1.2 Academic profile of the Collaborative Research Centre

1.2.1 Summary

The point of departure for the SFB 732 is the observation that the elements from which linguistic expressions are built are almost invariably ambiguous or underspecified when they are considered in isolation; but when those elements are combined into larger complexes, and thereby placed in increasingly informative contexts, most of the ambiguities get resolved. Usually few or none remain once a linguistic expression is used in a particular situation.

The SFB 732 treats ambiguity as a form of underspecification (and thus disambiguation as a form of a specification process). Its central goal is to study the mechanisms of specification in context at severalevels of linguistic description and language processing: speech perception and production, morphology, syntax and semantics.

By a specification process we understand any linguistic process that can transform a given linguistic representation into any one of several alternative more specific representations, making a choice between these on the basis of evidence taken from some relevant context. Specification processes are thus per definition always situated in a particular context which provides constraints and triggering conditions, and they make reference to different types of representations, varying in their degree of specification or underspecification.

Our research program deals with two main questions: (i) what is the nature of the transformation from underspecified to more specified representations? and (ii) what is the role of the context in this process, what kind of information does it provide, and when does this become relevant? When the SFB was first conceived we saw this research program, with its many parallel explorations of context-driven specification, as providing a unique opportunity for reaching a better understanding of how something as central and ubiquitous as incremental specification – both as an aspect of language processing in the brain and by machines and as a topic in linguistic theory – manifests itself at different processing levels, how different methodologies deal with it and how each can shed its own light on the nature and function of specification processes.

In the first and second funding periods of the SFB, we have seen the benefits of this concept; researchers with very different backgrounds and using radically different methods can find modes of interaction when they seek answers to the same kinds of questions. Comparing and combining empirical findings from different areas of the study of language, and especially findings that already bridge adjacent areas, has enabled us to see implications for the local and global architecture of a comprehensive theory of language processing and language structure. And it has shown us how much cross-fertilization is possible when different methods and assumptions are brought together pursuing a common goal.

In the third funding period, we will be able to build on insights from the previous funding periods that have mostly led to modular theories of some clearly delineated partial specification process, assuming controlled interface representations and making predictions for empirical language data that are considered canonical for the area under consideration. The advanced state of understanding we have reached will allow us (a) to investigate how our theories generalize to “non-canonical data” in the respective areas, and (b) to make comparisons across different modular approaches and investigate cross-module interactions. Such investigations will consider both parallel (redundant) modules that address overlapping areas of linguistic evidence with alternative methodologies and pairs or groups of modules whose specification processes can be construed as sequential in the sense that the output of one process feeds another process. Integrating canonical and non-canonical data will further our understanding of the process of specification in context in language.

Both these future directions will go along with the SFB's strong emphasis on a sustainable data infrastructure, which we will broaden in the third funding period. A multi-level corpus annotation scheme with quality-controlled automatic annotations of “silver standard” quality takes advantage of and further fosters the shared interest in the exploration of corpus data across projects and research paradigms.

A methodological framework that will receive a boost in emphasis in the third phase is distributional semantics. With the advanced computational machinery and know-how in place for analyzing large-scale corpus resources and given the constellation of research interests in a new generation of principal investigators the SFB 732 is the perfect environment for a thorough investigation of this framework alongside other research paradigms for the study of language.

In the next subsections, we turn to a presentation of this SFB’s development as a whole, followed by a presentation of its individual areas, and finally to a summary of the key ideas for phase 3 for the SFB 732.

1.2.2 Research Program

1.2.2.1 Organization of the SFB 732

We structured the research program of the SFB732 into four areas. Our activities in the third funding period will focus on three of them: A, B, and D. Some of the guiding questions of area C will be pursued further mainly in area B.

Area A: Speech, segmental and prosodic representations: area A deals with speech parameters at the segmental and prosodic level and their interaction with information structure, syntactic and semantic properties, as well as their theoretical and/or computational modeling.
Area B: Meaning and disambiguation at the interface between words and phrases: area B is concerned with the relationship between lexical and supralexical composition of meaning.
Area C: Noun phrases and context: sharing a focus on the nominal domain, the projects in area C had the aim to provide detailed analyses of the interaction of the different linguistic levels that lead to a particular interpretation of a nominal expression**.**
Area D: Disambiguation in context: the projects in area D focus on data-driven computational approaches to the modeling of contextually driven choice among alternative candidate analyses for a given surface form, and the reverse process of deciding among alternative realization options for a given underlying meaning representation.

The SFB's project areas and the individual projects have all been addressing our two main questions (the nature of transformations from underspecified to more specified representations and the role of context) at a wide range of levels of description and using a number of different methodologies and theoretical frameworks. This has led to rich empirical findings about specific languages, important insights about the ways in which the human language faculty operates in the broader context of our cognitive system, detailed descriptions of grammatical and lexical elements in particular languages, and computational models for analysis and generation. Importantly, it also led to a closer collaboration in methods and results between theoretical and computational linguistics, which we will further deepen in the third funding period.

Underspecification and specification processes

To bring out important aspects of the deeper common objective across projects, let us abstract away from the concrete empirical domain, to which we will turn in the next subsections, and discuss what we take to be the common ground among the methodologically diverse projects in the SFB 732. The schematic figure in 1 depicts the desired scope of a modeling approach to a speaker's or hearer's ability to make the contextually appropriate choice from a set of options. Any model of language interpretation in the face of ambiguity will ollow the general scheme in the top half in some way; models of choice in language generation follow the same scheme in the reverse direction, as seen in the bottom half.

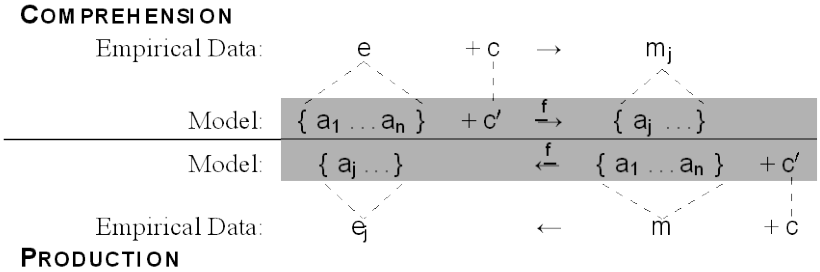

Figure 1: The general modeling scheme for specification in context

Upon encountering some linguistic expression *e* in a particular context *c*, the hearer must decide which is the appropriate interpretation m_j among a large set of interpretations which *e* could have in different contexts. Every theoretical or computational model will provide – either implicitly or explicitly – some characterization of a set of alternative analyses that would generally be compatible with *e*, and some function *f* which predicts that in the given context, represented as *c′*, the hearer will narrow down the set of choices. The reverse process models a speaker's choice among possible expressions for realizing some underlying thought or message in a given utterance context. The form of representation for the competing analyses has to be chosen in such a way that the denoted elements a_j of the narrowed-down choice reflect the hearer's interpretation m_j in the actual context, and the full set of alternatives captures the space of interpretations in all possible contexts (e.g., a_1 …a_n may all be grammatical syntactic structures for an observed string, and the local discourse context may lead the

hearer to exclude most of the available choices). As we have outlined in our previous proposals, the following possible aspects of context play a role in the specification process: linguistic vs. extralinguistic contexts, local vs. global contexts, dynamic vs. non-dynamic contexts. We have also hypothesized that we observe crosslinguistic variation in what counts as a possible context.

The schematic picture is very general and is – by design – completely open as to *how* a model characterizes the set of choices (disjunctive listings of candidates, compact underspecified symbolic representations, a vector space representation, probability distributions, etc.). But we note that it is the positioning of an observable expression *e* in a possible context (which has empirically observable properties) that provides theory-independent evidence for what are *relevant dimensions* and *characteristics* in the structuring of options available before and after considering some contextual factors. This makes the notion of an abstract process of Specification in Context an invaluable tool and meeting point in a methodologically diverse, interdisciplinary approach to the study of language.

In formal studies of language, the representations and functions assumed in a particular model process are obviously chosen in a way that ensures that we can accurately describe linguistic facts, obtained via linguistic experience of native speakers or from corpus evidence and/or experimental data and importantly make accurate predictions concerning the behavior of linguistic units. However, since there are considerable degrees of freedom in the choice of (intermediate) representations, the modeling process is heavily influenced by meta-theoretical principles such as the avoidance of unnecessary lexical stipulations as well as by principles of economy of expression or derivation. Thus an objective in systematic linguistic modeling can be characterized as follows: rather than assuming an explicit listing of the entire set of choices $\{a_1 \ldots a_n\}$ prior to contextual disambiguation, the model provides a single compact **aggregate representation** that denotes the entire set of alternatives. This is achieved by using some underspecified representation of the available choices (using the term underspecification in a non-technical sense),

In classical linguistic modeling working with symbolic formalisms, the form of an aggregate representation is some technical form of symbolic **underspecification**: in such a model, an entire set of choices that interacts with some contextual factors in a systematic way is represented by a single formal (feature) representation α, which leaves relevant decisions open (i.e., to be specified based on the concrete instantiation of contextual factors). Typical examples are the systematic underspecification of certain phonological features such as [+/– voiced] in some underlying lexical representation when they are contextually determined, and the underspecification of quantifier and operator scope in semantic representations. In the space of morphosyntactic modeling, underspecification of Voice morphology in the context of argument alternations would be another example.

This systematic usage of underspecification in modeling is defined in Alexiadou & Müller (2005) as follows:

(1) Linguistic expressions (LEs) may lack some property (or feature) x at some level of representation (Σ_i) and exhibit the same property x at some other level of representation (Σ_j).
The relation between LE(Σ_i) and LE(Σ_j) is determined by general rules or principles.

On this view, the grammar produces a representation which is subject to specification and thereby disambiguation at another component given some contextual information.

