähigkeit zur Modellierung könnte sowohl die Leistung von NLP- Anwendungen erhöhen als auch neuen Auftrieb für die Forschung in der theoretischen Linguistik schaffen.

Das Forschungsprogramm teilt sich in vier Bereiche auf (A–D). Diese Teilbereiche

behandeln jeweils die Arten von kontextueller Information und Spezifikationsprozessen, die in ihren enger gefassten Forschungsbereichen relevant sind. Im Besonderen werden die Forschungsbereiche folgende Punkte behandeln: die Rolle der Unterspezifikation bei der Ambiguitätsauflösung, die Natur der zu spezifizierenden Elemente, die Art der zu spezifizierenden Eigenschaften, die Natur der involvierten Kontextinformation, die Art, in welcher diese Information benutzt wird, die Art des involvierten Kontexts und die Interaktion der verschiedenen kontextuellen Parameter beim Prozess der Spezifikation.

Projektbereich A: Speech, Prosodie und exemplare Repräsentation

Im Bereich A entsprechen die unterspezifizierten Repräsentationen den durch unterspezifizierte distinktive Merkmale und distinktive prosodische Tunes kodierten phonologischen/prosodischen Repräsentationen, wohingegen die voll spezifizierten Repräsentationen den voll spezifizierten Exemplaren linguistischer Ausdrücke wie Phoneme, Silben und Wörter entsprechen. Die voll spezifizierten Repräsentationen akkumulieren detailliertes phonetisches Wissen, welches Sprecher über die linguistischen Ausdrücke ihrer Sprache besitzen; dies beinhaltet auch sprecherspezifische Stimmcharakteristika. Ein solches Wissen kann nicht mit den gewöhnlichen kategorialen Methoden der linguistischen Phonetik und Phonologie modelliert werden, da es nicht rein grammatischer, sondern eher stochastischer Natur ist und durch Generalisieren einer großen Zahl von Tokens erworben wird. Langfristig sollen die Projekte in diesem Bereich eine einzelne Prozedur entwickeln, die unterspezifizierte und voll spezifizierte Repräsentationen berechnet.

Projektbereich B: Bedeutung und Disambiguierung an der Schnittstelle von Wörtern und Phrasen

Alle Projekte in B beschäftigen sich mit lexikalischer und supra-lexikalischer Semantik; sie untersuchen dabei verschiedene Aspekte wie Wortbildung, die Ähnlichkeiten zwischen wortinterner Struktur und Phrasenstruktur sowie die Art der lexikalischen Information, aus der die Bedeutung von Phrasen und Sätzen gebildet wird. Die Projekte in B stehen mit dem globalen Forschungsprogramm des SFB im Hinblick auf Kontext und Spezifikation in vielfacher Hinsicht in Verbindung: sie diskutieren unterschiedliche Aspekte von Kontext (lokal vs. global, linguistisch vs. nicht-linguistisch, einzelsprachspezifisch vs. sprachübergreifend, dem Sprachwandel unterliegend vs. diachron stabil) und unterschiedliche Wege, wie Kontext im Spezifikationsprozess genutzt werden kann.

Projektbereich C: Nominalphrasen und Kontext

Alle drei Projekte in C behandeln die Spezifikation grammatischer Formen in der nominalen Domäne, wie zum Beispiel Wortstellungsmuster bei Adjektiv-Nomen Modifikation, Kasusmarkierung und Kongruenzmarkierung. Bei allen drei Projekten bestimmt hauptsächlich die morphosyntaktische Struktur den relevanten Kontext.

Projektbereich D: Disambiguierung im Kontext

Disambiguierung ist eines der zentralen Probleme der Computerlinguistik. Zwar hat die theoretische Linguistik mit zunehmend großem Erfolg die Menge möglicher Lesarten spezifiziert, die eine natürlichsprachliche Phrase oder ein Satz haben kann. Oftmals ist in

einem gegebenen Kontext jedoch nur eine dieser Lesarten relevant. Eine umfassende Theorie der menschlichen Sprache muss erklären, wie diese Lesart ausgewählt wird. Die Projekte in Bereich D untersuchen, wie dieser Disambiguierungsprozess als inkrementelle Spezifikation in einem Kontext erklärt und formalisiert werden kann.

Bibliography

Abney, S., 1987. *The English noun phrase in its sentential aspect*. Ph.D. dissertation, MIT.

Ackema, P., Neeleman, A., 2004. *Beyond morphology: Interface conditions on word formation*. Oxford University Press, Oxford.

Alexiadou, A. (ed.), 2003. *Perfect explorations*. Mouton de Gruyter, Berlin.

Alexiadou, A., Haegeman, L., Stavrou, M., forthcoming. *Issues in the morpho-syntax of noun phrases (Noun phrases in generative grammar)*. Mouton de Gruyter.

Alexiadou, A., Müller, G., 2005. *Introductory remarks on underspecification in morphology and syntax*. DGfS, Köln.

Alshawi, H., Crouch, R., 1992. *Monotonic semantic interpretation*. In: Proceedings 30th Annual Meeting of the Association for Computational Linguistics. 32–38.

