Discourse-Level Annotation for Investigating Information Structure

Ivana Kruijff-Korbayová and **Geert-Jan M. Kruijff**
Computational Linguistics, Saarland University, Saarbrücken, Germany
{*korbay,gj*}*@coli.uni-sb.de*

Abstract

We present discourse-level annotation of newspaper texts in German and English, as part of an ongoing project aimed at investigating information structure from a cross-linguistic perspective. Rather than annotating some specific notion of information structure, we propose a theory-neutral annotation of basic features at the levels of syntax, prosody and discourse, using treebank data as a starting point. Our discourse-level annotation scheme covers properties of discourse referents (e.g., semantic sort, delimitation, quantification, familiarity status) and anaphoric links (coreference and bridging). We illustrate what investigations this data serves and discuss some integration issues involved in combining different levels of stand-off annotations, created by using different tools.

1 Introduction

The goal of this paper is to present a discourse-level annotation scheme developed for the purpose of investigating information distribution in text from a cross-linguistic perspective, with a particular focus on the interplay of various factors pertaining to the realization of information structure. Information Structure (IS) concerns utterance-internal structural and semantic properties reflecting the speaker's/writer's communicative intentions and the relation of the utterance to the discourse context, in terms of the discourse status of the content, the actual and attributed attentional states of the discourse participants, and the participants' prior and changing attitudes (knowledge, beliefs, intentions, expectations, etc.) (Kruijff-Korbayová and Steedman, 2003). In many (if not all) languages, differences in IS motivate variations in surface realization of utterances, such as syntactic structure, word order and intonation. But languages differ in the extent to which they employ various combinations of IS-realization means (Vallduví and Engdahl, 1996; Kruijff, 2001). Modeling these phenomena and their interaction requires understanding IS and its role in discourse. IS is therefore an important aspect of meaning at the interface between utterance and discourse, which computational models of discourse processing should take into account. Unfortunately, there exists no theory that provides a comprehensive picture of IS, explaining its realization cross-linguistically, its representation at the level of linguistic meaning, and its interpretation in context. Employing corpora can help to deepen our intuitive understanding of IS, in order to construct explanatorily more adequate theories.

While the phenomena involved in discourse and IS are themselves complex and not yet fully understood, studying and modeling their interaction is made difficult by proliferating and often underformalized terminologies, especially for IS (cf. the diverging dichotomies, e.g., Theme-Rheme, Topic-Comment, Topic-Focus, Background-Focus, Given-New, Contextually Bound-Nonbound). What is needed is further systematization of terminologies, formalization and computational modeling, and empirical and corpus-based studies.

The goal of the MULI (MUltilingual Information structure) project is to contribute to this effort by empirically analyzing IS in German and English newspaper texts. For this, we designed annotation schemes for enriching existing linguistically interpreted language resources with information at the levels of syntax, discourse semantics and prosody.

The MULI corpus consists of extracts from the Tiger treebank for German (Brants et al., to appear)[1] and the Penn treebank for English (Marcus et al., 1994)[2]. It comprises 250 sentences in German (app. 3,500 tokens) and 320 sentences in English (app. 7,000 tokens). The MULI corpus has been created by extracting a continuous stretch of 21 relatively short texts from the Tiger treebank, and a set of 10 texts from the Penn Treebank. The selection was made so that the texts would be comparable in genre (financial news/announcements).

The morphological, part-of-speech and syntactic information encoded in the treebanks can be reused for our purposes. We add annotations of syntactically marked constructions, prosodic features and discourse semantics. Our approach to annotation at the levels of syntax, prosody and discourse is outlined in (Bauman et al., 2004a; Bauman et al., 2004b). In this paper, we provide

[1] `http://www.coli.uni-sb.de/cl/projects/tiger/`
[2] `http://www.cis.upenn.edu/~treebank/home.html`

more details about the discourse-level annotation.

In §2 we overview the methodological concerns and desiderata we adhere to in designing our annotation schemes. In §3 we present the discourse-level annotation scheme in detail. In §4 we illustrate the multi-level investigation perspective. §5 we briefly describe the annotation tools we use. In §6 we conclude and sketch future work.