It is important to note that any specific form of underspecification as a mode of representation has to be viewed against a set of theoretical modeling assumptions, such as the assumed relevance of particular levels of representation and particular meta-theoretical principles. Finding an empirically justified, logically perspicuous and computationally economical form of underspecification provides strong support for the adequacy of the relevant modeling assumptions. Hence the study of theoretically grounded instances of the *Specification in Context* model across the spectrum of all linguistic levels, as it has been undertaken and carried out within this SFB, amounts to a comprehensive investigation of architectural assumptions and their implications.

Form the very beginning of the SFB 732, the classical modeling approach making direct use of symbolic underspecification has been complemented by alternative views on the (pre-theoretical) notion of a contextually driven selection process among candidates. In projects in area B and C, symbolic models of the type in (1) have been refined and complemented with additional ones, and tools were developed to keep these models apart. On the one hand, it was shown that the grammar may produce different morphosyntactic representations, but that the elements/vocabulary items (VI) used to realize them in a particular language are underspecified, i.e. they are sensitive to one (or a limited number) of features that are shared by their morphosyntaxes. This leads to a situation where one and the same form is used in conjunction with different morphosyntaxes/semantics (again the case of Voice syncretism, as studied in B6, is a good example to distinguish between the two types of

underspecification: while passives and reflexives in Greek share the same morphosyntax and are disambiguated only at the conceptual level, along the lines of (1) from Alexiadou & Müller (2005), anticausatives differ in morphosyntax/semantics but share the same morphology as a result of the VI underspecification). On the other hand, it was shown that choices of different readings of LEs, as e.g. in the case of *-ung* nominals studied in B4, can be coded in the form of underspecified disjunctions, which are hierarchically organized in an LE entry. These projects have provided very elaborate models of very local relationships between representations (e.g. syntax-morphology, syntax-semantics, syntax-lexicon), which clarify the role of these levels of representations in processes of disambiguation. A major outcome of this enterprise is the development of a framework that combines insights of different schools of thought, lexical semantics, formal semantics and research adopting the strict modularity hypothesis, according to which all composition is syntactic, i.e. the internal structure of words is created by the same mechanisms of construction as the internal structure of sentences, as e.g. in work within Distributed Morphology and exoskeletal approaches in general. We will come back to architectural implications of certain interface-level assumptions below.

Besides these refinements in the use of symbolic underspecification, many SFB projects, especially in areas A and D, have adopted different technical ways of achieving the systematic aggregation (or underspecification in a non-technical sense) over the input to the contextual specification process in their models.

A number of computational projects follow the fundamentally different, usage-based approach called **Distributional Semantics**. Distributional semantics represents linguistic units by observing their occurrence contexts in large corpora and representing them as vectors in a high-dimensional vector space. The dimensions of these vector spaces are properties of the context (e.g., context words), and the components of the vectors are functions of the co-occurrence frequency between the target words and the properties of their contexts. Following observations by Harris and Firth, the distributional approach is based on the assumption that vector similarity can be interpreted as indicating semantic similarity between the represented linguistics units such as words or phrases. Vector spaces are a form of aggregate representation that is quite different from the prototypical examples of symbolic underspecification. In a suitably parameterized vector space, the different interpretations of ambiguous words *e* (as shown in Figure 1) correspond to *regions* a_1 through a_n in the vector space. The vector representing an ambiguous lemma e is typically located somewhere "in between" these sense regions: It combines the contextual characteristics of the different interpretations and is essentially underspecified with regard to these interpretations. The specification process *f* for an instance of an ambiguous lemma *e* in a concrete context *c* can then be defined in two steps: First, the ambiguous lemma vector is modified based on the instance's context to produce an instance-specific vector. Then, the similarity of the novel instance vector to the sense regions a_1 through a_n is analyzed. This process lends itself naturally to partial specification: Instance vectors can still be similar to multiple sense regions. The degree of specification can be measured as the reduction of uniformity in the distribution over the instance vector's similarities to the various sense regions (i.e., some interpretations have become preferred over others).

An additional difference from symbolic models of underspecification is that the senses are often not treated as given, but are determined dynamically from the data as part of the computational model (e.g., by clustering unambiguous lemmas or individual instances). The fact that there is a high degree of freedom in distributional modeling (mentioned above as "appropriate parametrization") confronts the model creator with decisions that share characteristics with the more classical choice of parameters for an aggregate representation. Again, the broader decisions concerning architectural modeling play an important role. An important area of current research is strategies and interfaces for the combination of distributional approaches with models following one of the other approaches.

The framework of Exemplar Theory explores the utility of assuming a memory that stores, in rich detail (across multiple linguistic strata), individual observed utterances, over which abstractions can emerge (e.g. phonological/syntactic categories). Here, the goal of modeling is to establish how storage and retrieval/access to multi-level exemplar information is organized, and to determine how linguistic categories and patterns develop and change over time. For example, in production, speakers retrieve a fully specified target on the basis of both contextual information and an underspecified target.

Finally, statistical natural language processing models (**Statistical NLP**) can be understood as extensions of the distributional approach in two general directions. First, they expand the dimensions of distributional spaces, which typically only consider co-occurrences in context, to other types of features ranging from properties of the target word (semantic class, frequency, etc.) to richer descriptions of the context (specific constructions, analyses provided by other processing tools, etc.) and global features that characterize properties of the whole sentence or discourse (genre, sentence complexity, etc.). Second, they place more emphasis on utilizing the resulting feature vectors for

machine learning tasks such as classification. This setup is more similar to the traditional model of specification in that it recovers the notion of discrete output analyses. Instead of making a binary choice, however, its goal is to estimate a probability distribution over the candidate output analyses, depending on the observed contextual features. Like in symbolic linguistic modeling, the structure of interdependence in the broader model architecture is a highly relevant factor for the effectiveness of an individual representational assumption. In addition, the effectiveness of statistical parameter estimation from the type of (annotated and/or unannotated) corpus data that are available becomes a relevant factoras well. By contrast,Meta-theoretical principles of economy or the like are not typically assumed to be relevant criteria for deciding on the adequacy of representation – it is system performance on representative test data that counts. Nevertheless, if appropriate design decisions are taken, linguistically informed high-level decisions about representation can well improve the overall performance.

To conclude our discussion of *Specification in Context* as a schematic abstract process, we note that the purely structural characterization has the advantage that we can identify the process as a typical building block appearing in any systematic, empirically grounded approach to modeling the phenomenon of ambiguity in the study of language. Often, we find complex modeling architectures that assemble several modular accounts each following the same structural scheme, and we will come back to this below in the discussion of the SFB's architectural findings.

For our SFB, the schematic view has proven immensely fruitful for comparison and project collaboration across levels of linguistic description, across disciplines, theoretical paradigms, as well as language families and languages. While the entities, representations and functions/processes under consideration may differ, the common scheme of specification in context makes it possible to pinpoint systematic similarities and differences – for instance the potential/justification for using underspecification in different modeling tasks. The view also provides motivation for the use of annotated corpus data, using relatively theory-independent "consensus" annotations as "hubs" in the exploration of empirical evidence across modeling paradigms.

Under the guiding idea of *Specification in Context*, detailed research in the SFB over the past eight years has led to a rich body of knowledge about general aspects of interface and context representations in linguistic models, as well as special insights relating to particular elements and choices in the actual modeling setup and to the empirical domain of study.

General results of the previous funding periods

In simple terms, we can say that during the ***first funding period (mid-2006 to mid-2010)***, the major focus of SFB research was on the **choice of representations**. We investigated the input and output of the specification processes, i.e., (a) underspecified representations vs. fully specified or less underspecified ones, (b) some other aggregate representation, or (c) a radically specified exemplar-theoretic representation. Related to this, we focused on determining what the relevant context factors are for particular specification processes and how they inform the choice of a suitable representation. The broader modeling architecture assumed in most individual projects during the first funding period was (consciously) incorporated a fair number of idealizing assumptions in order to allow for controlled in-depth studies that clarify the status of core representations and mechanisms of specification.

During the ***second funding period (mid-2010 to mid-2014),*** we have extended our focus to questions concerning the appropriate **model architecture** for the specification process under consideration: to what extent is it justified to view a sequence of specification processes as a pipeline of modules, each building on the output of the previous ones – and to what extent can a broader picture of simultaneous interaction be assumed, which allows "joint modeling" of effects across levels of description?

Figure 1 schematically captures the notion of *disambiguation*: many LEs can have various different interpretations or *readings*, but speakers/hearers normally have the competence to pick a single one in a given context. In order to be able to model this process systematically, the relevant properties of expressions have to be accessible in some representation, and since various properties are known to interact in the process of context-sensitive specification, or disambiguation, the simple scheme requires some further explication. To capture different properties in the general case, it is useful to unfold the candidate analysis a_i (as it is included in the set $\{a_1 \ldots a_n\}$ of analysis choices for some expression e) and think of it as a bundle $\langle \ell_i^1, \ell_i^2, \ldots \ell_i^k \rangle$ of properties (at k different levels of linguistic representation, or layers), each of which can in principle be aggregated over separately in some underspecified representation.

