Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Annotating descriptively incomplete language phenomena

Fabian Barteld, Sarah Ihden, Ingrid Schröder, and Heike Zinsmeister

Historical language corpora of German and language examples

Existing corpus projects for historical languages give examples for ambiguities concerning part of speech and each gives sets of rules to disambiguate the different analyses. This leads to a situation where the ambiguity cannot be recovered from the annotation.

▸ MERCURIUS
(Early New High German)

(1) zu **ruck** ge-brach-t
to back PTCP-bring-PTCP
KOMPE PTKVZ VVPP
'brought back'
(Pauly et al., 2012, p. 79, fig. 7)

(2) zu **grund** gang-en
to ground-M.DAT.SG go-PTCP
APPR NN VVPP
'gone down'
(Pauly et al., 2012, p. 79, fig. 7)

▸ Reference Corpora
(Old German, Middle High German, Early New High German)

(3) si zoch in **wider**
she-3SG.NOM tug-3SG.PST.IND he-3SG.ACC back
PPER VVFIN PPER PTKVZ
'she pulled him back'
(Dipper et al., 2013, 7, ex. 10a, MHG)

(4) si hie-ten **bider**-zog-en [...] ier hant
they-3PL.NOM have-3PL.PST.SBJV back-tug-PTCP [...] their-ACC.SG hand-F.ACC.SG
PPER VAFIN VVPP [...] DPOSA NA
'they would have withdrawn their hand'
(Dipper et al., 2013, 7, ex. 10b, MHG)

Ambiguous structures are not only found at the level of part of speech but also at the level of inflection. The examples are from the project 'Reference Corpus Middle Low German/ Low Rhenish', in which Middle Low German (GML) and Low Rhenish texts are transcribed and annotated (Schröder, In print).

In GML there are many identical forms of personal pronouns used for several morphological feature values. The word *en*, for example, can be either masculine or neuter and it can even be plural (where there is syncretism between the three genders). If *en* is plural or neuter, it can only be a dative form, but if it is masculine, it could be either dative or accusative.

		Masc	Neut	Fem
Sg	Nom	hê, hî	it, et	sê, sî, sů
	Gen	is, es, sîn, sîner	is, es	ere, er erer, örer
	Dat	***en***, eme, öme	***en***, em, eme, öm, öme	
	Acc	***en***, ene, ön, öne	it, et	sê, sî, sů
Pl	Nom	sê, sî		
	Gen	ere, er, erer, örer		
	Dat	***en***, em, öm, jüm		
	Acc	sê, sî		

Table 1 : GML pronouns - 3rd person; freely based on Lasch (1974)

(5) vppe dat god-es sone
upon that god-M.GEN.PL son-M.NOM.SG
ge-ere-t werd-e dor en
PTCP-honour-PTCP will-3SG.PRS.SBJV through EN
'so that god's son would be honoured through EN'
(BuxtehEv, Joh 11,4, GML)

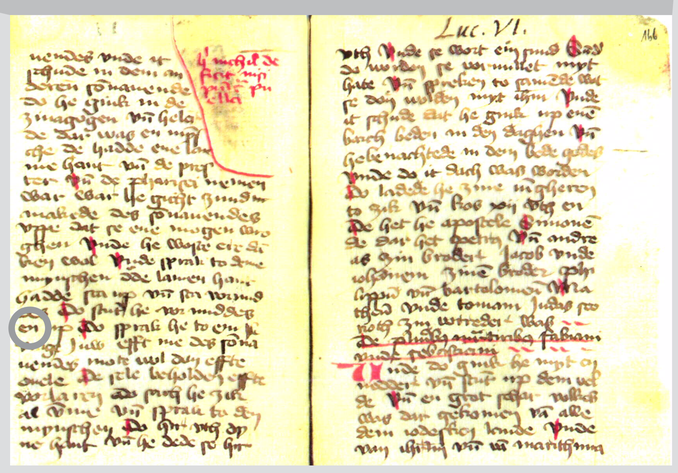
In example (5), the antecedent of *en* provides information on gender and number, but the ambiguity with respect to case can only be resolved in a local context – here, the prepositional phrase *dor en* ('through EN'). Due to the fact that in GML the preposition *dor* governs different cases, the case ambiguity in sentence (5) cannot be resolved.

A classification of descriptively incomplete language phenomena

Type of phenomenon	True analysis	Annotator	Token
Uncertainty	Dat	Dat?Acc?	*en*
Underspecification	Obj	Dat?Acc?	*en*
Ambiguity	{Dat,Acc}	Dat?Acc?	*en*

Table 2 : Types of descriptively incomplete language phenomena

Dat?Acc?