Archangeli, D., 1988. Aspects of underspecification theory. *Phonology* **5**, 183–207.

Bateman, J. A., 1995. On the relationship between ontology construction and natural language: a socio-semiotic view. *International Journal of Human-Computer Studies* **43**, 929–944.

Beard, R., 1987. Morpheme order in a lexeme-morpheme based morphology. *Lingua* **72**, 1–44.

Bergamaschi, S., Castano, S., De Capitani di Vimercati, S. Montanari, S., Vincini, M., 1998. *An intelligent approach to information integration.* In: N. Guarino (ed.), *Formal ontology in information systems*. IOS Press.

Bierwisch, M., 1967. *Syntactic features in morphology: General problems of so called pronominal inflection in German.* In: To honour Roman Jakobson. Mouton, The Hague, 239–270.

Billot, S., Lang, B., 1989. *The structure of shared forests in ambiguous parsing*. In: Proceedings of the 27th Annual Meeting of the ACL, University of British Columbia. Vancouver, B.C., Canada.

Bobaljik, J., 2002. *Syncretism without paradigms: Remarks on Williams 1981, 1994*. In: G. Booij, J. van Marled (eds.), *Yearbook of morphology 2001*. Kluwer, Dordrecht, 53–85.

Booij, G., Lieber, R., 2004. On the paradigmatic natur of affixal semantics in English and Dutch. *Linguistics* **42**, 327–357.

Borer, H., 2005. *Structuring sense. Vol. I: In name only*. Oxford University Press, Oxford.

Borer, H., forthcoming. *Structuring sense. Vol. III: Taking form*. Oxford University Press, Oxford.

Byrd, D., 1996. A phase window framework for articulatory timing. *Phonology* **13**, 139–169.

Casati, R., Varzi, A. (eds.), 1996. *Events*. Dartmouth, Aldershots, USA.

Casati, R., Varzi, A., 1997. *Spatial entities*. In: O. Stock (ed.), *Spatial and temporal reasoning*. Kluwer, Dordrecht.

Chomsky, N., 1970. *Remarks on Nominalization*. In: R. Jacobs, P. Rosenbaum (eds.), *Readings in English Transformational Grammar*. Ginn and Company, Waltham, Mass., 184–221.

Cimiano, P., Reyle, U., Šarić, J., 2004. Ontology-driven discourse analysis for information extraction. *Special Issue of the Data and Knowledge Engineering Journal.*

Clements, G. N., 1987. *Phonological feature representation and the description of intrusive stops*. In: A. Bosch, B. Need, E. Schiller (eds.), *CLS 23: Parasession on autosegmental and metrical phonoilogy*. Chicago: CLS, 29–50.

Cohn, A., 1990. Phonetic and phonological rules of nazalization. *UCLA Working Papers in Phonetics* **76**.

Cohn, A., 1993. Nasalisation in English. *Phonology* **10**, 43–81.

Collins, M., 1996. *A new statistical parser based on bigram lexical dependencies.* In: Proceedings of the 34th Annual Meeting of the ACL, University of California. Santa Cruz, Cal., electronically available at `http://xxx.lanl.gov/abs/cmp-lg/9605012`.

van Deemter, K. und Peters, S. (ed.), 1996. *Semantic ambiguity and underspecification*. CSLI Publications, Stanford.

Dogil, G., Möbius, B., 2001. *Towards a model of target oriented production of prosody*. In: Proceedings of the European Conference on Speech Communication and Technology. Vol. 1. Aalborg, Denmark, 665–668.

Ebert, C., 2005. *Formal investigations of underspecified representations*. Ph.D. dissertation, King's College, London.

Embick, D., Noyer, R., 2004. *Distributed morphology and the syntax/morphology interface*. In: G. Ramchand, C. Reiss (eds.), *To appear in: The Oxford handbook of linguistic interfaces*. Oxford University Press, Oxford.

Fellbaum, C., 1998. *WordNet – An electronic lexical database*. MIT Press.

Fine, K., 1975. Truth, vagueness and logic. *Synthese* **30**, 265–300.

Goldsmith, J., 1976. *Autosegmental phonology*. Ph.D. dissertation, MIT.

Groos, A., van Riemsdijk, H., 1981. *Matching effects with free relatives: A parameter of core grammar*. In: Theory of markedness in generative grammar.

Guenther, F. H., Hampson, M., Johnson, D., 1998. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* **105**, 611–633.

Halle, M., 1994. *The Russian declension: An illustration of the theory of distributed morphology*. In: J. Cole, C. Kisseberth (eds.), *Perspectives in phonology*. CSLI Publications, Center for the Study of Language and Information, Stanford, CA, 29–60.

Heim, I., 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. dissertation, University of Massachusetts, Amherst, ann Arbor: University Microfilms.

Johnson, M., Geman, S., Canon, S., Chi, Z., Riezler, S., 1999. *Estimators for stochastic "unification-based" grammars*. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics 1999. College Park, MD.