Since the cognitive process of picking a particular reading in context is extremely complex (and for instance involves extra-linguistic knowledge), it is common to focus attention on a subprocess with defined linguistic interface representations, typically situated at an established level of linguistic descriptions, such as syntactic constituent structure etc. The subprocess can then, quite conveniently,

be seen as a small-scale version of the full process; and thus it seems reasonable to construe the full process as a cyclic chain of formally similar subprocesses, as indicated in Figure 2:

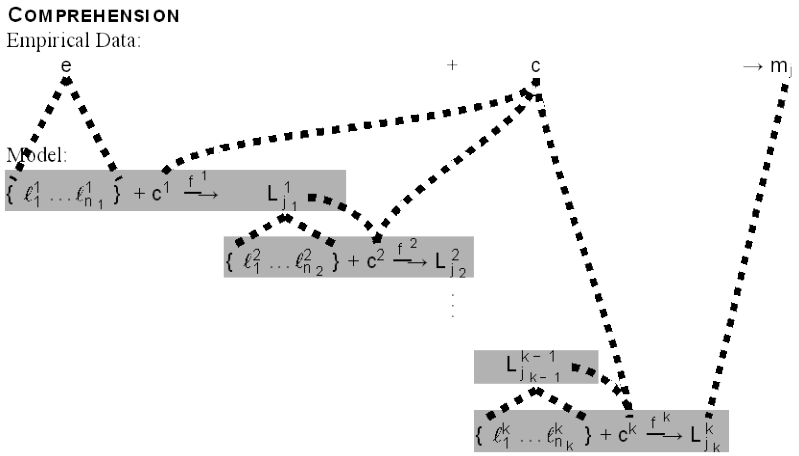

Figure 2: The cyclic (pipeline) model of specification in context

The underlying assumption is that at each layer *i*, a specification process f^i reduces a set of possible alternatives $\{\ell_1^i \ldots \ell_{ni}^i\}$ for this layer (represented in some aggregated form) to some subset L_{ji}^i, whose representation in turn defines the choice of options for the next layer *i* + 1. Note that if we view the cascade as a series of contextually driven specification steps, the relevant context for each step is not just determined by the empirically observed (presumably largely extra-linguistic) context *c*, but each layer contributes highly relevant bits of information to the specification context at the next layer. For instance, layer 2 may be the level at which inflectional feature values such as number (on verbs with subject agreement and on nominal elements) are determined, and layer 3 may be the level at which the syntactic structure for this input string is determined. Then due to agreement constraints the feature values determined at layer 2 will affect the specification at layer 3.

Classical feature underspecification at intermediate levels of representation is typically motivated by the observation that certain choices stay open across layers at which the relevant feature type would normally be resolved. Clearly, the modeling decision for interface representations is intimately tied to the assumed sequence of cyclic specification decisions, i.e., the architectural design. Modeling alternatives can be decided on the grounds of economy considerations.

The cyclic specification sequence goes along with one important characteristic: increase in specificity has to follow the same sequence across layers for all problems; in a classical pipeline architecture, specification decisions cannot normally be undone later. Often, the contextual clues at a particular layer give strong indications for a certain specification, but the decision can be overridden later, as for instance work in B1 and B4 on *-ung* nominalizations has shown. This effect cannot be modeled appropriately in a plain pipeline. While earlier work in Generative Linguistics (e.g., the GB model) was based on a clear concept of subsequent levels of information, more recent models (Minimalism, Distributed Morphology, exoskeletal approaches) have abandoned the idea of a step-by-step sequence of specification. Largely, problems of ambiguity are resolved at the interfaces with the articulatory and the perceptual system, respectively, a line of research adopted in several projects in area B and C.

Interestingly, despite its conceptual limitations the pipeline model forms the predominant skeleton (baseline) system in data-driven approaches in Natural Language Processing (NLP). Here, a layer corresponds to some analysis tool trained on annotated corpus data following the classical levels of linguistic representation. When applied to new input data, the tools make no strict choice of specification, but assign probability scores to the various options. In the typical pipeline set-up, the highest-scoring prediction is passed on to the next layer, which may of course occasionally have the effect that the subsequent layer can no longer make correct predictions, even though there may be strong local evidence. The pipeline approach is nevertheless predominant in NLP, for computational reasons and because training tools layer by layer on gold standard data, i.e., using manually annotated target representations, yields the most reliable prediction results (statistical effects from subsequent steps in the pipeline are often anticipated in earlier components such that it is not easy to improve over a pipeline baseline system). The risk of error propagation can be reduced by simulating the use of predicted outcomes from earlier layers during training. Various modifications of the pipeline are possible and have been studied in SFB projects, for instance in projects D2, D4 and D8.

We had already pointed out in our proposal for the second funding period that the pipeline architecture is not fully adequate to model situations where two independent linguistic subsystems constrain the space of possibilities for further specification. An abstract architecture that makes it possible to capture arbitrary cross-level effects is the joint model sketched in Figure 3, which does not specify in advance any particular sequence of subsequent specification. Procedurally speaking, it posits a simultaneous decision, in principle allowing for arbitrary global interaction across layers. This basic scheme reflects the abstract modeling setting assumed in most of the projects from areas B and C in the second funding phase, and it also reflects the assumptions of constraint-based formalisms like LFG, underlying the abstract modeling in D2. Moreover, the exemplar-based characterization of the multidimensional categorization is best captured with such an architecture.

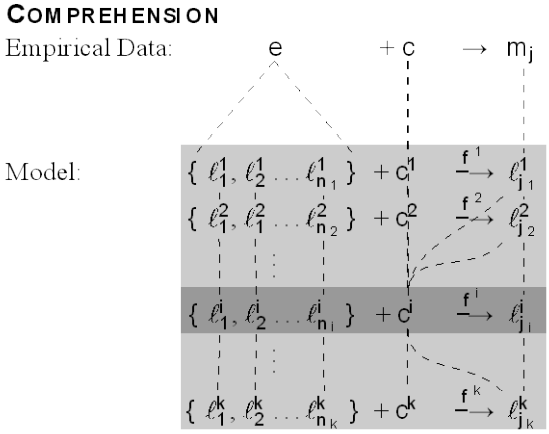

Figure 3: The joint model of specification in context

In the joint model, any subprocess of layer-specific specification can be informed by the output of any other subprocess; i.e., effectively the specification decisions mutually contribute context information to each other. Of course, a model based on this principle has to (a) break up the circularity in this conceptual design, and (b) solve the computational complexity issue that arises when the full search space is to be explored algorithmically. A possible simulation is for instance a modified pipeline that will keep a record of the *n* highest-scoring analyses and provide the alternatives in subsequent steps, so they can learn to correct the top predictions of earlier layers. A similar technique is to generate some subset of the cross-product of alternatives stemming from submodules, and then use discriminative (ranking) models to learn to pick the best combination of layer-specific choices (as used in D2 and D4). One can also use powerful search techniques for finding a combination of choices, subject to certain constraints, which may reflect cross-talk among layers (here, D8 has made a number of contributions). Over the past few years, the investigation of joint inference techniques has developed into a “hot topic” in the fields of computational linguistics and machine learning.

Studies into cross-layer interaction and in particular cross-project collaborations in the SFB have shown that even the well-understood aspects of “local” specification processes (such as the process of syntactic parsing, given a reliable tokenization and part-of-speech categorization of the input string pose non-trivial problems when the interaction of two or more such specification processes is regarded as a joint problem (e.g., morphological analysis and parsing of real world corpus data). Some dedicated activities in the first and second funding periods have already examined particular combinations of well-defined specification processes, leading to insights clearly beyond what can be achieved within the component models, e.g., the Multi-Level Exemplar Model (MLM) developed in A2 and the linguistically constrained search over locally trained statistical morphological and dependency-syntactic models in D8. Yet, by and large, our research work has so far obeyed certain idealized assumptions about suitable interface representations when crossing linguistic levels, partly due to the fact that established working assumptions are usually required as a starting point for placing the research contributions within the respective subfield.

Plans for the third funding period

During the two previous funding periods, investigations in our SFB have (a) mostly dealt with **canonical data** of the phenomena under consideration, and (b) focused on a **single layer-specific module** and its relation to the global architecture, or, when taking a broader architectural view, on modeling **units** that are **immediately adjacent** and directly compatible in terms of the modeling assumptions. These two idealized assumptions made it possible to control the space of options and

develop an advanced understanding of the systematic implications of representational and architectural choices.

The progress that has been made in the SFB thus far puts us in a position in which we can broaden our focus and pursue new research avenues: in the ***third funding period (proposed for mid-2014 to mid-2018)***, the projects will not only consolidate successful strands of research from the previous period and round up advances made in particular areas of study. They will also systematically explore more complex architectural interactions and investigate how the established modeling approach in each area responds to an application to **non-canonical data**, as defined below, and how assumptions may need to be adjusted. This turn has been anticipated in nearly all of our projects, for we have learned how valuable the inclusion of non-canonical data can be (see e.g. work on non-canonical argument alternations and realizations of the type studied in area B, in collaboration with projects in area D).

Canonical data are by definition data that straightforwardly conform to a rule or set of rules of grammar (phonetic, phonological, morphological, syntactic, semantic) or to some established annotation scheme. Non-canonical data are data that fail these criteria in one way or another. In general such deviations should not be simply dismissed as errors. They often reveal something about the underlying structure of the grammar, or about the dynamics of language use, that limiting attention to canonical data would have kept hidden. It should also be noted that at the level of expression *types* (as distinct from individual occurrences) canonical data represent a comparatively small proportion of all data. Therefore, including non-canonical data in the scope of one's theory and model is a way of substantially expanding its coverage. Thus, by examining non-canonical data in phase 3 we can make significant headway in terms of coverage of real language data. Moreover, the attempt to extend an account beyond its canonical scope is an important test of the validity of assumptions made so far, it contributes to the proper characterization of the phenomenon under investigation, and is thus an integral part of our goal to develop a more comprehensive understanding of the process of specification in context.

It is clear from our characterization of canonicity that this is a relative notion: Data can be canonical with respect to one type of description and modeling, to which some of their properties must conform, without also having to conform to some other description and modeling which selects for different properties. The same is true for 'non-canonicity': a piece of data can be non-canonical in one sense without being non-canonical in another. Thus, an utterance may be entirely canonical when viewed in terms of its prosodic properties and at the level of surface syntax, but may display a highly non-canonical usage of some construction, such that the canonical semantic interpretation is empirically inadequate. It is a natural development that by increasing the scope ofcross-level interactions applied to real corpus data, more and more submodules of the overall architecture have to be exposed to data that fall outside of the controlled spectrum of canonicity. Importantly, such data will also be brought to bear upon our understanding of theories of phonetics, phonology, morphology, syntax, semantics and the various interfaces between them.