2 Methodology

Text samples of varying origin, genre, language and size have been previously annotated with theory-specific notions of IS by various authors. Such data are typically not publicly available, and even if they can be obtained, it is very hard if not impossible to compare and reuse different annotations. More promising in this respect are annotations that include or add some aspect(s) of IS to an existing corpus or treebank. The most systematic effort of this kind that we are familiar with is the Topic-Focus annotation in the Prague Dependency Treebank (Buráňová et al., 2000).

In contrast to other projects in which IS is annotated and investigated, we do not annotate theory-biased abstract categories like Topic-Focus or Theme-Rheme. Since we are particularly interested in the correlations and co-occurrences of features on different linguistic levels that can be interpreted as indicators of the abstract IS categories, we needed an annotation scheme to be as theory-neutral as possible: It should allow for a description of the phenomena, from which 'any' theory-specific explanatory mechanisms can subsequently be derived (Skut et al., 1997). We therefore concentrate instead on features pertaining, on the one hand, to the surface realization of linguistic expressions (the levels of syntax and prosody), and, on the other hand, to the semantic character of the discourse referents (the discourse level).

In designing our annotation schemes, we followed the guidelines of the Text Encoding Initiative[3] and the Discourse Resource Initiative (Carletta et al., 1997). In line with these standards, we define for each annotation level (i) the markable expressions, (ii) the attributes of markables, and (iii) the links between markables (if any).

Syntax The Tiger treebank and the Penn treebank we use as the starting point already contain syntactic information. The additional syntactic features annotated in the MULI project pertain to clauses as markable units, and encode the presence of structures with noncanonical word order that typically serve to put the focus on certain syntactic elements. We include cleft, pseudo-cleft, reversed pseudo-cleft, extraposition, fronting and expletives, as well as voice distinctions (active, mediopassive and passive). We annotate these features explicitly (when not already present in the treebank annotation), to be able to correlate them directly with features at other levels. The annotation scheme draws on accounts of the analysed features in (Eisenberg, 1994) and (Weinrich, 1993) for German and in (Quirk et al., 1985) and (Biber et al., 1999) for English.

Prosody For the prosodic annotation, we recorded one German and one English native speaker reading aloud the texts of the MULI corpus.[4,5] The recordings were digitised and annotated using the EMU Speech Database System ((Cassidy and Harrington, 2001b); `http://emu.sourceforge.net/`).

The markables at the prosody level are intonation phrases, intermediate phrases and words. Their attributes encode the position and strength of phrase breaks, and the position and type of pitch accents and boundary tones, following the conventions of ToBI (Tones and Break Indices (Beckmann and Hirschberg, 1994)) for English and GToBI[6] (Grice et al., in press) for German, which are regarded as standards for describing the intonation of these languages within the framework of autosegmental-metrical phonology.

Discourse At the discourse level, we define as markable those linguistic expressions that introduce or access discourse entities (i.e., discourse referents in the sense used in DRT and alike) (Webber, 1983; Kamp and Reyle, 1993). Currently we consider primarily the discourse entities introduced by "nominal-like" expressions (Passoneau, 1996). We include other kinds of expressions as markable only when they participate in an anaphoric relation with a "nominal-like" expression. For example, a sentence is a markable when it serves as an antecedent of a discourse-deictic anaphoric expression (Webber, 1991); the main verb of a sentence is a markable when the subject of the sentence is a "zero-anaphor", etc. Our annotation instructions for identifying markables are an amalgamation and extension of those of the MUC-7 Coreference Task Definition[7], the DRAMA annotation manual (Passoneau, 1996), and (Wind, 2002).

The attributes of markables in our discourse-level annotation scheme are designed to capture a range of properties that semantically characterize the discourse entities evoked by linguistic ex-

[3] `http://www.tei-c.org/`

[4] We are aware that using recorded speech is not ideal. We nevertheless decided for this approach, as we wanted to work on top of existing treebanks. As far as we are aware, there does not exist a treebank for any of the publicly available speech corpora.

[5] Since prosodic annotation is very time-consuming, we had to concentrate mainly on one language. Thus, we analysed all German texts and restricted ourselves to some English examples. Since individual speaking preferences may vary from speaker to speaker, we will have to record additional speakers in order to be able to come up with generalizable results.

