The MULI Project: Annotation and Analysis of Information Structure in German and English

Stefan Baumann, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan Kruijff, Ivana Kruijff-Korbayová, Stella Neumann, Erich Steiner, Elke Teich, Hans Uszkoreit

Saarland University, Saarbrücken, Germany

Abstract

The goal of the MULI (MUltiLingual Information structure) project is to empirically analyse information structure in German and English newspaper texts. In contrast to other projects in which information structure is annotated and investigated (e.g. in the Prague Dependency Treebank, which mirrors the basic information about the topic-focus articulation of the sentence), we do not annotate theory-biased categories like topic-focus or theme-rheme. Trying to be as theory-independent as possible, we annotate those features which are relevant to information structure and on the basis of which typical patterns, co-occurrences or correlations can be determined. We distinguish between three annotation levels: syntax, discourse and prosody. The data is based on the TIGER Corpus for German and the Penn Treebank for English, since the existing information on part-of-speech and syntactic structure can be re-used for our purposes. The actual annotation of an English example sequence illustrates our choice of categories on each level. Their combination offers the possibility to investigate how information structure is realised and can be interpreted.

1. Introduction

MULI (MUltiLingual Information structure) is a pilot study for enriching treebanks with features relevant for investigating the distribution of information in texts. As its name suggests, the project looks from a contrastive angle on this investigation and incorporates different linguistic levels. Thus, 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 syntax, (discourse) semantics and prosody. The multilingual design of the study allows us to identify language-specific realisations and preferences of indicators of information structure. The annotation is restricted to a relatively small amount of data, since the experimental design of the study requires testing of tools as well as manual annotation.

We are particularly interested in the correlations and co-occurrences of features on different linguistic levels that can be interpreted as indicators of information structure. For this purpose, the annotation scheme has to be as theory independent as possible. We refrain from annotating abstract categories of information structure like topic/focus or theme/rheme, concentrating instead on more concrete linguistic phenomena that have been described as indicators of these abstract categories on the different levels.

In this paper we focus on the description of the annotation scheme and how this annotation can serve to enrich the existing treebank resources (§2). We illustrate the annotation in the discussion of an example from the corpus in §3. §4 concludes with the perspectives opened by this study.

2. Annotation

2.1. Corpus design

The MULI corpus consists of extracts from the Tiger Treebank for German ((Brants et al., to appear); http://www.coli.uni-sb.de/cl/projects/tiger/) and the Penn Treebank for English ((Marcus et al., 1994); http://www.cis.upenn.edu/~treebank/home.html). As the Tiger Corpus contains articles from the general newspaper *Frankfurter Rundschau*, we only select texts from the economics section in order to make the corpus as comparable as possible to the *Wallstreet Journal* texts which make up the Penn Treebank. Our corpus comprises 250 sentences in German (app. 3,500 tokens) and 320 sentences in English (app. 7,000 tokens). Part-of-speech information and syntactic structure in the treebanks help with interpreting the distribution of information in the texts.

2.2. Syntax

As the Tiger and Penn Treebank already contain syntactic information, the annotation on the syntactic level concentrates on those features which are specifically relevant for information structure. The Tiger Corpus encodes information on syntactic functions in the edges and phrase categories in the nodes. It also takes account of part-of-speech tags as well as morphology which is very important for an inflectional language like German. As a mixture of phrase structure and dependency analysis, the annotation combines the advantages of both grammars. Initiatives covering other linguistic phenomena on the basis of this annotation include the extraction of topological information (Becker and Frank, 2002) which can be used for the analysis of information structure on the syntactic level. The annotation of the Penn Treebank consists of a phrase structure analysis enriched by part-of-speech information. It differentiates between a number of adjuncts (e.g. temporal and local). Furthermore, the most frequent syntactic functions are included in the annotation.

Beyond these types of syntactic information, the MULI annotation scheme covers noncanonical word order and other syntactic structures that serve to put the focus on certain elements. It draws on accounts of the analysed features as described in (Eisenberg, 1994) and (Weinrich, 1993) for German and in (Quirk et al., 1985) and (Biber et al., 1999) for English. The annotation scheme comprises cleft, pseudo-cleft, reversed pseudo-cleft, extraposition, fronting, expletives, as well as active, medio-passive and passive. Where necessary, the annotation guidelines specify lan-

guage specific realisations of these features. This is particularly the case for the expletive *es* in German and its English equivalent *there*-insertion. The unit under investigation on the syntactic level is the clause, i.e. prior to the analysis the corpus was segmented into clauses.