Determining what counts as non-canonical data may be based on different parts of a grammar, on different annotation schemes or even to different practices of data selection, e.g. of the different languages that a crosslinguistic investigation is to take into account or of the corpora that are used for training and data extraction. Some of this variability is reflected in the different ways we use the terms 'canonical' and 'non-canonical' within the SFB. Here is a (somewhat tentative) list (more details will become clear in the next subsection and in the individual proposals):

- understudied languages/varieties (A2, A5, A6, A7)
- unintuitive phonetic phenomena (A2)
- data that appear to deviate from generally accepted descriptions (B4, B5, B6, B7, B8, D12)
- data at the boundary between two types of constructions (B1, B4, B6, D10, D11, D12)
- data that are subject to a high degree of variability within a language (synchronically and diachronically) and across languages (A5, A6, A7, B1, B5, B6, B7, B8)
- new text types (outside of the standard newswire domain) such as web corpora which include a high proportion of "uncleaned" text (B9, D2, D8, D11, D12)
- data that contains interjections, hesitations, pauses, self-corrections, orthographic variation and incomplete sentences (A4, A6, B5, D8)

One typical effect of studying non-canonical data is that there will now be a wider range of possible representations for a given linguistic input. We may expect this for instance when we move from news data to interviews, from formal to less formal registers or from synchronic data, which portray a language at one particular time, to diachronic data. In addition, crosslinguistic (non-canonical) approaches that seek to account for variation found across languages need to work with

underspecified crosslinguistically valid representations that can be further specified on a language-per-language basis. Thus incremental specification in context will become an even more important component of language processing, and given the central goals of the SFB 732 this means a special challenge for us. The new projects in areas A, B and D, as well as the stonger emphasis on psycholinguistic investigations in all three areas of our SFB will put us in a much stronger position to meet this challenge in phase 3.

Complex model architectures can combine ideas from the pipeline and from the joint model, yielding a vast space of possible system architectures. As the collaborations of the SFB have shown, understanding the strengths of architectural modeling decisions in the light of potentially divergent modeling goals can be the key to a successful combination of tools and resources. Symbolic model components, yielding a strict exclusion of options, have to be used in quite a different way than probabilistic components, but more fine-grained differences have to be taken into account, too. There is no single architecture that guarantees optimal behavior of an arbitrary process of specification in context, but implications of particular choices can be clearly stated.

The specific focal points for the third funding period go along with the need to access large amounts of corpus data for exploration, hypothesis testing, estimation of vector space models and weakly supervised or unsupervised training techniques. For this, the SFB's successful work on highly competitive automatic analysis tools at various levels and the intregration of multi-level annotations from different origins in a database will be further scaled up to devise a methodological framework for sharing a large common collection of corpus data with high-quality annotations, a transparent update and versioning scheme and multi-level exploratory facilities. Since there is a quality insurance scheme, we refer to these automatic annotations as **silver standard annotations**, following the terminology adopted by Rebholz-Schuhmann et al. (2010), among others, to indicate a quality level below gold standard (manual) annotation (which is out of reach for any corpus of the size required to support exploratory search for varied phenomena and in varied domains, text genres and styles of language). We will come back to this in section 1.3.3.

In the next subsections, we summarize the results and plans of our project areas before we turn once more to the broader perspective of the SFB as a whole.

1.2.2.2 Area A: Speech, segmental and prosodic representations

Characterization of area A

In general, the projects within area A are involved in the investigation of non-canonical data with respect to segmental and prosodic speech parameters, information structure, syntactic and semantic properties, as well as their theoretical and/or computational modeling. Beyond that, variability of prosody is one of the most prominent common topics among the A projects. While project A2 will continue to provide theoretical/computational models as input to further research into prosodic variation (focusing on exemplar-theoretic perspectives, which have been developed together with project A1), projects A5, A6, and A7 will investigate linguistic and paralinguistic variability in prosody due to information structure. Moreover, the projects in area A will contribute to enlarging the silver standard collection with corpora of unscripted, spontaneous speech (e.g. free conversations or radio interviews as in projects A4 and A6) and with data prepared for an in-depth look at prosody in less well-studied languages or varieties (projects A5 and A7).

Results of the second funding period

During the second funding period, the A projects were involved in the investigation of segmental and prosodic variability across different datasets (ranging from experimental articulatorily labeled data over spoken radio news corpora to spontaneous free conversations). Prosodic variability was looked at in relation to semantic and information-structural restrictions (A1), as a function of interpersonal dynamics within dialogic interaction-phonetic convergence (A4), and also examined from an exemplar-theoretic perspective including frequency effects (A1 and A2) and possible transfer phenomena between first and second languages (A2). Acoustic and articulatory variability was studied in a multilingual corpus of voicing profiles (A2). The method employed here was also exemplar-theoretic modeling, which proved very effective in accounting for these quite elusive phenomena.

One of the key achievements of project A1 was the creation of the DIRNDL database - the Discourse Information Radio News Database for Linguistic analysis (Eckart et al. 2012) - which was a result of the fruitful collaboration between projects A1 and B3. The unique value of DIRNDL for the investigation of prosody lies in its multiple layers of annotation covering syntactic annotations, prosodic information (pitch accents, prosodic boundaries, and - in the latest extension of DIRNDL - also acoustic details about intonation), segment, word and phonetic context information (e.g. syllable structure), information about the prosodic context, and also labels for referential as well as lexical information status and contrast. This richly annotated database with cross-linked layers was the basis for a number of studies within project A1, as well as for project A2, and is currently also used by

research groups outside our SFB (across Germany, in Zurich and New York). In the next phase of the SFB, the B3 database and DIRNDL will serve as the infrastructural backbone of the INF project, in which silver standard technologies for richly annotated databases will be developed. The sustainability of the methodological research is guaranteed by continuity of the personnel involved (Dr. K. Schweitzer, formerly A1, and K. Eckart, formerly B3, are intended as primary researchers of the INF project). The manually annotated prosodic part of the DIRNDL corpus was used to test the variability of pitch accent shapes with regard to frequency of occurrence effects (projects A1 and A2, in collaboration with A4 for shared statistical methodology and exemplar-theoretic interpretation). DIRNDL served as the dataset for a pitch accent categorization experiment in an exemplar-theoretic framework (collaboration of A1 and A2) as well, thereby also providing corpus data for testing the exemplar models developed in project A2. Project A1 was further involved in a study on the degree of prosodic prominence in relation to different referential and lexical information status (RefLex) combinations (Baumann & Riester, 2013) in read and spontaneous data. Thanks to its manual gold-standard annotations of information status, the database will also provide a sound empirical basis for modeling experiments in projects A5 and A6 in the next phase.

Another richly annotated database used in the investigation of prosodic features in dialog was created within project A4 - the IMS GECO corpus. The corpus includes silver standard annotations on segment, syllable and word level of 46 spontaneous dialogs in German and is complemented with social and psychological data about the participants. The database has so far been used within project A4 to test the mutual adaptation of dialog partners for a range of prosodic and segmental parameters in relation to social context and psychological factors and is planned to be officially released soon through the CLARIN-D project. The IMS GECO database is the first annotated corpus of free, spontaneous dialogs in German, and is especially well-suited for the investigation of phonetic convergence due to the natural spontaneous character and richness of the data provided. This rich dataset with additional social and personality variables served to test and expand the hybrid model of phonetic convergence proposed by Lewandowski (2012), which is also modeled in an exemplar-theoretic framework.

Concerning the SFB-wide discussion of model architectures, our results show interactions between levels in production that are usually not assumed to interact in a strict pipeline model of speech production, where phonetic encoding only operates on syllabified sequences of phonemes and associated pitch accents, or where articulation only depends on the immediately preceding phonetic encoding stage. For instance, A1 has confirmed that information structure, as well as the anaphoric and lexical properties of utterances, influence both phonological and phonetic encoding. Similarly, A2 has found that lexical properties as well as phonological context affect articulation, and A4 has shown that social context is relevant at the stage of phonetic encoding.

Plans for the third funding period

As the main goals of A1, namely developing an annotation scheme for Information Status (givenness and contrast), and providing a database with fully annotated prosodic, syntactic and discourse semantic relations (DIRNDL database), have been achieved, we will not continue this project. Project A2 has a new principal investigator since both Bernd Möbius and Hinrich Schütze accepted offers for professorships (at the Universität des Saarlandes and the Ludwig-Maximilians-Universität München, respectively). The new principal investigator, Grzegorz Dogil, has taken part in the project's research activities throughout the previous phases and has supervised the doctoral theses of Dr. Jagoda Bruni and Daniel Duran. In phase 3, project A2 will continue to work on their computational exemplar models, however from the novel perspective of applying them to non-canonical data. The two existing models - the Context Sequence Model and the Multilevel Exemplar Model - will be combined into one model, pooling the strengths of the two approaches. The model will be applied to the study of prosodic and segmental variability (including frequency of occurrence effects, vowel and pitch accent categorization). Additional computational modeling will involve simulations investigating language change. In particular, the emergence of unintuitive phonetic behavior within and across generations will be explored. Exemplar-based approaches will also be under investigation in project A4, e.g. in simulations of exemplar models and their dependence on attention and memory mechanisms within a hybrid model of convergence. The existing IMS GECO database will be expanded by psycholinguistic data on subjects' attention capacities, following the hypothesis that this might crucially impact a speaker's adaptation. Phonetic convergence will be studied in spontaneous free dialog contexts, ensuring a maximal richness of prosodic and segmental representations and providing a natural basis for the models tested in the subsequent simulations. Moreover, a new data set will be employed: several projects from different areas of the SFB will collaborate to acquire a new corpus of radio interviews (*Deutschlandfunk*) in the next phase. This data will be automatically annotated for syntax, prosody, and segment information, and observer-ratings of the speakers' personalities will be

collected with the goal of automatically annotating personality traits.