[6] `http://www.coli.uni-sb.de/phonetik/projects/Tobi/gtobi.html`

[7] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`

pressions. Thereby we differ from most existing discourse-level annotation efforts, which concentrate on the linguistic expressions and on identifying anaphoric relations between them (i.e., identifying anaphors and their antecedents). A notable exception is the GNOME project annotation scheme (Poesio et al., 1999): In GNOME, the aim was to annotate a corpus with information relevant for noun phrase generation. This included syntactic, semantic and discourse attributes of nominal expressions. The semantic attributes include, among others, animacy, ontological status, countability, quantification and generic vs. specific reference, which reflect similar distinctions as we make in our annotation scheme.

Besides the semantic properties that characterize discourse entities individually, our annotation scheme of course also covers referential relations between discourse entities, including both identity and bridging. We build on and extend the MUC-7 coreference specification and the coreference/bridging classifications described in (Passoneau, 1996), (Carletta et al., 1997), (Poesio, 2000) and (Müller and Strube, 2001). We represent anaphoric relations between linguistic expressions through links between the corresponding markables. The type of relation is annotated as an attribute of the markable corresponding to the anaphor.

3 Discourse-Level Annotation

Information structure theories describe the phenomena at hand at a surface level, at a semantic level, or at both levels simultaneously, i.e., an expression belongs to some IS partition, in virtue of some information-status of the corresponding discourse entity. For the investigation of IS at the (discourse) semantic level, we thus need more information about the character of the discourse entities introduced by linguistic expressions. We therefore annotated expressions with their discourse referents and their following properties:

Semantic type/sort reflects ontological character of a discourse entity: object, property, eventuality or textual entity. Since the primary focus of our current annotation are discourse entities evoked by nominal-like expressions, most of them denote objects. Objects are further classified according to semantic sorts: human/person, office/profession, organization, animal, plant, physical object, quantity/amount, date/time, location/place, group/collection, abstract entity, other. Properties are classified into either temporal or permanent. Eventuality has sub-classes phase (habit or state) and process (activity, accomplishment, achievement). Textual entities are for now not further classified.

Denotation characteristics of a discourse entity are captured by a combination of attributes, inspired by (Hlavsa, 1975). First, we distinguish between denotational (extensional, referential) and non-denotational (intensional, attributive) uses of linguistic expressions. Denotationally used expressions pick out (specify) some instance(s) of the designated concept(s). The instance(s) can be uniquely specified (=identifiable to the hearer), or specific but not identifiable, or even unspecific (arbitrary, generic – so any instance will do). Generic references are seen as denoting types. An expression is used non-denotationally when it attribute or qualifies, i.e., evokes the characteristic properties of a concept, without actually instantiating it. A typical example of a non-denotationally used expression is a predicative NP, as in "He was a painter".

The annotation of a group of denotation properties is motivated by the need to have a language-independent characterization of the referents as such, rather then the properties of the referring expression, such as (in)definiteness. The latter is a surface reflex of a combination of denotation characteristics, and sometimes may not even be overtly indicated by articles or other determiners.

For the denotationally used expressions, we then analyze what part of the domain designated by the expression is actualy included in the extension. These aspects are annotated in the determination, delimitation and quantification attributes.

Determination characterizes the specificity of the denoted concept instance. *Unique determination* means that the entity is uniquely specified, i.e., the hearer can (or is assumed to be able to) identify the entity/instance intended by the speaker. There may be just one such entity, e.g., as with proper names, or there are possibly more entities that satisfy the description, but the speaker means a particular one and assumes that the hearer can identify it. Anaphoric pronouns are also typically used as unique denotators. Finally, an entity can be uniquely specified through a relation to another entity, or through a relation between expressions in the text. In (Hlavsa, 1975) this is called *relational uniqueness*; it seems to correspond to Loebner's notion of NPs as functions, used in the GNOME annotation scheme.

Existential determination is assigned to entities that are not uniquely specified, that is, the speaker does not assume the hearer to be able to identify a particular entity, but in principle the speaker would be able to identify one. Maybe such unique identification by the hearer is not important for the interaction, it is enough to take "some instance".

