# Multi-dimensional annotation of linguistic corpora for investigating information structure

**Stefan Baumann, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan Kruijff, Ivana Kruijff-Korbayová, Stella Neumann, Elke Teich**
Saarland University, Saarbrücken

## Abstract

We present the annotation of information structure in the MULI project. To learn more about the information structuring means in prosody, syntax and discourse, theory-independent features were defined for each level. We describe the features and illustrate them on an example sentence. To investigate the interplay of features, the representation has to allow for inspecting all three layers at the same time. This is realised by a stand-off XML mark-up with the word as the basic unit. The theory-neutral XML stand-off annotation allows integrating this resource with other linguistic resources such as the Tiger Treebank for German or the Penn treebank for English.

## 1 Introduction

This paper reports on the project MULI (MUltiLingual Information structure), a pilot study in corpus annotation for the purpose of empirically investigating the distribution of information in texts. As the annotation happened on an empirical basis, it was restricted to a small scale, since the experimental design of the study required testing of tools as well as manual annotation. In a broader picture, we were especially interested in contrasting the annotation schemes as well as our findings for English and German, as the name of the project indicates. MULI is a step to enhance existing linguistically interpreted language resources like the Tiger Treebank for German or the Penn Treebank for English with information on the interface between prosody, syntax and (discourse) semantics. The multilingual design of the study allows us to identify language-specific realisations and preferences of indicators of information structure.

The initial interest was to look into the correlations and co-occurences of features on different linguistic levels that can be interpreted as indicators of information structure.

This resulted in a design of the annotation scheme which was as theory independent as possible. We refrained from annotating abstract categories of information structure like topic - focus or theme - rheme and concentrated on more concrete linguistic phenomena that have been described as indicators of these abstract categories on the different levels (cf. §2).

The challenge is to design, carry out and maintain a corpus annotation to facilitate interpretations of information distribution. Particularly, the different types of linguistic information that are relevant for the analysis of information distribution need to be combined in one resource, where the relevant types are drawn from different linguistic levels. The goal in a *multi-layer* annotation task like this is to keep each layer of annotation intact, while at the same time enabling relating the different layers when analysing the corpus (e.g., in querying) (cf. §3).

In this paper we focus on the description of the annotation scheme and the technical realization of alignment of annotation layers. §2 describes the annotation scheme, discussing an example from the corpus. §3 describes the tools we used for the annotation at each layer and presents our approach to relating the different layers. §4 concludes with open challenges.

## 2 Multi-layer annotation

The MULI corpus comprisess app. 7.000 words in 320 sentences extracted from the English Penn Treebank (Marcus et al., 1994) and app. 3.500 words in 250 sentences from the German Tiger Treebank (Brants et al., to appear). As the Penn Treebank consists of newspaper texts from the Wall Street Journal, we selected the German sub-corpus from the economics section of the newspaper Frankfurter Rundschau (which makes up the Tiger Corpus) to make the contrastive sub-corpora as comparable as possible. The use of existing syntactically annotated corpora enabled us to concentrate on those phenomena that are specific to information structure. The thus achieved annotation gives a comprehensive view on prosodic, syntactic and semantic characteristics of the corpus.

In the following, we discuss the annotation of prosody (§2.1), syntax (§2.2) and discourse semantics (§2.3) and illustrate the annotation on a sequence from the German corpus. The example, given in Figure 1, was chosen because it contains three subsequent sentences annotated with *fronting* on the level of syntax. This category was by far the most frequent in the German syntax annotation and therefore was deemed a suitable starting point for interpreting the integration of the three levels (§2.4).

### 2.1 Prosodic level

In spoken language, prosody (intonation, phrasing, stress, rhythm) is often used to realize information structure, e.g. the pragmatic structure (*focus/background*) or the degree of cognitive activation of individual discourse referents (*given/new*). Accent placement and phrasing are the primary means to mark information structural concepts, but pitch range, rhythm, and speech rate also play an important role.

In order to carry out the prosodic annotation, we recorded a German native speaker reading aloud the German texts of the MULI corpus.[1] These recordings were digitised and annotated on six different levels: (1) word boundaries and pauses, (2) punctuation of the written texts (which are not realised in the spoken version), (3) position and type of pitch accents, force accents and boundary tones, (4) position and strength of phrase breaks, (5) rhythmic phenomena, including non-canonical word stress placement, and (6) comments.