(i) Uncertainty
Incomplete information due to infrequent occurrence in the training material (automatic annotation), incomplete treatment in annotation guidelines or an incomplete understanding of the language system (manual annotation).
(ii) Underspecification
Incomplete information due to an undistinguished feature of the language system.
(iii) Ambiguity
Incomplete information due to an ambiguity in the language data.

These types of incomplete information "should be distinguishable by different mark-up" (EAGLES, 1996). But as all three types result in the same problem for the annotator, it is not always possible to decide at the time of annotation whether there is an ambiguity, an underspecification, or an uncertainty. Often, this can only be resolved (if at all) after the annotation has been completed and the quantitative results based on the annotated data have become available. Therefore, during the annotation process, the annotator should be able to assign any number of annotations to every possible feature.

Formats for encoding multiple annotations in XML markup

This section presents three encodings for multiple annotations of the ambiguous GML pronoun *en* 'him/ it' introduced in example 5.

```
<t f='en' i=''>
      <P t='PPER' r='1' c='1'>
            <j n='annotatorA' c='1'/>
            <j n='annotatorB' c='1'/>
            <b f='hî'>
                  <j n='annotatorA' c='0.5'/>
                  <j n='annotatorB' c='0.5'/>
                  <m d='3.Sg.Masc.Acc'/>
                  <m d='3.Sg.Masc.Dat'/>
            </b>
            <b f='it'>
                  <j n='annotatorA' c='0.5'/>
                  <j n='annotatorB' c='0.5'/>
                  <m d='3.Sg.Neut.Dat'/>
            </b>
      </P>
</t>
```

Figure 1 : A weighted set of alternatives (TüPP-D/Z DTD).

Figure 1 presents TüPP-D/Z DTD (Ule, 2004), an *inline-XML* specification that was designed to represent a ranked list of multiple competing tagger outputs resulting from ensemble tagging. Using such a structure of weighted alternatives, all possible interpretations could be encoded and made available for further analysis.

The other two options are encoded in *generic XML-standoff* formats that represent annotations as directed acyclic graphs such as PAULA (Dipper, 2005; Chiarcos et al., 2008) and GrAF (Ide and Suderman, 2007), both specifications of different versions of the Linguistic Annotation Framework (LAF) (Ide and Romary, 2004).

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE paula SYSTEM "paula_mark.dtd">
<paula version="1.1">
<header paula_id="mycorpus.doc1_morph_seg"/>
<markList
xmlns:xlink="http://www.w3.org/1999/xlink"
type="morph"
xml:base="mycorpus.doc1.tok.xml">
<!-- en -->
<mark id="morph_8"
xlink:href="(#xpointer(id('tok_8')))"/>
<!-- en -->
<mark id="morph_9"
xlink:href="(#xpointer(id('tok_8')))"/>
<!-- en -->
<mark id="morph_10"
xlink:href="(#xpointer(id('tok_8')))"/>
</markList>
</paula>
```