Variable determination is assigned when an expression not only does not uniquely specify an entity, but a particular entity cannot in principle be identified, rather, the speaker means an arbitrary ('any') instance. Typical examples are generics, or references to type.

Delimitation characterizes the extent of the denoted concept instance with respect to the domain designated by the expression. The posible values are

total and *partial*, indicating the entire domain designated by the expression is included in the extension, or only a part.

Quantification captures the countability of the denotated concept instance, and if countable, the quantity of the individual objects included in the extension:

- *uncountable* is assigned when it is impossible to decompose the extension into countable distinguishable individual objects, e.g., with mass nouns;
- *specific-single* means quantity of one, e.g., "one x", "the other x";
- *specific-multiple* means a concrete quantity larger than one, e.g., "two x", "both x", "a dozen";
- *unspecific-multiple* means an unspecified number larger than one, e.g., "some x", "many x", "most x".

Familiarity Status is a notion that most approaches to IS use as one dimension or level of the IS-partitioning, for example Given/New in (Halliday, 1985), Background/Focus in (Steedman, 2000), or as the basis for deriving a higher level of partitioning (Sgall et al., 1986).

It is therefore important to capture it in our annotation as an independent feature, so that we can correlate it with other features at the discourse level and at other levels. We apply the familiarity status taxonomy from (Prince, 1981), distinguishing between new, unused, inferrable, textually and situationally evoked entities. We are aware that operationalizing Prince's taxonomy is a tough issue. For the time being, our annotation guidelines give intuitive descriptions of the different statuses, roughly as follows:

- *brand new*: create a new discourse referent for a previously unknown object;
- *unused*: create a new discourse referent for a known object;
- *inferable*: create a new discourse referent for an inferable object;
- *evoked* (textually or situationally): access an available discourse referent.

Annotators' uncertainty or discrepancies between annotators help us to identify problematic cases, and to revise the guidelines where necessary.[8]

Linguistic form encodes the syntactic category of the markable expression. This is not an attribute encoding a semantic property of a discourse entity. We have found it useful to distinguish the following categories:

- *nominal group* is a "normal" NP with a head noun;
- *pronominal* subsumes expressions headed by a personal, demonstrative, interrogative or relative pronoun;
- *possessive* covers possessive premodifiers (typically a possessive pronoun, e.g., "our view", or possessive adjective, e.g., "the Treasury's threat" or in German "newyorker Burse";
- *pronominal adverb* in German, e.g. "daraus" (from that);
- *apposition* and *coordination*;
- *clitic* is used for clitics and in those cases when an expression contains a clitic affix (though not frequent in English and German newspaper text);
- *ellipsis* is used for elliptical (reduced) expressions, which function as nominal-like groups, but contain no nominal head (e.g., "the first"); in case a discourse entity is evoked by a zero argument, e.g., in case of subject- or object pro-drop, a markable is created on a surrogate non-nominal expression, labeled as *zero-arg*; finally, *clause* or *text* are used for markables which are clause and simple sentences, or text segments, respectively (note that these are only markable, when they serve as antecedents to nominal anaphors).

These categories classify the linguistic forms of expressions independently of the categories employed in the syntactic-level annotation. There are also technical reasons for introducing a form-feature, e.g., when some other expression serves as a markable to annotate the attributes of the discourse entity corresponding to a "zero-anaphor" or to a clitic affix.

Referential link encodes the type of relation between the discourse entity corresponding to an anaphoric expression, and the one corresponding to the (most likely) antecedent. The referential links we distinguish are identity (representing coreference) and bridging, further classified into set-membership, set-containment, part-whole composition, property-attribution, generalized possession, causal link and lexical-argument-filling.

The attributes of information status and referential link are related, but we include them both, because the former is a property of a discourse entity, while the latter directly reflects anaphoricity as a property of an expression (the size of it ranging, ultimately, from a word to a segment of a discourse). The relation between anaphoricity and IS is not a straightforward one, and needs further investigation, enabled by an annotation like ours.