The annotation of level 3 and 4 follows the conventions of GToBI (German Tones and Break Indices), which can be regarded as standard for describing German intonation within the framework of autosegmental-metrical phonology (Grice et al., to appear).

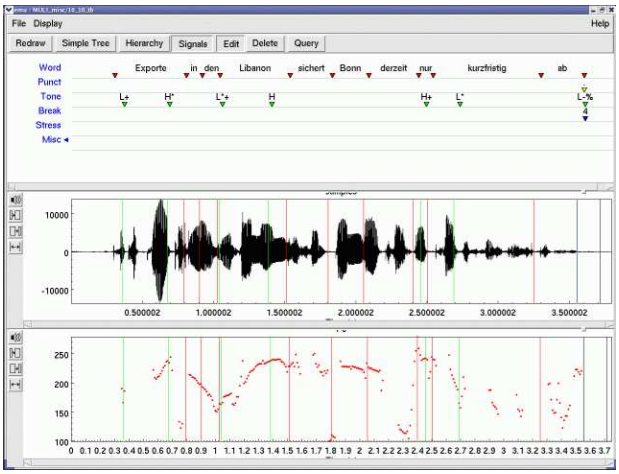


Figure 2: Prosodic annotation in EMU

Below we discuss, the prosodic annotation for the German example sentences.

**Phrasing** Though the "Exporte" sentence is realised as one intonational phrase, we can still divide it into two parts. The first part, ending after *Libanon*, marks the theme of the sentence by a rising intonation contour on the focussed constituent (*Libanon*). The second part marks the rheme by a falling intonation on the focussed constituent (*kurzfristig*). The overall contour is called 'hat pattern', common for regular theme-rheme sentences in German read speech.

**Accent Placement, Accent Type** The two most prominent accents in the sentence are on *Libanon* and *kurzfristig*, marking the end of theme and rheme, respectively. There is another accent on the word *Exporte*, however. At first

[1] Since prosodic annotation is very time-consuming, we had to concentrate on one language, choosing German.

"Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen."
*(This-DATIVE opposes, as Lebanese interlocutors deplore, limited export-guarantees PARTICLE.)*
"Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab."
*(Exports in the Lebanon safeguards Bonn presently only short-term PARTICLE.)*
"Bei aussichtsreichen mittel- und langfristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung."
*(For promising medium- and long-term projects promised Rexrodt now an individual-case-examination.)*

Figure 1: Example sentences taken from the Tiger Corpus

sight, the accent does not seem to be necessary, neither for information structural nor rhythmic reasons. Nevertheless, this weaker accent is appropriate for marking a referent that is inferable from the preceding discourse. In this case, *Exporte* is accessible via a bridging inference from the term *Ausfuhrgarantien* mentioned in the sentence before (cf. §2.3).

The most important information in the sentence is provided by the word *kurzfristig*, since it represents the only new part. Furthermore, it is preceded by the focus particle *nur*, underlining its informativeness. However, the word *kurzfristig* does not receive a peak accent (H*), which is the default marker of new or newsworthy information. It is instead marked by an H+L* pitch accent. (Kohler, 1991) describes this contour as typical for the end of an argumentation or elaboration on something already known. This nuclear contour (H+L* L-%) is also common for marking the end of a paragraph. This can be seen in the third sentence, in which the H+L* accent is assigned to the paragraph final word *Einzelfallprüfung*, although the referent is brand-new in the discourse.

### 2.2 Syntactic level

On the syntactic level, the annotation concentrated on those structures which are relevant for the information structure. These include structures deviating from the canonical word order such as extraposition as well as those structures which serve to focus on certain elements such as clefting and voice, as far as the features were not already explicitly annotated in the treebanks. The syntactic annotation scheme builds on descriptions 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. It comprises cleft, pseudo-cleft, reversed pseudo-cleft, extraposition, fronting, expletive *es* for German and there-insertion for English, as well as active, medio-passive and passive. The unit under investigation on the syntactic level is the clause, i.e. prior to the analysis the corpus was segmented into clauses.

We annotated *fronting*, the feature found in all three sentences in the sample sequence, when an element of the clause structure which typically follows the finite verb is moved to the clause initial position. For German the flexible word order necessitates identifying the element according to its status in the canonical sequence of arguments rather than specifying the element in question according to its respective syntactic function. In clauses annotated with the feature *fronting* another than the most inherent argument precedes the finite verb.