Kamp, H., 1981. *A Theory of Truth and Semantic Representation*. In: J. Groenendijk, T. Janssen, M. Stokhof (eds.), *Formal Methods in the Study of Language*. Vol. 1. Amsterdam Center, Amsterdam, 277–322.

Keating, P. A., 1988. Underspecification in phonetics. *Phonology* **5**, 275–292.

Kiparsky, P., 1985. Some consequences of lexical phonology. *Phonology Yearbook* **2**, 85–138.

Klein, F.-J., 1981. *Lexematische Untersuchung zum französischen Verbalwortschatz im Sinnbezirk von Wahrnehmung und Einschätzung*. Kölner romanische Arbeiten. Droz, Genf.

Levin, B., Rappaport Hovav, M., 1988. Lexical subordination. *Papers from the Regional Meeting of the Chicago Linguistic Society* **24, Part 1**, 275–289.

Lieber, R., 2004. *Morphology and lexical semantics*. University Press, Cambridge.

Marantz, A., 2001. *Words and things*. handout, MIT.

McCarthy, J., Taub, A., 1992. Review of "the special status of coronals: Internal and external evidence", by Carole Paradis and Jean-Francois Prunet. *Phonology* **9**, 363–370.

Müller, G., Gunkel, L., Zifonun, G. (eds.), 2004. *Explorations in nominal inflection*. Mouton de Gruyter.

Müller, K., 2002. *Probabilistic syllable modeling using unsupervised and supervised learning methods*. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 8 (3). University of Stuttgart.

Paradis, C., Prunet, J.-F. c., 1991. *Asymmetry and visibility in consonant articulations*. In: C. Paradis, J.-F. c. Prunet (eds.), *The special status of coronals: Internal and external evidence*. Academic Press, San Diego, CA, 1–28.

Pierrehumbert, J. B., Beckman, M. E., 1988. *Japanese tone structure*. MIT Press.

Pinkal, M., 1996. *Radical underspecification*. In: P. Dekker, M. Stokhof (eds.), *Proceedings of the 10th Amsterdam Colloquium*. 587–606.

Plag, I., 2004. Syntactic category information and the semantics of derivational morphological rules. *Folia Linguistica* **38** (3–4), 193–225.

Plank, F. (ed.), 1991. *Paradigms: The economy of inflection*. Mouton de Gruyter, Berlin.

Reyle, U., 1993. Dealing with Ambiguities by Underspecification: Construction, Representation and Deduction, journal = Journal of Semantics **10**, 123–179.

Reyle, U., Roßdeutscher, A., Kamp, H., . *Ups and downs in the theory of temporal reference*, to appear in *Linguistics and Philosophy*.

Riezler, S., Holloway King, T., Kaplan, R. M., Crouch, R., Maxwell III, J. T., Johnson, M., 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 2002. Philadelphia.

Ringen, C., 1975. *Vowel harmony: Theoretical implications*. Ph.D. dissertation, Indiana University.

Sauerland, U., 1996. *The late insertion of Germanic inflection*. Manuscript. MIT.

Schiehlen, M., Bos, J., Dorna, M., 2000. *Verbmobil Interface Terms (VITs)*. In: W. Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag, Berlin, 183–199.

Stemberger, J., 1991. Radical underspecification in language production. *Phonology* **8**, 73–112.

Stemberger, J., 1992. Vocalic underspecification in language production. *Language* **68**, 492–524.

Stemberger, J., 1993. Glottal transparency. *Phonology* **10**, 107–138.

Steriade, D., 1987. Redundant values. *CLS* **23** (2), 339–362.

Stevens, K. N., 1989. On the quantal nature of speech. *Journal of Phonetics* **17**, 3–45.

Stevens, K. N., 1998. *Acoustic phonetics*. MIT Press,, Cambridge, MA.

Stevens, K. N., Keyser, S. J., Kawasaki, H., 1986. *Toward a phonetic and phonological investigation of redundant features*. In: J. Perkell, D. H. Klatt (eds.), *Symposium on invariance and variability of speech processes*. Lawrence Erlbaum, Hillsdale, NJ, Ch. 20, 426–463.

Stump, G., 2001. *Inflectional morphology*. Cambridge University Press.

Trubetzkoy, N., 1939. *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague VII.

Vollmer, K., 1997. *Koartikulation und glottale Transparenz*. Doctoral dissertation, University of Stuttgart, arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, AIMS 3(5).

Wunderlich, D., 2004. *Is there any need for the concept of directional syncretism?* In: G. Müller, L. Gunkel, G. Zifonun (eds.), *Explorations in nominal inflection*. Mouton de Gruyter, 373–395.

Wurzel, W., 1998. *Drei Ebenen der Struktur von Flexionsparadigmen*. In: R. Fabbri, et al.

(eds.), *Modelle der Flexion*. Niemeyer, Tübingen, 225–243.

Zsiga, E. C., 1997. Features, gestures, and Igbo vowels: An approach to the phonology-phonetics interface. *Language* **73**, 227–274.