Annotating descriptively incomplete language phenomena

Fabian Barteld, **Sarah Ihden**, **Ingrid Schröder**, and **Heike Zinsmeister**
Institut für Germanistik
Universität Hamburg
Von-Melle-Park 6
20146 Hamburg, Germany
{ fabian.barteld, sarah.ihden, ingrid.schroeder, heike.zinsmeister }
@uni-hamburg.de

Abstract

When annotating non-standard languages, descriptively incomplete language phenomena (EAGLES, 1996) are often encountered. In this paper, we present examples of ambiguous forms taken from a historical corpus and offer a classification of such descriptively incomplete language phenomena and its rationale. We then discuss various approaches to the annotation of these phenomena, arguing that multiple annotations provide the most appropriate encoding strategy for the annotator. Finally, we show how multiple annotations can be encoded in existing standards such as PAULA and GrAF.

1 Introduction

In grammatical annotations, a lack of ambiguity is of great benefit: The more distinctive the relationship between a token and its morphological and syntactic attributes, the more successful and reliable the annotation. However, especially in corpora of non-standard language varieties annotators are confronted with a significant number of cases of doubt and ambiguity. This problem has been more relevant in semantic and syntactic analyses than in PoS tagging and morphological annotation, and consequently has already been addressed in the former processes (Kountz et al., 2008; Bunt, 2007; Spranger and Kountz, 2007; Regneri et al., 2008) and incorporated into tools such as SALTO (Burchardt et al., 2006). With respect to corpora of non-standard languages, ambiguous forms must be taken into consideration in morphosyntactic tagging as well. This has been confirmed by current corpus projects of historical varieties of German – for example, the "MERCURIUS Corpus of Early New High German" (ENHG[1]) (Pauly et al., 2012) and the "Historical Tagset" (HiTS) (Dipper et al., 2013), which provide different options for dealing with ambiguities at the level of part of speech. Below we will discuss examples of ambiguities at the morphological level.

Within the extensive field of non-standard language annotations, we have concentrated on historical linguistics, showcasing the kinds of ambiguities that historical corpus linguists must confront and how they can be managed. Historical corpus linguistics based on annotation necessarily faces the challenge of avoiding circular argumentation. The description of a historic language must be based on the annotated texts of the corpus, since they are the only sources of linguistic material in historical grammatography. However, no annotation of the material can be accomplished without a basic knowledge of the language and its structure. Thus, an annotator confronted with a dubious case cannot know whether it is actually a case of ambiguity in the language system or whether the grammatical categories adopted for the annotation do not fit the grammatical system of the non-standard language. Transferring the annotation standards developed for a standardized language such as written New High German (NHG) to a historical corpus might at first seem tempting, but this process would conceal the actual grammatical characteristics of the language to be described.

[1] All language abbreviations in this article correspond to ISO 639.

LAW VIII - The 8th Linguistic Annotation Workshop, pages 99–104,
Dublin, Ireland, August 23-24 2014.

		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)

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

2 Cases of descriptively incomplete phenomena

The project "Reference Corpus Middle Low German/ Low Rhenish (1200–1650)"[2] transliterates and grammatically annotates the Middle Low German (GML) texts from which we take our examples. Because GML is a non-standardized language that is not well described, ambiguous forms occur frequently, and accurately interpreting them is a matter of high priority for any annotation. First, with regard to nouns and pronouns, GML's case syncretism[3] should be mentioned. For personal pronouns, in particular the syncretism of the dative and accusative forms in the first- and second-person singular and plural leads to problems in annotation. However, in this section, we concentrate on the third person.

Table 1 illustrates the many identical forms of third person personal pronouns that are used for several morphological feature values. Moreover, it reveals the distribution of case syncretism across the three different genders of the third-person singular.[4] While the neuter paradigm shows syncretism in the nominative and accusative forms, for the feminine pronouns there are ambiguous forms not only for nominative and accusative but also for genitive and dative. The masculine paradigm includes a partial syncretism of dative and accusative for the pronoun *en* ('him').

In addition, there is syncretism in the dative forms of the third-person singular masculine and neuter and in the third-person plural. Hence, in example (1),[5] the word *en* could be either masculine or neuter if there is no context providing reliable information on the gender of the referent, or it could 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.

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

Even where the context provides additional information, often not all ambiguities can be resolved. In example (1), the antecedent of *en* provides information on gender (masculine) and number (singular), but the ambiguity with respect to case can only be resolved in a local context – here, the prepositional phrase. The problem is that in GML the preposition *dor* ('through') can govern different cases. Consequently, the case ambiguity in (1) cannot be resolved.