4 Multi-level Investigation of IS

We illustrate the different levels of annotation and analysis with an example sequence taken from our English corpus (Figure 1). We considered the syntactic annotation as a suitable starting point for the

[8] Our reason for applying the familiarity taxonomy from (Prince, 1981) is that it addresses the status of discourse entities as such, not other referential properties. For example, the givenness hierarchy in (Gundel et al., 1993) interleaves information status with uniqueness and specificity.

analysis. Where relevant features are detected, we compare the annotation at other levels.

(1) In the 1987 crash, remember, the market was shaken by a Danny Rostenkowski proposal to tax takeovers out of existence. (2) Even more important, in our view, was the Treasury's threat to thrash the dollar. (3) The Treasury is doing the same thing today; (4) thankfully, the dollar is not under 1987-style pressure.

Figure 1: Example from the English corpus

Of the four clauses in the example sequence, three show noncanonical word orders. In (1), the temporal adjunct is fronted, followed by the main predicate *remember* (in imperative mood). Additionally, (1) contains a passive construction bringing the patient in subject position. In (2), subject complement and adjunct (marking stance) are fronted. In (4), an adjunct (againmarking stance) is fronted.

The discourse entity (DE) introduced in the fronted temporal phrase *the 1987 crash* in (1) is extensional, abstract, unique, specific singular, and has the information status of unused (also indicated by *remember*). The DE introduced in the unmarked subject position is extensional, abstract, unique, specific singular, but has the status of inferrable: *the market* can be seen as a bridging anaphor to *the crash*, by means of an argument filling (*crash of the market*). The DEs introduced by the sentence-final expressions in (1) and (2) are also extensional, abstract, unique, specific singular, and both have the information status of new.[9] What appears sentence-final in (1) and (2) are thus two negative things that happened during the 1987 crash. The fronted expression(s) in (2) are not annotated as a DE. The DEs in the unmarked subject positions in (3) and (4) both have the information status of textually evoked, as both expressions are coreferential anaphors to parts of *the Treasury's threat to thrash the dollar*. While the DE referred to by *the Treasury* is an extensional, office, unique, specific singular, that of *the dollar* is intensional, abstract, unique, uncountable. The expression *the same thing* in (3) is anaphoric to *the Treasury's threat ...* in (2), but it introduces a new DE of the same type; its information status is that of inferrable. Finally, the DE introduced in the sentence-final expression *1987-style pressure* in (4) is intensional, abstract, existential, uncountable, and also has the information status of inferable; it is however hard to code it as a bridging anaphor, because it is not clear what relation it would have to what antecedent: if anything, then *a Danny Rostenkowski proposal ...* in (1).

The prosodic analysis shows that the fronted phrase in (2) is not only syntactically but also prosodically prominent (cf. Figure 2): Two peak accents on *even* and *more* highlight these words (with the more pronounced accent on *more* expressing a contrast), whereas the word *important* is deaccented, since the concept of 'importance' is inferable from the context. Furthermore, the adjective construction forms a phrase of its own, delimited by an intonation phrase boundary, which is in turn signalled by a falling-rising contour plus a short pause. The following parenthesis *in our view* also constitutes a single intonation phrase. Here again, *our* is assigned a contrastive accent, while *view* is unaccented.

All remaining content words of the clause receive accents. However, the most 'newsworthy' word, *threat*, is the only one marked by a rising pitch accent (L+H*), indicating its higher degree of importance for the speaker. This interpretation is further supported by the insertion of a phrase break directly after this word. Finally, the high-downstepped nuclear accent (H+!H*) on *dollar* marks this item as being accessible by speaker and hearer (Pierrehumbert and Hirschberg, 1990).

5 Technical Realization

Above we presented a multi-level view on IS annotation, where each layer is to be annotated independently, to enable us to investigate interactions across the different levels. Such investigations involve either exploration of the integrated data (i.e., simultaneous viewing of the different levels and searching across levels) or integrated processing, e.g., in order to discover or test correlations across levels. There are two crucial technical requirements that must be satisfied to make this possible: (i) stand-off annotation at each level and (ii) alignment of base data across the levels. Without the first, we would not be able to keep the levels separate and perform annotation at each level independently, without the latter we would not be able to align the separate levels.

