

An Empirical Study across Multiple Layers of Annotation

Julia Ritz, Christian Chiarcos, Heike Bieler
SFB 632 "Information Structure", University of Potsdam, Karl-Liebknecht-Str. 24-25
14476 Potsdam, Germany.

{jritz, chiarcos, bieler}@uni-potsdam.de

1 Background

Information Structure manifests itself on various levels of grammar, including phonology, syntax and semantics (Krifka 2007), where it often accounts for the "packaging" of sentences (Chafe 1976), i.e., the same abstract meaning can be realized in different ways:

- (1) a. Es liegen zur Zeit etwa 4.500 Bewerbungen vor. (TüBa-D/Z corpus, s19771)
b. Zur Zeit liegen etwa 4.500 Bewerbungen vor.
c. Etwa 4.500 Bewerbungen liegen zur Zeit vor.

It is obvious that the *semantic* content of an expletive sentence like (1a) can be conveyed more efficiently, e.g. in (1b/c). In this empirical study, we investigate possible *pragmatic* motivations for violating the Gricean principle of quantity.

For the example of German expletives, this study addresses the interrelations between **constituent order/realisation** and **coreference/information structure**. We show how corpus linguistic techniques can be applied to the study of information structural phenomena.*

* We take a strong focus on methodological issues here. How our findings relate to existing theoretical models is left for subsequent research.

2 The Multi-Layer Corpus TüBa-D/Z

Information structure is manifested by many different phenomena, so its empirical study requires corpora annotated on multiple layers. The TüBa-D/Z corpus, for example, combines morphosyntactic, morphological, and syntactic annotations with anaphoric annotations

The TüBa-D/Z corpus (<http://www.sfs.uni-tuebingen.de/tuebadz.shtml>) consists of 2,213 articles from the German newspaper *die tageszeitung (taz)*, 45,200 sentences and 794,079 tokens in total, completely annotated for syntax (Telljohann et al. 2009) and coreference (Naumann 2007). The corpus comprises 101 sentences with expletive *es* in Vorfeld. Their information structural status is assessed here by means of the coreference annotation.

Creation

Anaphoric and syntactic annotations are fundamentally different, so that different specialized tools are required for their creation, in this case Palinka (Orasan 2003) for anaphoric annotations and Annotate (Brants and Plaehn 2000) for morphological and syntactic annotations. TüBa-D/Z thus represents a prototypical multi-layer corpus.**

** By multi-layer corpora, we specifically mean corpora whose creation requires the application of several specialized annotation tools.

3 Working with multi-layer corpora

Physical Integration

Linguistic research with multi-layer corpora requires that the independent layers are transformed onto a common level of representation, i.e., a data model and a format that can represent both pointing relations and syntactic dominance hierarchies.

For this task, we suggest the application of PAULA XML (Dipper 2005), an XML-standoff format, whose data model also forms the basis of a relational data base implementation, ANNIS (Chiarcos et al. 2009).

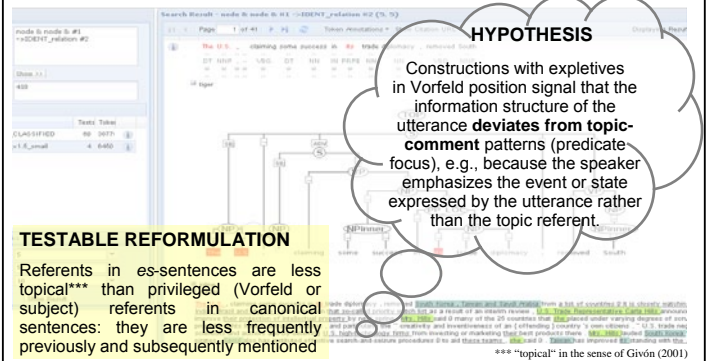
Querying, Visualization and Retrieval of Results

Based on PAULA data structures, ANNIS provides generic means of visualization capable to represent flat, layer-based annotations, dominance trees and pointing relations in separate views. The same structural differentiation is underlying the definition of operators for markable extension, dominance, and pointing relations in the ANNIS query language.

Results can be exported as a table of matches that can be further processed using tools like WEKA (Witten and Frank 2005) or R (Venables and Smith 2002).

4 Evaluating TüBa-D/Z with ANNIS and WEKA

With the help of PAULA and ANNIS-QL, we can now evaluate our research query with the linguistic annotations contained in TüBa-D/Z.