There are many other examples of ambiguous forms, for instance, the gender of nouns or the inflection paradigm of verbs. For all these cases of ambiguity the annotation should provide as much grammatical information on a given form as possible.

[2]The "Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200–1650)" ("Reference Corpus Middle Low German/ Low Rhenish", or "ReN"), supported by the German Research Foundation (DFG)) and in development since February/ March 2013 at the universities of Hamburg and Münster, is part of the "Corpus of Historical German Texts", together with the corpora "Altdeutsch" (Old German), "Mittelhochdeutsch" (Middle High German), and "Frühneuhochdeutsch" (Early New High German). More information on the structure of ReN can be found in Nagel and Peters (In print) and on the website `www.referenzkorpus-mnd-nrh.de`. For information on the annotation used in ReN and possible grammatical analyses, see Schröder (In print).

[3]Baerman (2006) asserts that "syncretism refers to the situation when a single inflectional form corresponds to multiple morphosyntactic feature values" (363). With respect to the feature case, this means that identical forms are used for different cases, e.g., for dative and accusative.

[4]The order of the pronouns was chosen for presentational reasons. The example *en* that we refer to in this paper is shown in bold italics.

[5]This glossing is based on the Leipzig Glossing Rules (`http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf`).

3 Types of descriptively incomplete language phenomena

In cases of descriptively incomplete language phenomena such as those described above, the annotator (which could be a tool or a human) is unable to unambiguously assign an analysis to the language data. This inability can have various causes. Consequently, EAGLES (1996) distinguishes between two types of "descriptively incomplete phenomena": *underspecification* and *ambiguity*. In the first case, the inability arises because "the distinction between the different values of an attribute is not relevant". The second case is characterized as "the phenomenon of lack of information, where there is uncertainty between two or more alternative descriptions". For both of these types, EAGLES provides subtypes; however, in the case of ambiguity, these subtypes also differ with respect to the reason for the uncertainty. In one subtype, the apparent ambiguity could be resolved given more information. In the other, the uncertainty results from a real ambiguity in the language or the given text and therefore cannot be resolved. Consequently, we propose a differentiation between three types of descriptively incomplete language phenomena that can occur during annotation: (i) **uncertainty**, i.e., 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**, i.e., incomplete information due to an undistinguished feature of the language system; and (iii) **ambiguity**, i.e., incomplete information due to an ambiguity in the language data.

Returning to example (1), further analyses could provide evidence that the preposition *dor* ('through') unambiguously takes the accusative case, such that this would represent a case of *uncertainty*. In English personal pronouns, there is no distinction made between dative and accusative, both of which are represented by the objective case (Obj) (Quirk et al., 1985). If this were also true for GML, the example would be a case of *underspecification*. However, it could also represent a true case of *ambiguity*. As long as this categorization is unclear, the types cannot be distinguished.

Table 2 summarizes the distinction between these three types. Although all of them result in the same situation for the annotator (machine or human), they differ with respect to the *true analysis*, which is unknown to the annotator; it is therefore impossible for him or her to definitively assign a tag to the token, as exemplified in Table 2. In situations of *uncertainty* or *underspecification*, an unambiguous, true analysis exists. In the case of uncertainty, it is a matter of redefining the annotation guidelines to help the annotating system to find this true analysis. In the case of underspecification, the tagset is too fine-grained to provide the true analysis. Only by adjusting the tagset would the annotator be able to determine the true analysis. Adjustments to the annotation guidelines and the tagset during the process of annotation can be accomplished through the use of an annotation development cycle such as the MATTER methodology (Pustejovsky and Stubbs, 2012, 23–32). In the case of *ambiguity*, however, both analyses are true. They should be retrievable for further interpretation and thus should both be assigned to the token.

Optimally, the different types of incomplete information "should be distinguishable by different markup" (EAGLES, 1996). But as we have argued, when annotating historical languages (or less-studied languages in general), it is not always possible to decide at the time of annotation whether there is an ambiguity, an underspecification, or an uncertainty, as all three result in the same problem for the annotator. Thus, in many cases, the annotator can only distinguish between the three types (if at all) after the annotation has been completed and the quantitative results based on the annotated data have become available. The three types must therefore be dealt with similarly during the annotation process, and the possible interpretations should be retrievable from the annotations. Consequently, the annotator should have the possibility to assign any number of annotations to every possible feature. This would require special tools to create and retrieve these annotations, but existing standards to encode annotations are already flexible enough to allow annotations. Some examples are shown in the next section.

4 Encoding multiple annotations in markup standards

This section presents three formats for encoding multiple annotations of descriptively incomplete structures in XML markup. We return to the ambiguous GML pronoun *en* 'him/ it' introduced in example (1) in Section 2.