The three sentences of the sample sequence were segmented into four clauses. The second clause beginning with *wie libanesische Gesprchspartner* is not marked with respect to information structure. The fronted elements in the other three clauses are the indirect object realised by the anaphoric pronoun *Dem* in the first sentence, the direct object *Exporte in den Libanon* in the second sentence and the circumstantial *Bei aussichtsreichen mittel- und langfristigen Vorhaben* in the third sentence. All four clauses in the three sentences are annotated in active voice. However, the relevance of voice for realising information structure has to be seen with precaution, since this feature has also other functions such as underspecifying the agent. [2]

[2]The multi-dimensional interpretation in MULI combining three linguistic levels may help to clarify the role of the passive voice in the realisation of information structure, but this is beyond the scope of the present paper.

### 2.3 Discourse level

At the level of discourse semantics we concentrated on linguistic expressions introducing or accessing discourse entities, annotating them with their referential properties and the anaphoric links between them. Below we briefly motivate the choice of referential properties.

Information structure theories describe the phenomena at hand at a surface level, at a semantic level at both levels simultaneously, i.e., an expression belongs to some IS partition, in virtue of some the information-status of the corresponding discourse entity For the investigation of IS at the semantic level, we need more information about the character of the discourse entities introduced by linguistic expressions. We annotated: `Type (intensional or extensional object, property, eventuality or textuality)` and more finegrained `Semantic Sort`; referential properties of `Delimitation (unique, existential, variable, non-denotational` use (Hlavsa, 1975)) and `Quantification (uncountable, unspecific non-singular, specific-nonsingular or specific singular)`; `Information Status (new, unused, inferable, evoked` (Prince, 1981)) and `Form` –because there are correlations with the other features, though it does not necessarily belong to this level.

Besides the properties of individual discourse referents, we annotate anaphoric links. We distinguish between `coreference` and `bridging` anaphoric links. The former is identity of reference, the latter involves an associative relationship between the anaphor and the the antecedent, such as `set containment, part-whole composition, property attribution, generalized appurtenance, causality or lexical argument filling`.

In line with the recommendations of the Text Encoding Initiative[3] and the Discourse Resource Initiative[4], we define what expressions are *markables*, what properties they have as *attributes* and what *links* can hold among them. Our annotation scheme and guidelines build on MUC-7[5] Coreference Specification, DRAMA (Passoneau, 1996), the MATE project[6] and (Müller and Strube, 2001).

In the example sequence, we find instances of both (i) coreference (*dem* und *Rextrodt* –both have unique, textually evoked referents) and (ii) bridging (*libanesische Gesprächspartner* has a unique referent of inferrable status (general appurtenance to the business trip of minister Rextrodt); *Bonn* has a unique referent (the German government) of inferrable status, being the superset containing also the minister; *Exporte in den Libanon* has a referent of the variable type ("any") which has inferrable status given its attribute, the export guarranties mentioned in the preceding sentence; *aussichtsreichen mittel- und langfristigen Vorhaben* has again a variable referent of inferrable status ( general appurtenance to exports). *beschrñkte Ausfuhrgarantien* and *eine Einzelfallprüfung* have referents of existential type of brand new status. All these referents are of extensional type.

### 2.4 Integrated view on the annotation

To see whether the syntactic markedness is matched by distinctive features on the other levels under investigation, we look in more detail at the second sentence. The fronted direct object *Exporte in den Libanon* establishes a thematic relationship to the immediately preceding element in the previous sentence. Semantically, *Exporte* is linked by a bridging relation of the type `attribution` to *Ausfuhrgarantien* in the NP *beschränkte Ausfuhrgarantien.* This way the preceding rheme *beschränkte Ausfuhrgarantien* is thematised. The accent on *Exporte* interpreted as pointing to an inferable referent in (§2.1) substantiates this anaphoric link. In the third sentence this theme is further specified by *langfristige Vorhaben* in thematic position, which is related to it by a bridging relation of the type `general appurtenance`.

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

[4] `http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html`

[5] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html`

[6] `http://mate.nis.sdu.dk/`

In the "Exporte" sentence the subject *Bonn* follows the finite verb. The position in the *Mittelfeld*, syntactically the least focussed position in the sentence, already suggests its unfocussed character. The findings on the level of prosody corroborate this assumption. Another accent might have been expected on *Bonn*, since in read speech normally every NP receives one. Here, however, it has obviously been regarded as being too 'given' to 'deserve' intonational prominence. Indeed the agent does not contribute much information to the message. Were the sentence passive, the syntactic structure would remain the same and the loss of additional information expressed by the agent would be minimal. The object in fronted position realises a link to the immediately preceding discourse entity and thus serves to make the present sequence coherent. In canonical word order with the subject in initial position the passage would lose its focus on the subject matter of exports.