HYPOTHESIS

Constructions with expletives in Vorfeld position signal that the information structure of the utterance deviates from topic-comment patterns (predicate focus), e.g., because the speaker emphasizes the event or state expressed by the utterance rather than the topic referent.

TESTABLE REFORMULATION

Referents in *es*-sentences are less topical*** than privileged (Vorfeld or subject) referents in canonical sentences: they are less frequently previously and subsequently mentioned

*** "topical" in the sense of Givón (2001)

5 Selected Statistical Experiments (exp. 1, exp.2)

Tested sentences with VF constituents and subject in MF position

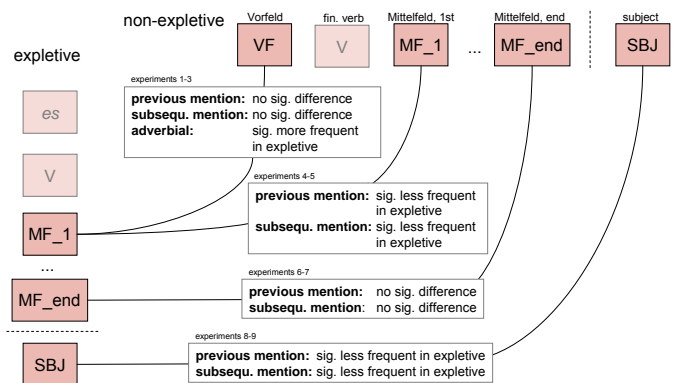
Table 1 shows the number of previously mentioned (discourse-old) subjects in sentences with an expletive in VF position as compared to other sentences. It is significantly lower in expletive constructions ($\chi^2=30.11$, $p<.0005$).

mentioned	ES	other VF	total
subj prev.	2	3325	3327
subj not	89	8195	8284
	91	11520	11611

Table 2 shows the number of subsequently mentioned subjects in sentences with an expletive in VF position as compared to other sentences. It is also significantly lower in expletive constructions ($\chi^2=7.62$, $p<.01$).

mentioned	ES	other VF	total
subj subseq.	12	3059	3071
subj not	79	8461	8540
	91	11520	11611

6 Findings



7 Preliminary Conclusions

- ANNIS and WEKA can be applied to investigate linguistic research questions that involve structurally different annotations
- The topicality*** of the first postverbal constituent (MF_1) in expletive sentences is comparable to Vorfeld constituents (VF) in non-expletives (exp. 1-2; cf. exp. 4-5), but this does not necessarily reflect identical grammatical roles (exp. 3).
- Subjects (SBJ) in expletive sentences are less topical*** than subjects in non-expletive sentences (exp. 8-9)

*** "topical" in the sense of Givón (2001)

Brants, T. and Plaehn, O. (2000). Interactive corpus annotation. *Proc. LREC 2000*. Athens, Greece, June 2000.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. Li, *Subject and topic*. Academic Press, New York, 25-55.

Chiarcos, C., S. Dipper, M. Götz, U. Leser, A. Lüdeling, J. Ritz, M. Stede (2009). A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. *TAL (Traitement automatique des langues)*, 49(2).

Dipper S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tag 2005 (BXML 2005)*, pages 39-50, Berlin, Germany.

Givón, T. (2001). *Syntax*, 2 volumes. John Benjamins, Amsterdam, Philadelphia.

Krifka, M. (2007). Basic notions of information structure. In Féry, C. et al., *The notions of information structure*. Interdisciplinary Studies on Information Structure (ISIS); 6, Universitätsverlag Potsdam.

Naumann, K. (2007). *Manual for the annotation of in-document referential relations*. version of May 2007, technical report, Universität Tübingen, <http://www.sfs.uni-tuebingen.de/resources/tuebadz-coreference-manual-1-2007.pdf>

Orasan, C. (2003). PALinka: A highly customisable tool for discourse annotation. *Proc. of the 4th SIGDial Workshop on Discourse and Dialogue*, p. 39-43

Telljohann, H., E. Hinrichs, S. Kübler, H. Zinsmeister, K. Beck (2009). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*, version of November 2009, technical report, Universität Tübingen, <http://www.sfs.uni-tuebingen.de/resources/tuebadz-sty-2009.pdf>

Venables, W., D. Smith (2002). *An introduction to R. Network Theory*. Bristol, UK.

Witten, I., E. Frank (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufman, San Francisco