We have chosen XML for the representation and maintenance of annotations. Each level of annotation is represented as a separate XML file, referring to (sequences of) tokens in a common base file containing the actual text data. We keep independent levels of annotation separate, even if they can in principle be merged into a single hierarchy. Parallel aligned texts (e.g., the written and spoken versions of our corpus) are also represented via shared token IDs. A related issue is that of annotation tools. We are not using one generic tool for all levels for the simple reason that we have not found a tool that would support the needs of all levels and still be efficient (Bauman et al., 2004b; Müller and Strube, 2001). Therefore, we prefer to use tools specifically designed for the task at hand. We describe the tools of our choice below.

Prosodic Level The speech data was annotated with the EMU Speech Database System[10] (Cassidy and Harrington, 2001a), which produces

[9] We assume a layman reader. For an economy expert, these entities may have the status of unused.

[10] `http://emu.sourceforge.net/`

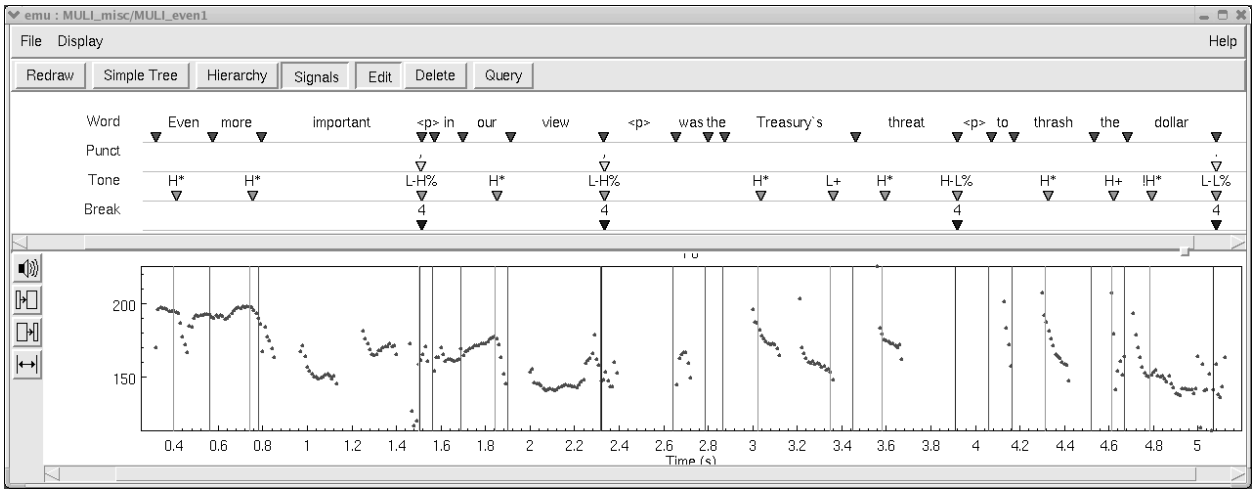

Figure 2: Prosodic annotation of example sentence (2) in EMU

several files in which time stamps are associated with the respective annotated labels.

Syntactic Level For the syntactic annotation, we used the XML editor XML-Spy[11]. The annotation scheme is defined in a DTD, which is used to check the well-formedness and the validity of the annotation.

Discourse Level The discourse-level annotation is done with the MMAX annotation tool developed at EML, Heidelberg (Müller and Strube, 2003). MMAX is a light-weight tool written in Java that runs under both Windows and Unix/Linux. It supports multilevel annotation of XML-encoded data using annotation schemes defined as DTDs. MMAX implements the above-mentioned general concepts of markables with attributes and standing in link relations to one another. To exploit and reuse annotated data in the MMAX format, there is the MMAX XML Discourse API.

Integration The tools inevitably employ different data formats: on the prosodic level data is stored in the EMU data format, on the syntactic level in Tiger XML and on the discourse level in MMAX XML format.