The circumstantial *kurzfristig* again realises an anaphoric link to the rhematised subject of the previous sentence. It specifies the limitation of the export guarantees and is emphasised lexically by the focus particle *nur*. As explained in (§2.1) against all expectation *kurzfristig* does not receive a peak accent. The combination of the levels reveals that the sentence as a whole is an elaboration of the preceding rheme.

When looking at the prosody of the example, we obtain yet another view on the information structure. As seen in (§2.1) the fronted element does not receive a marked accent but rather the phrasing typical for an unmarked theme. Thus, from the point of view of prosody, the specific pattern of *fronting* is not discernable.[7]

With respect to the following sentence the circumstantial *kurzfristig* implies a link to the premodifier *mittel- und langfristig* which together form some kind of a member-set relationship.

The example shows that rather than the features on the different levels all concentrating on certain elements in the sentence, the specificities on each level spread over different elements in the text. Level-specific annotation thus forms a complex picture giving comprehensive evidence on how the information is structured in the text. This division of work between the levels requires an integrated view on information structure that is best represented in a stand-off annotation.

## 3 Technical realisation

In the previous sections we presented a multi-layered view of information structure, where each layer is to be annotated independently, in order to allow us to investigate interactions across the layers. 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 correlations across layers. There are two crucial technical requirements that must be satisfied to make this possible: (i) stand-off annotation at each layer and (ii) alignment of base data across the layers. Without the first, we would not be able to keep the layers separate, without the latter we would not be able to align the separate layers.

There are currently basically two kinds of approaches dealing with multi-layer corpus annotation: one is committed to a hierarchical data structure — either implemented as relational databases (e.g., EMU (Cassidy and Harrington, 2001)) or in XML (e.g., XCES (Ide et al., 2000) or MATE (McKelvie et al., 2001)) — using XPointers or ordered directed acyclic graphs (ODAGs) to represent intersecting hierarchies. The other is based on the concept of temporality using directed arcs representing nodes at certain time offsets on a given timeline — this concept is reflected in the annotation graph (AG) as proposed by (Bird and Liberman, 2001).

Our approach to multi-layer annotation is based on both of these concepts (see also (Teich et al., 2001)): from AGs we adopt the principle of modularity, i.e., we keep independent layers of annotation separate, even if they can in principle be merged into a single hierarchy. From ODAGs

[7] The corpus evidence seems to indicate that *fronting* on the level of syntax is correlated with the realisation of an intermediate phrase in combination with rising accent on the prosodic level. But as this is a typical property of initial accents it has only limited explanatory power.

we adopt the possibility to explicitly represent structural hierarchies — if they are meaningful. Like in MATE, we have chosen XML as the concrete data format to maintain and represent annotations, in order to take advantage of the rich set of readily available tools to edit, validate, transform, and query XML. This means the following for our stand-off annotation: The raw text of the corpus is represented as a flat text file — each token carrying a unique ID. Each layer of annotation is represented as a separate XML file. Each markable containing the linguistic information is equipped with the ID or the IDs of the corresponding text sequence. Parallel aligned texts (e.g. spoken and written versions of the corpus) are also represented via shared IDs in order to refer to overlapping segments.

A related issue is that of annotation tools. One possibility would be to use a generic tool supporting the annotation of either of the levels we are concerned with (and any levels that we may want to add in the future). The most ambitious projects we are familiar with are the MATE Workbench[8] and the follow-up NITE Workbench[9] for multi-level, cross-level and cross-modality annotation of language data. The MATE Workbench has been developed as a highly customizable tool for parallel annotation of arbitrary and possibly non-hierarchical layers of linguistic description. It is an open source tool written in Java and handles XML-encoded data. Unfortunately, practical experiences outside the MATE project itself have been rather negative (e.g., (Müller and Strube, 2001)). The NITE Workbench has equally ambitious goals, however, the available prototype version is implemented in C++ and only for Windows.