The new projects A5, A6 and A7 establish an additional joint interest within area A, namely the investigation and modeling of phonological and/or phonetic features of prosody in connection with information structure. A basic assumption shared by all these projects is that information-structural notions are universal but are implemented in a language-specific way and/or in different grammatical areas (prosody, morphology, syntax) across different languages. The scope of research in A5 extends to studying fine-phonetic detail in affective prosody in combination with focus particles which induce an evaluative meaning. This phenomenon is investigated in a set of three languages: German, Mandarin Chinese and Vietnamese. The latter is understudied in terms of its prosodic features. Project A6 will analyze spoken dialog data in French and German with respect to discourse structure and information structure and study the prosodic and syntactic differences. German and French differ substantially in terms of prosodic and syntactic marking of information-structural notions. Project A7 investigates the prosodic expression of information-structural categories in Black South African English, a second language used by native speakers of Bantu languages, and L2 English by speakers of L1 German. The main focus here lies on models of the production and processing of intonation in a second language, developed on the basis of empirical data for the languages under consideration.

Investigating prosodic phenomena in non-canonical language data, i.e. in less well-studied languages or varieties, like Vietnamese (in the case of project A5), Tswana (in A2 and A7), Black South African English (A7) or L2 English by speakers of German (A7) is thus a common feature of studies in the next phase. Project A2 will use DIRNDL in its investigations into prosodic behavior and both projects A5 and A6 will employ DIRNDL as a gold-standard dataset against which to test their hypotheses regarding information structure.

Second language acquisition/usage also forms an important component of future research within the A area. For A7, second language usage is the central phenomenon. In project A2, L2 will also be examined, in this case with regard to the acquisition of local syntactic knowledge (employing the MLM). This builds upon earlier work carried out in A2 on exemplar-theoretic syntax. Additional work at the syntactic level to be performed in the A area includes collaborations between B6 and A2 (examining the behavior of psychological verbs) and between A4 and D8 (exploring parser adaptation and syntactic convergence).

For the next phase, we expect our hypothesis, viz. that a strict pipeline model is not sufficient to explain some interactions observed in the area A, to be further strengthened by results from continuing and new projects. For instance, A5, A6, and A7 will contribute to understanding the relation between information structure, discourse particles, and prosodic implementation. A4 will extend its research into extra-linguistic influences on phonetic implementation and complement social context factors with individual differences in the participants' attention capacities. As one common interest in area A lies in extending our analyses to L1 influences on L2 phonetic implementation (A2, A7), this will constitute an interesting challenge for model architecture, where one has to deal with interactions between L1 and L2 model components.

Collaborations

The extensive collaborations of individual projects within area A as well as between area A and individual projects from areas B and D are highly specific and thus listed and described in detail in the individual project sections.

As in the other areas in the SFB, we will organize an annual workshop series, since a workshop is the best way of providing a common ground for a fruitful exchange of current advancements and ideas, and for a genuine discussion of methodologies as well as both theoretical and practical results of value for the whole area A. The workshop planned for 2015 will be concerned with methodology-driven research on prosodic variability such as Exemplar Theory, Rich Representations and Simulation Technology. The second workshop, planned for 2016, will focus on the use and analysis of non-canonical data in research on prosody and information structure.

1.2.2.3 Area B: Meaning and disambiguation at the interface between words and phrases

Characterization of area B

The projects in area B are concerned with the relationship between lexical and supralexical composition of meaning. They examine the similarity between word-internal structure and phrase structure and the type of information that contributes to the building blocks of the meaning of phrases and clauses. The empirical domain of investigation is the ambiguity related to word formation processes in both the nominal and verbal domain (deverbal and deadjectival nominalizations, different types of conversion, deadjectival/denominal verbs), argument alternations and their relationship to Voice morphology, particle and prefix verbs, as well as the variation in event entailments in the verbal domain. The main questions the projects in this area deal with are the following: If a word can mean different things in different contexts, then i) how can these different interpretations be teased apart, ii)

how do they come about, and iii) what kind of information is necessary to select a particular reading? The overall goal of area B is to explain the ways meaning is built and the variety of information that influences the meaning of words in context and the variety and sources of information, the nature of these dependencies and the ways in which dependencies on different contextual information may interact. In pursuing this goal, the projects in area B aim to formulate and formalize algorithms of disambiguation of word meaning in context. To successfully reach these goals, they make use of various complementary methods. These include a crosslinguistic perspective in combination with in-depth analyses of individual languages via synchronic and diachronic investigations, data collection from corpora and lexical resources as well as experimental studies.

Results of the second funding period

Our commitment to the strict modularity hypothesis, mentioned in 1.2.2.1, according to which the same principles that guide the formation of sentences also guide the formation of words proved to be a very fruitful perspective for the analysis of constructions that do not seem at first glance to be compatible with a syntax-driven approach to meaning composition. Specifically, we have been pursuing the idea that compositional analyses of meaning for particular linguistic constructions presuppose a syntax which determines the order and the types of the composition steps. Over the past few years there has been a close collaboration between projects B1 and B6, which work within purely morphosyntactic frameworks (Minimalist Syntax and Distributed Morphology), and projects B4, B5, and B7, which are concerned with semantic lexical sub-structure. It was this particular collaboration that gave us the impulse to examine in depth how aspects of the meaning of the phenomena under investigation are composed from elements present in their morphosyntactic representations and to raise the question what the correspondences are between morphological pieces (roots, stems, prefixes, suffixes) and lexical semantic roots and primitive predicates, integrating insights from different theoretical frameworks. Work in Distributed Morphology has stressed how elements of meaning or functional properties encoded in morphology apply to roots to build up meanings of derived words compositionally; work in lexical semantics has concentrated more on the elements of meaning encoded in the roots themselves and on the organization of conceptual information, and work in formal semantics has focused on model-theoretic semantic mechanisms of meaning composition. In area B, we have successfully brought together all these approaches and integrated their important different perspectives in determining word meaning.

All projects work on processes that involve category change (e.g. different types of nominalization, different types of conversion, deadjectival/denominal verbs). Such processes raise at least three questions. First, how much of the meaning of the original category is preserved in the new element? Second, can we observe readings of the new category that were not available within the source category? Third, what is the minimal source element that enters word formation (stem or root)? We have been able to show that the meaning of derived words only partially depends on the source element. An array of factors, which are associated with different linguistic interfaces, play a role, such as the template structure the source element combines with (which can be read off its morphological makeup), the type of arguments it combines with and the kind of type-shifting operations/coercion it can undergo.

A large body of our research in phase 2 focused on the realization of argument structure under category change and complex word formation, associated with particular readings of the nominal/adjective only. In parts of this research, we have successfully established a fruitful dialogue between lexicalist and syntactic treatments. Projects B1 and B7 have been able to show the advantages and the limitations that each of the two views has in accounting for the subtle differences among e.g. deadjectival nominalizations. For instance, while the lexicalist view can formulate finer semantic distinctions among different types of bare nominalizations, the syntactic account fares better in explaining the complex internal syntax of such nominals.

An important issue related to argument structure is what happens to argument slots and the phrases filling them when a word of one category is transformed (by 'derivational morphology') into a word of another category (e.g. the 'transformation' of verbs into deverbal nouns or adjectives). Another issue is the way the argument structures of adjective-noun combinations relate to the argument structures of the nouns and adjectives of which they are composed. In pursuing these questions, we hypothesized that in all cases there is a common core, namely the root, which combines with certain syntactic constructionsthat ideally are parallel across domains (verbal/nominal/prepositional). The projects in area B, as well as C2, have looked at these questions from different perspectives, and arrived at very similar conclusions. In this sense, our work is unique in comparison to other linguistic centers that have similar research goals, as in our research we integrate morphological, syntactic, and semantic tools and analyses, naturally reflecting the expertise available in our institutions.

In dealing with the ambiguity of lexical elements, the projects in area B have focused on the role of various disambiguating factors, which we have primarily located in the morphosyntactic context

the core element is inserted into and in the way this morphosyntactic context interacts with conceptual information; for instance, work in B4 and B5 has shown that information such as outer Aspect and modality trigger particular interpretations of the verbal element. Our commitment to the view of meaning composition outlined above enables us to proceed incrementally in determining meaning. Work in B6 has demonstrated that Voice syncretisms could result from two different grammatical strategies: In some cases, the grammar produces an identical morphosyntax that allows further specification and thereby disambiguation at the conceptual-interpretative component; in others, it produces different morphosyntaxes and the elements used to realize them in a particular language are underspecified, i.e. they are sensitive to one (or a limited number) of features that are shared by their morphosyntaxes. Natural languages make use of both these strategies in the verbal domain and we developed tests and tools to keep them apart. B1 has shown that the latter strategy is available in the context of ambiguous nominalizations. Work in B4 has provided evidence that it is even possible to 'reambiguate' a word: the context in which an ambiguous word occurs disambiguates it to one sort, but the next sentence targets a different reading. A formal account of reambiguation put forth in B4 makes reference to an underspecified lexical entry whose various available readings are represented in dependency relations.

In view of the fact, however, that meaning is also dependent on the core element that is contained within words, projects in area B approached the question of how much and what type of information is included in this core element, and more importantly, what type of status it has. For projects B1, B4, B5 and B6, core elements are roots that belong to different types (e.g. manner, result). These are subject to different requirements in interaction with template structure. However, what we label roots depends on the language under investigation. Work in B7 has raised the question whether determination of typical associations of conceptual fields with stems on the one hand, and roots on the other, is necessary. Other work in B6, B5 and C2 has shown that languages differ as to the role a root has in determining meaning. For instance, in English, the interpretation of the root is primarily determined by its encyclopedic meaning. In other languages, such as Greek, German and the Romance languages morphosyntactic information plays the key role. Further examples of root and morphosyntactic information interaction are provided in the reports of all our projects.