2.3. Discourse

Information structure (IS) 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 semantic level, we need more information about the character of the discourse entities introduced by linguistic expressions. We therefore annotate expressions with their discourse referents and their following properties: *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)*; the *Form* of an expression (although it does not necessarily belong to this level, but there are correlations with the other features); *Information Status (new, unused, inferable, evoked)* (Prince, 1981). Coding information status is motivated by the fact that IS theories often employ some notion of information status as one dimension of the partitioning on its own, or as the basis for deriving a higher level of partitioning. We use Prince's familiarity taxonomy, which clearly addresses the status of discourse entities as such, not other referential properties.

Besides the properties of individual discourse referents, we annotate anaphoric links between expressions. We distinguish between *coreference* and *bridging*, where there exists an associative relationship between the referents of the anaphor and the antecedent, such as *set-containment, part-whole composition, property-attribution, possession, causality or lexical-argument-filling*. The relation between anaphoricity and IS is not a straightforward one, and needs further investigation, enabled by an annotation like ours.

Our annotation scheme follows the Text Encoding Initiative recommendations (http://www.tei-c.org/) and the Discourse Resource Initiative guidelines (Carletta et al., 1997). In line with these standards, we define what expressions are markables, what attributes they have and what links can hold between them. At the discourse level, markables are "nominal-like" (Passoneau, 1996) linguistic expressions that introduce or access discourse entities (i.e., discourse referents in the sense used in DRT and alike). We build on and extend the reference annotation schemes for MUC-6 and MUC-7 (MUC Coreference Specification), DRAMA (Passoneau, 1996), the MATE project ((Poesio et al., 1999); http://mate.mip.ou.dk), the DRI guidelines (Carletta et al., 1997), (Poesio and Vieira, 1998) and (Müller and Strube, 2001). The corpus has been annotated by two annotators (one of the developers and one only instructed by the annotation guidelines), using the MMAX annotation tool (http://www.eml.villa-bosch.de/english/Research/NLP/Downloads).

2.4. Prosody

In spoken language, prosody (intonation, phrasing, stress, rhythm) is often used to realise the information structure of a text, e.g. the pragmatic structure (*focus/background*) or the degree of cognitive activation of individual discourse referents or propositions (*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 one German and one English native speaker reading aloud the texts of the MULI corpus.[1] Since individual speaking preferences may vary from speaker to speaker, our results are not generalisable, reflecting the experimental character of the study.

The recordings were digitised and annotated on six different levels using the EMU Speech Database System ((Cassidy and Harrington, 2001); http://emu.sourceforge.net/): (1) word boundaries and pauses, (2) punctuation of the written texts, (3) position and type of pitch accents and boundary tones, (4) position and strength of phrase breaks, (5) rhythmic phenomena, including non-canonical word stress, (6) comments.

The annotation of level 3 and 4 follows the conventions of ToBI (Tones and Break Indices (Beckmann and Hirschberg, 1994)) for English and GToBI ((Grice et al., in press); http://www.coli.uni-sb.de/phonetik/projects/Tobi/gtobi.html) for German. They can be regarded as standards for describing the intonation of these languages within the framework of autosegmental-metrical phonology, in which pitch contours are decomposed into high and low tonal targets (symbolised by H and L). Diacritics are listed in Table 1, the tonal and break index inventories are summarised in Table 2.

*	target on the accented syllable
+	target before or after the accented syllable
–	boundary tone of an intermediate phrase (ip)
%	boundary tone of an intonation phrase (IP)
!	downstep of an H tone
^	upstep of an H tone

Table 1: (G)ToBI diacritics

	ToBI	GToBI
pitch accents	H*, L*, L+H* L*+H, H+!H*	H*, L*, L+H* L*+H, H+!H*, H+L*
force accents	–	H(*), L(*)
boundary tones	L–, H–, L–L% H–L%, H–H% L–H%, %H	L–, H–, L–% H–%, H–^H% L–H%, %H
break indices	0, 1, 2, 3, 4	2r, 2t, 3, 4

Table 2: (G)ToBI inventories of tones and break indices

[1] Since prosodic annotation is very time-consuming, we had to concentrate on one language. Thus, we analysed all German texts and restricted ourselves to some English examples.

3. Example

We illustrate the different levels of annotation and analysis with an example sequence taken from our English corpus (Figure 1). We consider the syntactic annotation a suitable starting point for the analysis. Where relevant features are detected, we compare the annotation to 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 sequence from the English corpus

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

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 inferable: *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.[2] What appears sentence-final in (1) and (2) are thus two negative things that happened during the 1987 crash. The evaluation-ascribing adjective phrase in (2) is 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 inferable. 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) (according to one of the annotators).

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

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 due to givenness.

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 (cf.(Pierrehumbert and Hirschberg, 1990)). This means it can neither count as brand new (which normally requires a H* peak accent), nor as immediately given, since it is not deaccented (as is the case with the word *important* above).