It is quite likely that any tool aimed at being entirely generic will run into problems of efficiency for (most of) the individual layers, because of too much additional processing overhead. Therefore, we prefer to use tools specifically designed to support the particular annotation task(s) at hand. We describe the tools of our choice below.

[8] `http://mate.nis.sdu.dk`

[9] `http://nite.nis.sdu.dk`

**Prosodic Level** Two people annotated the spoken data with the EMU Speech Database System[10] ((Cassidy and Harrington, 2001)) following our annotation scheme (§2.1). For each of the six annotation levels, EMU produces a file in which time stamps are associated with the respective annotated label.[11] The EMU files have to be converted into stand-off XML. To be able to align the prosodic annotation with the syntax and the discourse level, we chose the word as common basic unit. The prosodic annotation itself can be organised in a strictly hierarchical fashion: Each intonation phrase (IP) carries one boundary tone and consists of one or more intermediate phrases. Each intermediate phrase (ip) carries one boundary tone and consists of one or more words. Each word can be associated with one or more (pitch or force) accents.

The word as basic unit poses several problems for the prosodic level, though. 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. 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.

**Syntactic Level** For the syntactic annotation, we used the XML editor XML-Spy[12]. The annotation scheme is defined in a DTD (see below), which is used to check the well-formedness and the validity of the annotation.

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

[11] I.e. for the prosodic level the basic unit is actually the sample.

[12] http://www.xmlspy.com/

**Discourse Level** Two people annotated the MULI corpus for the discourse annotation; (one of the developers of the annotation scheme and one annotator who was only instructed by the annotation guidelines), using the MMAX annotation tool developed at EML, Heidelberg (Müller and Strube, 2003). MMAX is a lightweight tool written in Java rhat 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 exploite and reuse annotated data in the MMAX format, there is the MMAX Discourse API.

**Integration** Of course, the tools output different data formats: on the prosodic level in the EMU data format, on the syntactic level in Tiger XML and on the discourse level in MMAX XML format. Hence it is necessary to either convert these "native" representations into a single data format that can be processed by one common tool, or implement an interface for each format that can retrieve the relevant information. Initial steps in both directions have been made by implementing a Java API using Tiger XML as its common data format.

Figures 3-6 show how the raw text of the sample is encoded and how the annotation spans of each annotated markable refer to the IDs of the raw text in the example. The annotation levels (prosody, syntax, disourse) are kept separate. The time spans of the prosodic annotation are aligned with base words.

```
<word id="s5981_1">Exporte</word> <word id="s5981_2">in</word>
<word id="s5981_3">den</word> <word id="s5981_4">Libanon</word>
<word id="s5981_5">sichert</word> <word id="s5981_6">Bonn</word>
<word id="s5981_7">derzeit</word> <word id="s5981_8">nur</word>
<word id="s5981_9">kurzfristig</word> <word id="s5981_10">ab</word>
<word id="s5981_11">.</word>
```

Figure 3: Corpus representation

```
<markable id= "markable_192" span="s5981_1..s5981_11"
ISsyntax="fronting" voice="active"/>
```

Figure 4: Syntactic level

```
<intonphrase id="s5981_IP1" boundarytone="L-%">
   <intermedphrase id="s5981_ip1" boundarytone="L-">
      <word id="s5981_1" starttime="0.29775" endtime="0.791935">
         <gtobiaccent id="s5981_1_acc1" type="L+H*" strength="pitch">
         </gtobiaccent></word>
      <word id="s5981_2" starttime="0.791935" endtime="0.901458"></word>
      <word id="s5981_3" starttime="0.901458" endtime="1.02643"> </word>
      <word id="s5981_4" starttime="1.02643" endtime="1.51226">
         <gtobiaccent id="s5981_4_acc1" type="L*+H" strength="pitch">
         </gtobiaccent>
      </word>
      <word id="s5981_5" starttime="1.51226" endtime="1.80432"></word>
      <word id="s5981_6" starttime="1.80432" endtime="2.05496"></word>
      <word id="s5981_7" starttime="2.05496" endtime="2.40108"></word>
      <word id="s5981_8" starttime="2.40108" endtime="2.50288"></word>
      <word id="s5981_9" starttime="2.50288" endtime="3.25129">
         <gtobiaccent id="s5981_9_acc1" type="H+L*" strength="pitch">
         </gtobiaccent> </word>
      <word id="s5981_10" starttime="3.25129" endtime="3.55107"
       pauseduration="unknown" punctuation="."></word>
   </intermedphrase>
</intonphrase>
```

Figure 5: Prosodic level

## 4 Perspectives

The challenge in the MULI project is 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 ((Skut et al., 1997) is methodologically related). 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 need to derive such explanations from the data.