```
...
<multiFeat xlink:href="#morph_8"> <!-- en -->
      <feat name="Person" value="3"/>
      <feat name="Number" value="Sg"/>
      <feat name="Gender" value="Masc"/>
      <feat name="Case" value="Acc"/>
</multiFeat>
<multiFeat xlink:href="#morph_9"> <!-- en -->
      <feat name="Person" value="3"/>
      <feat name="Number" value="Sg"/>
      <feat name="Gender" value="Masc"/>
      <feat name="Case" value="Dat"/>
</multiFeat>
<multiFeat xlink:href="#morph_10"> <!-- en -->
      <feat name="Person" value="3"/>
      <feat name="Number" value="Sg"/>
      <feat name="Gender" value="Neu"/>
      <feat name="Case" value="Dat"/>
</multiFeat>
...
```

Figure 2 : A separate layer with multiple markables linked to the same token (PAULA).

Figure 2 shows that features are related to annotation objects ("markables") only by links. Markables themselves are linked to text tokens. Multiple markables (i.e. multiple annotations) can be related to the same token, as each markable is uniquely identified by its ID.

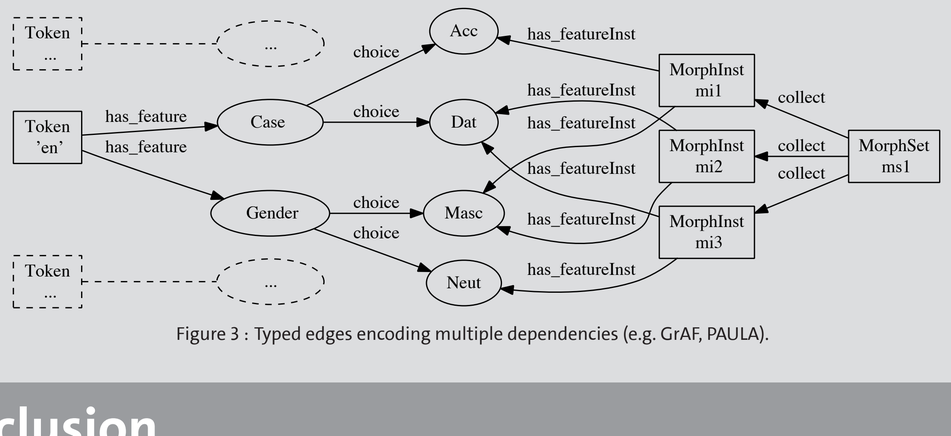


Figure 3 : Typed edges encoding multiple dependencies (e.g. GrAF, PAULA).

Figure 3, finally, sketches how dependencies between multiple ambiguous features could be encoded in the generic LAF-derived standoff formats by annotating edges. We combine a choice structure with a collect structure; thereby explicitly encoding the dependencies between the multiple ambiguous features of gender and case. Kountz et al. (2008) propose an extension to GrAF in which such dependencies are explicitly modeled.

Conclusion

In order to avoid circular argumentation and to reveal the actual grammatical characteristics of the language under investigation, historical corpus linguistics must go beyond simply adapting the rules of a standardized language, both by disambiguating ambiguous forms but also by encoding ambiguities. In historical language corpora such as "ReN", annotators must deal with descriptively incomplete language phenomena.

Furthermore, they need to decide what type of phenomena these are, i.e., real ambiguities, underspecifications or uncertainties. Often this decision is impossible at the time of the annotation, since all three types result in the same problem for the annotator.

In markup formats such PAULA or GrAF, the straightforward encoding of multiple annotations and their dependencies is possible. Nevertheless, linguists still lack sufficient tools to create, query, and visualize the multiple annotations represented in the underlying data structure. For these reasons, corpus projects such as "ReN" are currently unable to use multiple annotations, even though this is the most appropriate encoding strategy for the grammatical annotation of historical languages.

References

Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*, 49(2):271–293.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. HiTS: ein Tagset für historische Sprachstufen des Deutschen. [HiTS: A tagset for historical varieties of German]. *JLCL*, 28(1):1–53.

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

EAGLES. 1996. Recommendations for the morphosyntactic annotation of corpora. EAGLES document EAG-TCWG-MAC/R. Technical report.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, 10(3-4):211–225.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotation. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Manuel Kountz, Ulrich Heid, and Kerstin Eckart. 2008. A LAF/GrAF-based encoding scheme for underspecified representations of dependency structures. In *Proceedings of LREC-2008, Linguistic Resources and Evaluation Conference*, Marrakesh.

Agathe Lasch. 1974. *Mittelniederdeutsche Grammatik. [Middle Low German Grammar]*. Sammlung kurzer Grammatiken germanischer Dialekte. A. Hauptreihe, 9. Niemeyer, Tübingen, 2nd edition.

Dennis Pauly, Ulyana Senyuk, and Ulrike Demske. 2012. Strukturelle Mehrdeutigkeit in frühneuhochdeutschen Texten. [Structural ambiguities in Early New High German texts]. *JLCL*, 27(2):65–82.

Sabine Pettke and Ingrid Schröder. 1992. Eine Buxtehuder Evangelienhandschrift. Die vier Evangelien in einer mittelniederdeutschen Übersetzung des 15. Jahrhunderts aus dem Alten Kloster. [A Buxtehude handwritten gospel. A GML translation of the four gospels from the fifteenth century]. In Bernd Utermöhlen, editor, *Qvatuor Evangeliorum versio Saxonica. Eine mittelniederdeutsche Evangelienhandschrift aus dem 15. Jahrhundert. Textedition*, Buxtehuder Notizen Nr. 5, pages 99–266. Stadt Buxtehude, Buxtehude.

Ingrid Schröder. In print. Das Referenzkorpus: Neue Perspektiven für die mittelniederdeutsche Grammatikographie. [The reference corpus: New perspectives for GML grammatography]. In *Jahrbuch für germanistische Sprachgeschichte 5*. De Gruyter.

Tylman Ule. 2004. Markup manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z). Technical report, University of Tübingen.

Sources of Attested Examples

BuxtehEv Qvator Evangeliorum versio Saxonica. A GML handwritten gospel from the fifteenth century. Transliterated by the DFG-funded project "ReN". For further information, see Pettke and Schröder (1992).

The image shows Thott 8, 8°, Det Kongelige Bibliotek, Copenhagen (Pettke and Schröder, 1992, 54sq.).

All language abbreviations used on this poster correspond to ISO 639.
The paper for this poster can be found at
http://www.aclweb.org/anthology/W/W14/W14-4915.pdf