Our strong commitment to collaboration between computational and theoretical linguistics has allowed a more objective approach to the empirical domain by means of the investigation of large corpora of data and has also provided a solid testing ground for the analyses we have pursued. In particular, projects B4 and B6 have established fruitful collaborations with more computationally oriented projects, e.g. B4/B3, B6/D2 (see the part *collaborations within the SFB*; see also the individual collaborations within the individual project reports). B5 embodies this theoretical-computational strategy within one, applying statistical methods to the lexical resources available in order to verify theoretical assumptions (e.g. correlation between affixation and transitivity). For its theoretical investigations, project B1 has developed databases of psychological verbs after the model of B5 , which will be administrated by INF and will be used by Project B6 in phase 3. B7, too, benefited from the computational expertise in B5 for their corpus searches and will continue doing so during the next funding period. In collaboration with B3, B4 has tested their classification of *nach*-verbs against information that can be extracted with current NLP tools from corpora and in collaboration with the Heisenberg project of Schulte im Walde, NLP methods were applied to corpora in order to extract four-fold classification of an-particle verbs. B3 has created a relational database (B3DB) which provides data structures for the handling of primary data and multi-layer annotations on the one hand, and to track the workflow which led to the creation of these annotations on the other hand. Moreover, in collaboration with A1, B3 combined speech- and text-related annotations in the B3DB, which resulted in the DIRNDL corpus, described in the section on area A. Having successfully reached its objectives in phase 2, and having provided an excellent example of cross-fertilization between theoretical and computational linguistics, project B3 will not be continued in the next phase. All the resources made available by B3 will be managed by INF, and Kerstin Eckart, researcher in B3, will be a Postdoc in INF. These collaborations show that corpus-based methods are powerful tools for validating and enhancing descriptive and exploratory hypotheses and directly feed into linguistic theorizing.

In addition to joint presentations and publications, projects in area B organized a series of thematic workshops, which furthered the dialog among the various approaches.

Plans for the third funding period

In phase 2, we have been working on identifying the pieces necessary to build word meaning by predominantly looking at what counts as canonical data for the description of the various linguistic phenomena we took into account. We have also been examining very local relationships between representations (e.g. syntax-morphology, syntax-semantics, syntax-lexicon) and clarifying their role in processes of disambiguation. Several projects have been developing a model that combines the insights from Distributed Morphology and Discourse Representation Theory (B1, B4 and B6). In the

third phase of this SFB we aim to offer a model of the interfaces between morphosyntax and semantics (in B1, B4, B5, B6, B7, and B8), and the way in which linguistic expressions relate to ontology, a topic that will primarily be investigated in B4. This project will focus on the question whether it is possible to model the composition of word meaning as the syntactically driven organization of ontological units, which are directly assigned with roots or which result through application of compositional operations to other such units. Projects B1, B5, B6, B7, and B8 will be mostly concerned with the interaction between morphosyntactic representations and conceptual information in determining meaning, which involves an in-depth investigation of the individual pieces that function as building blocks of meaning.

Second, we aim to offer a model of parallel structural relations between V, N, A, and P domains building on the results of our work from phases 1 and 2 on the basis of the so-called parallelism hypothesis, according to which the structures and the mechanisms available for the formation of verbs, nouns, adjectives and prepositions show a significant amount of overlap. The study of such parallelisms is at the forefront of linguistic research, as it is instrumental for our understanding of how functional material interacts with core elements in the formal organization of meaning across languages. Projects B1, B4, and B7 will be primarily concerned with such parallelisms.

Third, we aim to test the models developed in phases 1 and 2 by integrating non-canonical data, namely phenomena that are subject to a high degree of gradience and remain poorly understood either because they are rare or hard to detect, or because they are subject to a great deal of variation – dialectical, idiolectical, register-based, and even in a single speaker's productions and perceptions over time. One particular domain in which we want to enlarge the empirical basis of our work through the inclusion of non-canonical material is that of compounds based on derived nominals (B1). The study of these types of elements, at the boundary between words and phrases, will be highly informative for theories of nominal meaning and we expect it to tell us more about the strategies that are employed by different languages to cope with ambiguities in the nominal domain. Another new domain, investigated in B9, are meaning shifts that do not involve systematic changes in argument structure but change the contexts in which the linguistic expression is used (e.g. the feminine suffix *-in* or the diminutive suffix). The investigation of understudied types of nominalizations (B4, B7) will complement our model of nominal meaning. In addition, B9 will offer a distributional model of *-er* and *-ung* nominals of the type studied in B1 and B4 in the previous funding periods, which will be able to capture the various readings these can have and also incorporate non-syntactic kinds of meaning shifts as described above. Projects B1 and B7 will further investigate adjectival compounds whose head is a participle and whose non-head acts as a manner or agentive modifier, the aim being to contribute to the better understanding of idiomatic interpretation across categories (parallelism hypothesis).

In the verbal domain, we will turn to an investigation of phenomena that challenge the model we have been pursuing: B5, B6, and the new project B8 will deal with understudied argument alternations, B4, B5, B6, and B9 with (often marked) meaning shifts, and B5 and B6 with unexpected readings of verbal predicates in certain contexts. In particular, B5 and B6 propose to approach the domain of psychological predication under a new perspective by shifting the focus of attention from the *Linking Problem* to argument alternations. B8 will investigate the stative-locative alternation and similar locative alternations in German and across languages. B5 will examine predicates with conative construals. B4, B5, B6 and B9 will look at the availability of novel readings of predicates in specific contexts, often triggered by the presence of particles (especially in the case of B4 and B9). In B5 such meaning shifts include verbs with non-culminating construals, in B6 they involve change-of-state verbs and physical verbs that can be coerced intopsychological interpretations. B4 and B5 share a common interest in the question of how outer or supra-lexical Aspect interferes with the interpretation of verbal predicates. B6 and B8 share a common interest in understanding the role of Voice in argument alternations. The envisaged collaboration between B6 and the new project B8 will enable us to further develop our integrated theory of Voice, and determine the role of Voice in argument alternations and binding.

Diachronic work packages will be included in projects B5, B6 and B7, which will help us determine whether the unsystematic synchronic picture is related to general processes of language change, or particular language development. One project, B3, will not continue in phase 3, as explained above.

Non-canonical data can be empirically hard to evaluate. To tackle them, we will continue benefiting from a variety of research methods. In some cases, the empirical identification of the phenomenon can only proceed via searches of large corpora that deviate from the standard written and spoken ones, and are based on more spontaneous spoken and written texts (e.g. interviews, etc.). For instance, B9 will base its work on web corpora. These include “non-canonical” phenomena such as very long, very short, or ill-formed sentences, and B9 will investigate to what extent this is a problem. In addition, psycholinguistic investigations will provide insights that will complement corpus-

baseddiachronic and synchronic research. All projects will contribute either directly or indirectly to the **silver standard**, by providing feedback on the accuracy of the annotation for their purposes. All projects will make resources available to INF.

The model of the syntax-semantics interface we have been developing in area B, strengthened in the third phase by the addition of B8, is based on the principle of compositionality: meaning is built on the basis of syntactic operations. Our research has attempted to combine this particular view on building word meaning with the ways the immediate context surrounding a particular expression gives information about its meaning. Other models developed and used within this SFB follow the primarily usage-driven strategy of "distributional modeling" that observes words' contexts in large corpora and represents their meaning in terms of algebraic objects (vectors, matrices, etc.). Here, the principle of compositionality arises just as in symbolic meaning construction, but its application to algebraic objects is still a topic of intense research. The algebraic (data-driven) and the symbolic (theory-driven) approach have a great potential to complement each other, but this requires an intense dialogue and adaptation on both sides. Our SFB is ideally suited for such an exchange and can profit from combinations of different perspectives, especially because of the addition of project B9 to area B (as well as several projects adhering to distributional semantics in area D).

Collaborations

Our commitment to close collaboration between computational and theoretical linguistics will continue in the third phase, with multiple collaborations across areascomprising three facets. First, several projects share a common interest in the same topic, e.g. compounds in the case of B1 and D11, prefix verbs in the case of B4, B9 and D12, and meaning shifts in the case of B5, B6, B9 and D12. Second, projects in area A and D will make tools available to projects in area B that we will use to test some of our hypotheses. For instance, B1 will carry out corpus search of the multi-lingual parallel Europarl Corpus in collaboration with D11. In collaboration with INF, B1 will also search canonical (newspaper-based) and non-canonical (web-based) Romanian and English corpora for compounds that lack the genitive/possessive marking. Project B4 together with projects B8, B9, D12 and INF will aim to empirically validate the theoretical hypotheses it has developed against large data. B5 will provide INF with lexical data, corpus data and tools for earlier stages of French. Additionally, B5 will be responsible for creating the metadata necessary for integrating the data from other B projects into INF. Project B6 will collaborate with A2 and D12, and apply tools developed in these projects in order to provide a taxonomy of psychological predicates. B7 will test its hypotheses with the help of the resources developed by B5 and contribute to these resources by sharing their analyses. Projects B1 and B8 will collaborate with D8 in the annotation of non-canonical data, which will contribute to the **silver standard.** Third, a dialog between distributional approaches and symbolic approaches will help systematize the empirical picture in areas where this is rather murky, e.g. psych predicates.

To further strengthen collaborations within area B and across areas, we will co-organize two thematic workshops in 2015 and 2016. The workshop in 2015 will focus on the parallelism hypothesis introduced above. Our second workshop in 2016 will be concerned with the use of non-canonical data in promoting computational and formal models, of the type developed and envisaged in area B.

1.2.2.4 Area C: *Noun phrases and context*

Characterization of area C

The projects in area C aimed to provide detailed analyses of the interaction of the different linguistic levels that lead to a particular interpretation of a linguistic expression with a special focus on the nominal domain. In particular, area C pursued an approach to disambiguation that takes this to be the result of an interaction of various parameters at distinct levels such as syntax, semantics, discourse and information structure. As in area B, several methods were used, including experimental studies as well as diachronic investigations and crosslinguistic comparison. We will not continue the projects in area C, their guiding questions and scientific results will be incorporated in other projects within our SFB, most notably in area B.