An important strand of research that the MULI corpus facilitates is the linguistic investigation of how phenomena at different annotation layers interact. For example, how do syntactic structure and intonation interact to realize information structure? Or, how does information structure interact with anaphoric relationships? In this way, 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.

Also, the corpus makes it possible to construct computational models from the corpus data. For example, we can consider the integration of information structure into approaches to grammar learning to provide a bridge between surface phenomena and deeper levels of meaning. Another challenge would be the construction of lexicalized models of salience tracking, useful to determine e.g. the appropriate answer context

```
<markable id="markable_31" span="s5981_1..s5981_4" bridging_link="attribution" obj_subtype="extensional" obj_delimitation="variable"
   sem_sort_of_object="abstr" obj_quantification="unspecific_multiple" attribution_link="attribute-carrier"
   referential_link="bridging" type="object" obj_info_status="inferrable" ling_form="nominal" pointer="markable_29" />
<markable id="markable_32" span="s5981_3..s5981_4" obj_subtype="extensional" obj_delimitation="unique" sem_sort_of_object="loc"
   obj_quantification="specific_single" type="object" referential_link="identity_of_referent" obj_info_status="text_evoked"
   ling_form="nominal" pointer="markable_12" />
<markable id="markable_33" span="s5981_6" containment_link="superset" bridging_link="containment" obj_subtype="extensional"
   obj_delimitation="unique" sem_sort_of_object="organiz" obj_quantification="specific_single" referential_link="bridging"
   type="object" obj_info_status="inferrable" ling_form="nominal" pointer="markable_22" />
```

Figure 6: Discourse level

in Q&A-systems. Furthermore, we can use the data to construct shallow parsers for information structure and discourse structure, possibly using co-training or active learning given the relatively small amount of data we are working with.

Theory-neutrality not only raises interesting challenges when for building explanatory models based on the data. It also 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.

## References


[Biber et al.1999] D. Biber, J. Stig, G. Leech, S. Conrad, and E. Finegan. 1999. *The Longman Grammar of Spoken and Written English.* Longman, Harlow.

[Bird and Liberman2001] S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60.

[Brants et al.to appear] S. Brants, S. Dipper, P. Eisenberg Peter, S. Hansen-Schirra, E. Knig, W. Lezius Wolfgang, R. Christian, G. Smith andH. Uszkoreit. to appear. Tiger: Linguistic interpretation of a german corpus. In E. Hinrichs and K. Simov, eds, *J. of Language and Computation (JLAC), Special Issue.*

[Cassidy and Harrington2001] S. Cassidy and J. Harrington. 2001. Multi-level annotation in the EMU speech database management system. *Speech Communication*, 33(1-2):61–78.

[Eisenberg1994] P. Eisenberg. 1994. *Grundriss der deutschen Grammatik, 3. Aufl.* Metzler.

[Erk et al.2003] 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.*

[Grice et al.to appear] M. Grice, S. Baumann, and R. Benzmüller. to appear. German intonation in autosegmental phonology. In S.-A. Jun, ed., *Prosodic typology.*

[Hlavsa1975] Z. Hlavsa. 1975. *Denotating of objects and its means in contemporary Czech [in Czech]*, volume 10 of *Studie a práce lingvistické.* Academia.

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

[Kohler1991] K.J. Kohler. 1991. Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology, and semantics. *AIPUK*, 25:115–185.

[Marcus et al.1994] M. Marcus, G. Kim, M. A. 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*

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

[Müller and Strube2001] Ch. Müller and M. Strube. 2001. Annotating anaphoric and bridging relations with MMAX. In *Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue*

[Müller and Strube2003] Ch. Müller and M. Strube. 2003. Multi-level annotation in MMAX. In *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*

[Passoneau1996] R. Passoneau. 1996. Instructions for applying discourse reference annotation for multiple applications (DRAMA). draft, December 20.

[Prince1981] E. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, ed., *Radical Pragmatics*, pages 223–256. Academic Press.

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

[Skut et al.1997] 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.

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

[Weinrich1993] H. Weinrich. 1993. *Textgrammatik der deutschen Sprache.* Dudenverlag, Mannheim u.a.