4. Conclusions

First experiences with our multilingual multi-layer annotation lead to conclusions with respect to how to automize the annotation process using statistical methods and learning procedures. On the grammatical level, the syntactic annotation of the Tiger Corpus and the Penn Treebank, for example, can be used to determine passive constructions. In connection with the discourse annotation, for instance, the existing part-of-speech tags can be used in order to identify pronominal co-reference.

We are also working on robust methods for identifying information structure, following the annotation scheme (which focuses on the information status of individual markables) as well as investigating the idea of informativity zoning, i.e. the division of clauses/sentences into parts that are more or less informative in the given context.

Furthermore, conclusions can be drawn from the co-occurrence of particular categories on different levels of annotation in MULI, indicating how these different levels are deployed in order to mark information structure.

However, 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.

Using our findings as a tertium comparationis, theory-dependent information structure annotation and thus existing theories on information structure can be compared to and validated against our theory-neutral approach and vice versa. Finally, using our findings on the co-occurrence of the annotated categories, it is possible to compare how different languages use different grammatical, discursive and prosodic means to structure information.

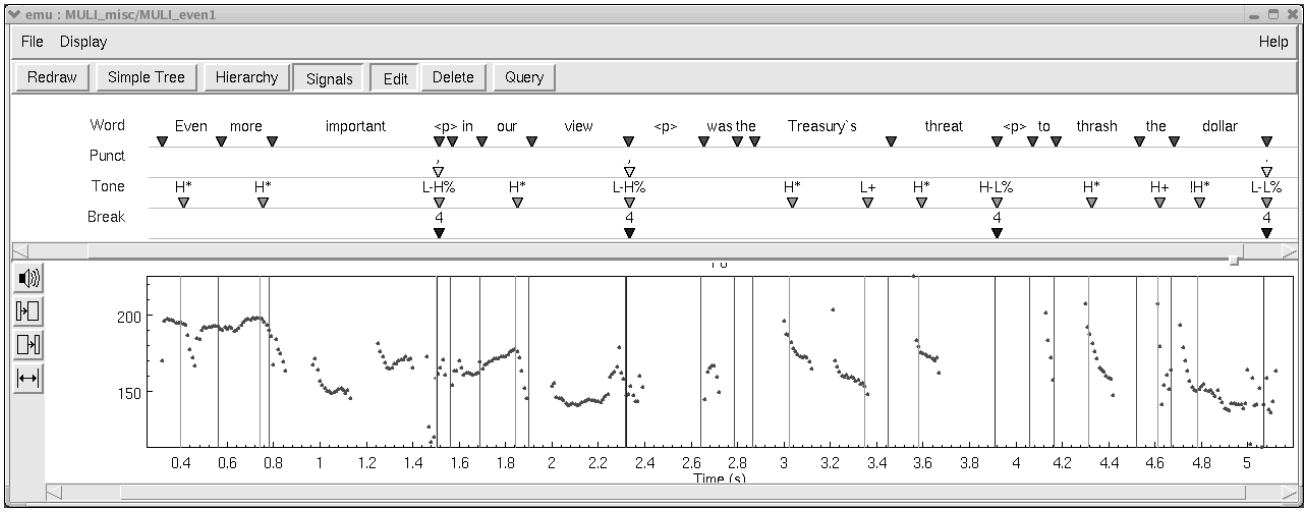

Figure 2: Prosodic annotation of example sentence (2) in EMU

5. References

Becker, Markus and Anette Frank, 2002. A stochastic topological parser of German. In *Proceedings of COLING 2002*. Taipei, Taiwan.

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

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan, 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit, to appear. TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation (JLAC), Special Issue*.

Carletta, Jean, Nils Dahlbäck, Norbert Reithinger, and Marylin A. Walker, 1997. Standards for dialogue coding in natural language processing. Report on the dagstuhl seminar, Discourse Resource Initiative.

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

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

Grice, Martine, Stefan Baumann, and Ralf Benzmüller, in press. German intonation in autosegmental-metrical phonology. In Sun-Ah Jun (ed.), *Prosodic Typology: Through Intonational Phonology and Transcription*. OUP.

Hlavsa, Zdeněk, 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.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop*. San Francisco, Morgan Kaufmann.

Müller, Christoph and Michael Strube, 2001. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark.

Passoneau, Rebecca, 1996. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Draft.

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

Poesio, Massimo, Florence Bruneseaux, Sarah Davies, and Laurent Romary, 1999. The MATE meta-scheme for coreference in dialogue in multiple languages. In Marylin Walker (ed.), *Proceedings of the workhops on "Towards Standards and Tools for Discourse Tagging" at the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*. University of Maryland.

Poesio, Massimo and Renata Vieira, 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Prince, Ellen, 1981. Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical Pragmatics*. Academic Press, pages 223–256.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartik, 1985. *A comprehensive grammar of the English language*. London: Longman.

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