Results of the second funding period

Project C2 refined the notion of specificity and its relationship to topicality and discourse prominenceby experimentally investigating the properties of peculiarly marked indefinites, extending the domain of Differential Object Marking (DOM) to Differential Subject Marking (DSM), and looking at verbal object alternations. Further examination of the parameter of specificity responsible for DOM in many languages has provided new insights into the relation between specificity and topicality. C2 proposed a particular formal analysis of specificity in terms of referential anchoring, which proved helpful for a comparison of specificity markers in different languages. A detailed investigation of special markers of indefiniteness such as *pe*-marked indefinites in Romanian and indefinites with *this*

in Englishrevealed that, in addition to marking semantic prominence in terms of specificity, such markers may have a special discourse-pragmatic function marking prominence of referents in the subsequent discourse, i.e., their referential persistence. This finding has opened a new perspective on specificity, normally seen as a backward-looking phenomenon, as having also a forward-looking discourse structuring function. In collaboration with C4, C2 has proposed parameters for the characterization of discourse persistence and developed a method to test these parameters experimentally. The investigation of the DOM in Mongolian, which serves as a specificity marker in some contexts, has suggested that in the context of DSM it indicates semantic prominence of embedded subjects to ensure their distinguishability from the matrix subject. All in all, our investigations have led to systematic insights about the interaction of the sentential level with the discourse level in specifying the contribution of special markers of indefiniteness.

Project C4 has developed semantic and discourse-pragmatic analyses of discourse particles such as *nämlich*, *aber*, *auch*, and *doch* in 'Nacherstposition' (post-initial position) as opposed to other positions in German. A second major result is introducing the concept of Discourse Structuring Potential (=DSP), which is a well-structured alternative to notoriously vague concepts such as salience and thus enables more fine-grained analyses and experimental investigations. This new concept may account for the discourse effects of several referential expressions (definites and indefinites) in German, English, and Romanian. In collaboration with C2, the processing of discourse functions of optional object marking (accusative case) in Turkish was also examined in corpus-related work and a series of experiments.

Another domain of investigation has been the study of specific and unspecific readings for different kinds of indefinites. In a comparative experimental study, we examined the interpretation of indefinites with heavy accent on the determiner as opposed to indefinites with unaccented determiner in German and contrasted different lexical options in English (e.g.: *a* vs. *one*). The results of this study also bear on several theoretical issues in the domain of exceptional wide scope readings and specificity. Furthermore, we have investigated several – predominantly indefinite – understudied determiners in German (like *son*, *dieser*, or *der und der*) and other languages with respect to their semantic and discourse-related properties. The results of C4 add to the recent development of providing more fine-grained analyses of the determiner systems in various languages.

Several factors have led us to reorganize this area of our SFB. First, there were several changes relating to the participating PIs, see section 1.3. Second, a large part of our research in this area has been made available to the linguistic community by the publications listed in the reports of C2 and C4 as well the publication (in 2014) of Alexiadou's book on *multiple determiners* (a book that resulted from the research in project C1 in collaboration with C2 in phase 1). Third, the questions that have not yet been systematically addressed such as object alternations in C2 can be better dealt with in area B and thus will be incorporated in the research program of this area (especially projects B4, B6 and B8; the continuation is guaranteed by the fact that Dr. Ljudmila Geist, currently Postdoc in C2, will be a Postdoc in B8).

From a methodological point of view, the in-depth analyses of individual languages in a crosslinguistic perspective on the basis of corpora searches, experimental methodology and under consideration of the diachronic development of individual languages will be incorporated into the research agenda of several projects in area B, and to a certain extent also in area A. Projects in area C have provided important insights into the nature of the relationship between sentence semantics and discourse semantics as well as tools to incrementally determine structure building, which feeds discourse, an important question for projects A5, A6, and B4. Work in this area also highlighted the role of conceptual and pragmatic parameters in determining interpretation, an insight we will build on in the research program in the third phase of area B. The results of the work relating to the structure of noun phrases, discourse particles, specificity and fine-grained properties of determiner systems will certainly prove profitable for projects A5, A6 and to a certain extent B1, although the overall goals of these projects are, indeed, distinct. Finally, the results of the research on DOM will feed certain WPs in B6 (which will examine the function of clitic-doubling, viewed in the literature as a DOM marker).

1.2.2.5 Area D: Disambiguation in context

Characterization of area D

The projects in area D focus on *computational* approaches to the modeling of contextually driven choice among alternative candidate analyses for a given surface form, and the reverse process of deciding among alternative realization options for a given underlying meaning representation. The predominant paradigm for capturing this abstract specification process in current computational linguistics and Natural Language Processing (NLP) research adopts a data-driven perspective, i.e., model architectures are developed in such a way that the actual disambiguation decision (the "decoding") is based on some statistical model(s) whose parameters have previously been estimated

on a sample of corpus data (during “training”). Projects in area B adopt this data-driven view and operate with state-of-the-art statistical modeling techniques. However, the goal is not just to tune models in order to maximize performance on standard test data, but to deepen our systematic understanding of the implications of particular representational and architectural decisions for models of the specification process.

Within the data-driven paradigm, the SFB’s fundamental research questions translate as follows: what representations and hard constraints should characterize the set of candidates and their properties so as to make (a combination of) data-driven models effective in the prediction of the outcome for unseen data? Besides the objective of following systematic modeling principles (e.g., in terms of crosslinguistically justified interface representations), which is shared with theoretically oriented projects, tractability of the algorithmic traversal of the search space and the effectiveness of parameter estimation from corpus data (with or without a gold standard annotation) play a decisive role. Since practically any concrete disambiguation/specification task involves effects that span more than one linguistic level, data-driven NLP research sheds a very interesting light on dependencies among sub-modules and on the justification of certain interface representations: leaving certain specification decisions undecided in a particular processing step, or modeling a number of decisions from different levels in an integrated, “joint” model has empirical as well as computational implications. The investigation of joint inference techniques is currently receiving great attention in the NLP community. Moreover, in some of the projects in area D, the computational models are construed to directly reflect certain aspects of the cognitive processing architecture and are hence validated against psycholinguistic evidence. The range of linguistic phenomena and of types of input-relations covered in the D area is wide; and although all projects in the second and third phase use statistical or machine learning techniques, the formalisms, model classes and training approaches are diverse. Lastly, since the models developed in some of the D area projects can be applied as robust analysis tools on corpus data (part-of-speech taggers, parsers, coreference resolution tools), these projects play a role in collaborations across areas making use of automatic annotation or training specialized statistical tools for corpus-linguistic investigations.

Results of the second funding period

A major focus for several D projects in phase 2 (D2, D4, D8) was on advanced computational models for syntactic processing, exploring the exploitation of multiple, often subtle contextual factors in the development of sophisticated data-driven architectures for the various preprocessing steps and submodules of parsing and generation – mostly based on standard treebank datasets from the news domain. Some considerable success on this track is demonstrated for instance by the highly successful shared task participation of a group of researchers from these projects (Björkelund et al. 2013): the SFB’s contribution reached the best parsing results of all contestants both for constituent and for dependency parsing in nine typologically distinct languages.

Project D2 departed from an LFG-based bidirectional architecture with discriminative models characterizing the choice among alternatives in parsing and generation, which used functional structures as the relevant abstraction over realization alternatives. In a continued collaboration with A1, the project explored the interrelation between the information status of referring expressions and choices in grammatical realization. Going beyond the core LFG architecture, D2 undertook a series of studies (which have received considerable attention in the Natural Language Generation community to augment candidate generation with robust techniques from dependency generation, adding mapping steps that introduce a more abstract underlying representation and facilitate an explicit modeling of coherence factors from the local discourse context. Also in analysis, D2 has integrated the ParGram LFG grammar of German with robust dependency parsing and explored the use of grammar induction techniques for the mapping between a non-linguistic structured representation and surface realizations. With its dedicated multi-level bidirectional architecture D2 has been able to perform controlled and linguistically grounded corpus-driven experiments studying subtle interactions among a diverse range of contextually influenced realization decisions, such as word order, diathesis and argument realization/generation of referring expressions.

Project D4 has managed to further refine the highly developed BitPar constituent parser with a subtree reranking component and features reflecting earlier or redundant parallel analysis steps, such as linguistically grounded morphological analysis of German and features from a dependency parser. D4 has also developed very effective morphological language models and an implementation for higher-order conditional random field models that is able to deal with large tagsets (> 1000 tags) and large contexts (> 4 context tags). An additional focus that D4 has developed in the second phase has been on statistical machine translations; following the SFB approach of careful consideration of the relation between different processing steps, the Operation Sequence Model can condition the probability of the next translation action on the preceding actions, thereby allowing for an intergrated modeling of translation and reordering steps. Because of the move of its principal investigator Helmut

Schmid to LMU München, D4 will not be continued in the third phase; however, analytic tools from the project will be applied in the silver standard corpus annotation.

Project D8 was added to the SFB through a *Nachantrag*, since the PI, Jonas Kuhn, joined the Universität Stuttgart in 2010. With its main focus on dependency parsing, a topic of considerable current interest in the NLP community, the project has been able to make some important contributions to recent debates and considerations. In line with the architectural considerations in the SFB, the D8 researchers have investigated consequences of various distinct ways of feeding information from different relevant levels of representation into a data-driven approach to part-of-speech tagging, morphological and (dependency) syntactic analysis, andcoreference resolution. This has led to improved approaches in transition-based and graph-based dependency parsing, and combinations of the feature schemata from these approaches. Considerable effort has been put into the investigation of different decoding algorithms and in particular the combination of morphological and syntactic information sources. Carefully designed joint models were shown to have an advantage over a pipeline optimized on system predictions, and we performed some successful detailed studies on the best ways of enforcing hard structural and formal constraints over the space of possible candidate analyses, while allowing for statistical modeling at various levels to pick up even subtle tendencies from the data: work produced by D8 reports a linguistically constrained Integer Linear Programming parser that outperforms state-of-the-art purely data-driven parsers for German and Czech, by translating underspecified morphosyntactic representations that reflect syncretisms into an appropriate, conservative set of constraints. Generalizing the modeling approach beyond dependency parsing in a narrow sense, and covering a slightly larger window of the local, structurally influenced context of linguistic expressions, D8 has extended its scope during phase 2 to also include co-reference resolution. Here, too, the project achieved very competitive results in a shared task including three distinct languages, and the modeling approach has been further developed to allow for the inclusion of richer feature schemata and parameter estimation techniques that make better use of limited amounts of training data. D8 has generally provided the analysis tools it has been developing to the community, for instance in the CLARIN resource infrastructure, and various projects in the SFB and beyond have built on corpus data annotated with the D8 tools.