The EMU files have to be converted into stand-off XML format. To be able to align the prosodic annotation with the syntax and the discourse level, we chose the word as common basic unit. This poses several problems. First, punctuation marks count as separate words, but are not realised in spoken language. To be able to correlate prosodic phrasing and punctuation marks, we store the punctuation marks as attributes of the respective preceding word. Second, pauses occur very often in speech, but as they are not part of the written texts, they do not count as words. Because they are an important feature for phrasing and rhythm, we also code them as attributes of the preceding word. Third, in some cases a single word carries more than one accent, e.g. long compounds (*Getränkedosenhersteller*), or numbers. In these cases, it would be interesting to know which part(s) of the word get accented, which requires some way of annotating parts of words (e.g., syllables). Finally, for some multi-word units, e.g. *18,50 Mark*, the spoken realisation (*achtzehn Mark fünfzig*) cannot be aligned with the orthographic form, because spoken and orthographic form differ in number and order of words.

[11] `http://www.xmlspy.com/`

6 Conclusions and Perspectives

We presented the details of the discourse-level annotation scheme that we developed within the MULI project. This project is a pilot project: As such, the annotation has so far been restricted to a relatively small amount of data, since the experimental design of the study required testing of tools as well as manual annotation. We plan to extend the size of the corpus by manual and semi-automatic annotation in a follow-up project.

The challenge in the MULI project has been to define theory-neutral and language-independent annotation schemes for annotating linguistic data with information that pertains to the realisation and interpretation of information structure. An important characteristic of the MULI corpus, arising from its theory-neutrality, is that it is *descriptive*. The corpus annotation is not based on explanatory mechanisms: We have to derive such explanations from the data. (See (Skut et al., 1997) for related methodology pertaining to syntactic treebanks.)

The MULI corpus facilitates linguistic investigation of how phenomena at different annotation levels interact. For example, how do syntactic structure and intonation interact to realize information structure? Or, how does information structure interact with anaphoric relationships? Such linguistic investigations can help to extend existing accounts of information structure, and can also be used to verify (or falsify) predictions made by such accounts. The corpus also makes it possible to construct computa-

tional models from the corpus data.

Theory-neutrality enhances reusability of linguistic resources, because it facilitates the integration with other, theory-neutral resources. To some extent we have already explored this in MULI, combining e.g. Tiger annotation with discourse-level annotation. Another possibility to explore is the to integrate MULI annotation with, e.g., the SALSA corpus (Erk et al., 2003), which provides more detailed semantico-pragmatic information in the style of FrameNet.

Our initial investigation also reveals where additional annotation would be needed. For instance, the text example discussed above constitutes a concession scheme, which we cannot identify without annotating discourse/rhetorical relations. This in turn requires extending the annotation scheme to non-nominal markables.

Acknowledgements

We would like to thank Saarland University for funding the MULI pilot project. Thanks also to Stella Neumann, Erich Steiner, Elke Teich, Stefan Baumann, Caren Brinckmann, Silvia Hansen-Schirra and Hans Uszkoreit for discussions.

References

S. Bauman, C. Brinckmann, S. Hansen-Schirra, G.-J. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich. 2004a. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *Proc. of the Workshop on Frontiers in Corpus Annotation, held at the NAACL-HLT 2004 Conference*.

S. Bauman, C. Brinckmann, S. Hansen-Schirra, G.-J. Kruijff, I. Kruijff-Korbayová, S. Neumann, E. Teich, E. Steiner, and H. Uszkoreit. 2004b. The MULI project: Annotation and analysis of information structure in German and English. In *Proc. of the LREC 2004 Conference*.

M. E. Beckmann and J. Hirschberg. 1994. The ToBI annotation conventions. Ms. and accompanying speech materials, Ohio State University.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Longman, Harlow.

S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60.

S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. to appear. TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation (JLAC), Special Issue*.

E. Buráňová, E. Hajičová, and P. Sgall. 2000. Tagging of very large corpora: Topic-focus articulation. In *Proc. of the 18th Conference on Computational Linguistics (COLING'2000), July 31 - August 4 2000*, pages 139–144. Universität des Saarlandes, Saarbrücken, Germany.

J. Carletta, N. Dahlbäck, N. Reithinger, and M. A. Walker. 1997. Standards for dialogue coding in natural language processing. Report on the dagstuhl seminar, Discourse Resource Initiative, February 3–7.