Projects D6 and D7 share an interest with D2, D4 and D8 in the interpretation options for contextually embedded syntactic configurations – however, each with a very specific modeling goal: D6 takes a psycholinguistic view on a specific set of phenomena, and D7 investigates how the sentiment analysis task can be informed by contextual information.

In corpus studies, psycholinguistic experiments and computational modeling work, project D6 has addressed the phenomenon of logical metonymy in verb+object combinations like *begin the book*, which imply a covert event (*reading*). The focus of D6 has been on investigating the actual range of covert events produced and understood and the role of context in this specification process. We were able to make a novel contribution by showing that metonymy interpretation is shaped largely by world knowledge in a way that goes far beyond what is usually assumed as part of the lexicon. Furthermore, the influence of world knowledge emerges very early and dynamically in the reading process. These findings are problematic both for traditional lexicon-based and for pragmatics-based accounts of metonymy processing. After the original principal investigator, Sebastian Padó, accepted an offer from the University of Heidelberg and left the SFB during the early days of phase 2, Sabine Schulte im Walde took over as a PI. There will not be a direct continuation of the D6 project in the third phase, but strands from the project will be continued in new projects, e.g. D10 and D12, led by Sebastion Padó, who has now returned to the IMS, and Sabine Schulte im Walde, respectively.

In project D7, the SFB's method of scrutinizing the role of different types of contextual information has been applied to the sentiment analysis task, making a systematic distinction between local contexts (represented through structured linguistic units) and global contexts (such as document-wide context, topics, and cultural influences), and lastly crosslingual contexts, i.e., properties of sentiment across languages. D7's principal investigator, Hinrich Schütze, received and accepted an offer from LMU München, so this project will not be continued in the third phase.

Plans for the third funding period

Both continuing projects from the previous phase are in a position in which they can build on very successful modeling results using non-trivial combinations of model components in a canonical scenario, i.e., addressing typical newswire data. D2 will take advantage of the insights from phase 2 to build extended modeling architectures for bidirectional processing capturing grammatical alternations, realization of referents in argument positions and local discourse coherence. The extensions will (a) allow researchers to systematically approach corpus-attested instances of paraphrases or textual edits, viewed as alternative realizations of the same content in a shifted context, and (b) further extend the space of candidates considered as competing realizations by adapting unsupervised techniques from recent work in semantic parsing and thereby relaxing some of the controlling constraints

assumed in phase 2. D2 was led by Christian Rohrer in the first phase, who was joined by Jonas Kuhn as a co-PI for the second phase. For the third phase, Jonas Kuhn will act as the single principal investigator for this project.

Project D8, which has reached state-of-the-art levels of accuracy for its major analysis models (taggers, morphological disambiguators, dependency parsers, and coreference resolution components) on canonical news data will extend its scope to the systematic improvement of cross-domain application of the models and to domain adaptation techniques, including ways of exploiting unlabeled and partially labeled data. This research will not only lead to improved performance of the available analysis tools (which is to the benefit of all users of the automatic annotation), but it will shed further light on the relation between the various levels of linguistic description: with added noise at most levels, how can certain robust pieces of information be identified that provide enough signal to override misleading sources of error? The automatic analysis tools will play a key role in the SFB's silver standard annotation methodology; D8 will provide the best available models for annotating the data, and it will in return benefit from feedback resulting from the shared use and consistency checking of the annotations.

Three new projects in area D (D10, D11, D12) adopt a distributional approach to computational semantics. All three projects are concerned with compositionally predicting the meaning of complex linguistic structures (sentences, particle verbs, compound nouns) from the meaning of the parts and can thus be understood as characterizing specification processes of meaning in context in the continuous vector space-based interpretation outlined above. All three projects pursue a data-driven methodology that is consistent both with the requirements of natural language processing and non-canonical data: they start out from largely unlabeled corpus data and use supervision sparingly, interleaving steps of data-driven modeling with steps of manual evaluation in order to arrive at robust models of interpretation that are applicable to novel occurrences of these, if possible without the need for manual adaptation.

In project D10, the focus is on broad-coverage incremental computation of sentence meaning: new incoming words are interpreted in the context of the preceding words of the sentence. The resulting representations will be evaluated intrinsically (against human similarity judgments) as well as in two extrinsic tasks, (a) assessing the plausibility of incremental parsing hypotheses (in collaboration with D8) and (b) predicting human reading times.

The other two projects, D11 and D12, investigate morphological processes, predicting the meaning of words with internal structure from the meaning of their parts, plus regular meaning shifts arising from the combination. D11 considers noun compounds. Its first specific focus is on architectural considerations for solving the different phases of compound interpretation (splitting – disambiguation – determining implicit relations); its second focus is the use of crosslinguistic information from a parallel corpus as an additional, complementary context for the task. To this end, in collaboration with Project B1 it will carry out corpus search of the multilingual parallel Europarl Corpus. D12 investigates particle verbs, where the major challenge is the ambiguity of base verbs, of particles, and their combinations into particle verbs. The project will first induce senses distributionally and then model the interpretation process by extending the distributional perspective to include psycholinguistic (in particular visual) evidence as well as other types of linguistically motivated features, thus testing the boundaries of the purely distributional approach.

Aspects of the D6 project from the second SFB phase can be found in all three projects: D10 continues the human sentence processing aspect, D11 picks up the question of implicit relations, and polysemy also plays a central role in D12.

Collaborations

The D projects have been involved in a web of bilateral and larger project collaborations, which are described in section 1.2.2.7 and in the individual project descriptions. Some far-reaching collaborations are mediated through the project INF, which ensures that the best available computational models resulting from ongoing research in the D projects are put to use in the annotation of large-scale corpus resources of common interest for the SFB and beyond, the so-called silver standard annotation. The systematic annotation methodology is not only the basis for exploiting intra-level and cross-level tool redundancies for improved exploratory search and quality control of the silver standard. It also provides a controlled channel for crucial feedback from the users of system annotations to the researchers working on modeling, including partial gold standard annotation which can be used for distant supervision of components and can thus inform bootstrapping of the silver standard annotation.

Another recurring aspect of collaborations that the D projects are involved in is related to the status of vector space modeling and its relation to other modeling approaches and analysis components. By addressing a variety of specific phenomena and tasks under this view, the SFB will again be in a position to draw higher-level conclusions at a systematic representational and

architectural level.

Like in areas A and B, there will be dedicated workshops at the end of the first and second year of phase 3 that highlight one of the common research questions in area D. The first workshop will focus on distributional approaches and their relation to models from other paradigms, either at the same level of description or in a configuration of one feeding the other. The second workshop addresses issues in annotation of large heterogeneous collections of corpus data and specifically the ways and limitations for migrating models from canonical data to non-canonical data.

1.2.2.6 The SFB 732 as a whole

We regard our SFB a success: Many projects of the SFB share an interest in the same phenomena but look at them from different angles, defined by different frameworks or assumptions. On the other hand there are projects working on different phenomena which share the same data bases or corpora, or make use of the same LP tools (many of which were developed, adapted or fine-tuned by SFB participants, and often as part of work done within the SFB). These interactions have led to cross-fertilisation within the SFB along different dimensions and have allowed all of us to come to learn and understand not only things of which we knew that we didn't know or understand them well enough, but also many things of which we had previously not even been aware.

In this way we have been able, in ways we would not have been without our cooperations within the SFB, to improve our hypotheses and theories by making them responsive to larger varieties of facts and to competing explanations developed in alternative frameworks. In view of all this we feel that the SFB 732 has been a success – not just because of the results obtained within its individual projects, but also, because of the various ways in which many of those results hang together – a success, in other words, as long-term, large-scope interdisciplinary enterprise.

Let us briefly summarize the key directions our research will take in phase 3. In phases 1 and 2, most projects have focused on canonical data to establish their representational and architectural assumptions. Also, interaction with neighboring modules in the broader sequential model architecture had to be viewed under somewhat idealized assumptions about the interface representation and the relevant context factors, likewise the combination of "parallel" modules that model the same cross-level relation under different views and using different research paradigms.

These developments, which have led to the interactive environment that now prevails within the SFB, have– the point has been stressed more than once already – mostly involved what we have been referring to as 'canonical data'. The extension to 'non-canonical data' that will play a central part in phase 3 will also add a further level of complexity to the interactions between the individual projects and thus give a new impetus to the dialog among the different theoretical frameworks and modelling architectures represented within the SFB.

We will continue this intensive dialog among different modeling and theoretical frameworks. To this end, we will put emphasis on

- applying the theoretical and computational models to concrete corpus data and providing annotations at the respective levels,
- compilating lexical databases that record the findings of systematic linguistic studies, and
- building of computational tools that benefit from our improved understanding of specification processes and can (semi-)automatically generate silver standard analyses for non-canonical data.

There has been a growing awareness, in linguistics and language processing research, of the paramount importance that results continue to be accessible once they have been obtained, and in a form in which they can be of direct use to the community. For the results of this SFB, with their wide diversity and the complex relations between them, setting up an environment within which this is possible is a challenge in its own right. To meet this challenge we are proposing, for this third phase of the SFB, a project that will develop the technological and methodological infrastructure that a transparent presentation of these results and an easy access to them presuppose. Essential to the usefulness of such an environment is its compatibility with the data base environments that have been and are being developed by others: Users should not be forced to switch between frameworks when they browse for data from distinct sources and they ought to be able to download data from different sources within a single document. Ensuring such continuity with developments outside the SFB will be one of the infrastructure project's central tasks.