S. Cassidy and J. Harrington. 2001a. Multi-level annotation in the emu speech database management system. *Speech Communication*, 33(1-2):61–78.

S. Cassidy and J. Harrington. 2001b. Multi-level annotation in the EMU speech database management system. *Speech Communication*, 33(1-2):61–78.

P. Eisenberg. 1994. *Grundriss der deutschen Grammatik, 3. Aufl.* Metzler, Stuttgart, Weimar.

K. Erk, A. Kowalski, S. Padó, and M. Pinkal. 2003. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In *Proc. of ACL 2003*, Sapporo, Japan.

M. Grice, S. Baumann, and R. Benzmüller. in press. German intonation in autosegmental-metrical phonology. In Sun-Ah Jun, editor, *Prosodic Typology: Through Intonational Phonology and Transcription*. OUP.

J. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, (69):274–307.

M. A.K. Halliday. 1985. *Introduction to Functional Grammar*. Edward Arnold, London, U.K.

Z. Hlavsa. 1975. *Denotace objektu a její prostředky v současné češtině [Denotating of objects and its means in contemporary Czech]*, volume 10 of *Studie a práce lingvistické [Linguistic studies and works]*. Academia.

N. Ide, P. Bonhomme, and L. Romary. 2000. Xces: An xml-based standard for linguistic corpora. pages 825–830, Athens, Greece.

H. Kamp and U. Reyle. 1993. *From discourse to logic*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Geert-Jan M. Kruijff 2001. *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*, Faculty of Mathematics and Physics, Charles University. Prague, Czech Republic.

I. Kruijff-Korbayová and M. Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information: Special Issue on Discourse and Information Structure*, 12(3):249–259.

M. Marcus, G. Kim, M. Ann Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger˙1994. The Penn treebank: Annotating predicate argument structure. In *Proc. of the Human Language Technology Workshop*, San Francisco, Morgan Kaufmann.

D. McKelvie, A. Isard, A. Mengel, M.B. Moller, M. Grosse, and M. Klein. 2001. The MATE workbench — an annotation tool for XML coded speech corpora. *Speech Communication*, 33(1-2):97–112.

C. Müller and M. Strube. 2001. Annotating anaphoric and bridging relations with MMAX. In *Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 90–95, Aalborg, Denmark, 1–2 September. `http://www.eml.villa-bosch.de/english/Research/NLP/sigdial`

C. Müller and M. Strube. 2003. Multi-level annotation in mmax. In *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 4-5 July. http://www.eml.villa-bosch.de/english/Research/NLP/Publications.

R. Passoneau. 1996. Instructions for applying discourse reference annotation for multiple applications (DRAMA). draft, December 20.

J. Pierrehumbert and J. Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P.R. Cohen, J. Morgan, and M.E. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT press.

Massimo Poesio, Renate Henschel, Janet Hitzeman, Rodger Kibble, Shane Montague, and Kees van Deemter 1999. Towards An Annotation Scheme For Noun Phrase Generation In *Proc. of the EACL Workshop on Linguistically Interpreted Corpora*. Bergen, Norway.

Massimo Poesio 2000. *The GNOME Annotation Scheme Manual* Available online `http://www.hcrc.ed.ac.uk/~gnome/anno_manual.html`

E. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartik. 1985. *A comprehensive grammar of the English language*. Longman, London.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht, The Netherlands.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Applied Natural Language Processing 1997*, pages 88–95.

M. Steedman. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689.

E. Teich, S. Hansen, and P. Fankhauser. 2001. Representing and querying multi-layer annotated corpora. pages 228–237, Philadelphia.

E. Vallduví and E. Engdahl. 1996. The linguistic realisation of information packaging. *Linguistics*, 34:459–519.

B. L. Webber. 1983. So what can we talk about now? M.I.T. Press.

B. L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

H. Weinrich. 1993. *Textgrammatik der deutschen Sprache*. Dudenverlag, Mannheim u.a.

L. Wind. 2002. Manual zur Annotation von anaphorischen und Bridging-relationen. European Media Laboratory GmbH, August 9.