Our first option is 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 the same kind of structure, all possible interpretations of the pronoun *en* could be encoded and made available for further analysis and disambiguation.

The other two options are generic *XML-standoff* formats that represent annotations as directed acyclic graphs: PAULA (Dipper, 2005; Chiarcos et al., 2008), derived from early drafts of the Linguistic Annotation Framework (LAF) (Ide and Romary, 2004), and GrAF (Ide and Suderman, 2007), a more recent specification of the LAF. Each level of annotation is represented separately, such that features are related to annotation objects ("markables") only by links. Markables themselves are defined on the basis of text tokens or other markables. Multiple markables can be related to the same token, as each markable is uniquely identified by its ID. These options also allow us to encode all interpretations of *en*.[6]

In certain cases, there are dependencies between multiple ambiguous features. Concerning 'en', if the gender is *Neut*, the case is not ambiguous, but if the gender is *Masc*, the case could be either *Dat* or *Acc* (cf. Table 1). The above strategies do not allow us to encode these dependencies. However, the generic LAF-derived standoff formats can be employed to do this because they also allow us to define labels for edges, such that they can be annotated and typed. Kountz et al. (2008) propose an extension to GrAF in which such dependencies are explicitly modeled. As depicted in Figure 1, we make use of this property to combine a choice structure with a collect structure. In this way, each token correlates with one MorphSet object that can be instantiated by a set of MorphInst objects, thereby explicitly encoding the dependencies between the multiple ambiguous features of gender and case.

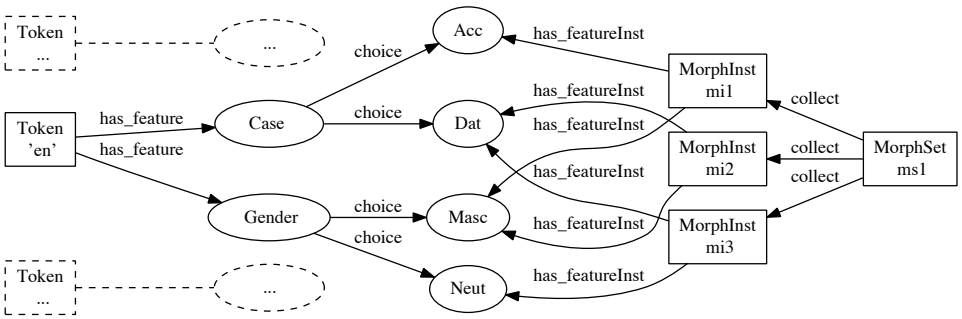

Figure 1: Representation of an encoding of the ambiguous GML pronoun *en* 'him/it' with typed edges

5 Conclusion and Outlook

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. By means of data taken from the "ReN" corpus, we have demonstrated that in historical language corpora, 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, as discussed in Section 3. In Section 4, we have shown that 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.

[6] In addition, PAULA offers a *multiFeat* structure (Zeldes et al., 2013, 14f.) for linking sets of fully-specified features to one markable. However, each piece of information must be unambiguous.

Acknowledgements

We would like to thank Kerstin Eckart for very helpful discussion and suggestions, and also Claire Bacher for improving our English. All remaining errors are ours. Figure 1 was created with GraphvizFiddle (`https://stamm-wilbrandt.de/GraphvizFiddle/`) an online editor for Graphviz (`http://www.graphviz.org/`). Part of this work was funded by the German Research Foundation (DFG).

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).

References

Matthew Baerman. 2006. Syncretism. In Keith Brown, editor, *Encyclopedia of Language and Linguistics.*, volume 12, pages 363–366. Elsevier, Amsterdam [a.o.], 2nd edition.

Harry Bunt. 2007. Semantic underspecification: Which technique for what purpose? In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning*, volume 83 of *Studies in Linguistics and Philosophy*, pages 55–85. Springer.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Pado. 2006. SALTO: A versatile multi-level annotation tool. In *Proceedings of LREC 2006*, pages 517–520.

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.

Norbert Nagel and Robert Peters. In print. Das digitale "Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (ReN)". [The digital Reference Corpus of Middle Low German/ Low Rhenish (ReN)]. In *Jahrbuch für germanistische Sprachgeschichte 5*. De Gruyter.

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.

James Pustejovsky and Amber Stubbs. 2012. *Natural language annotation for machine learning*. O'Reilly, Beijing [a.o.].

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Michaela Regneri, Markus Egg, and Alexander Koller. 2008. Efficient processing of underspecified discourse representations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 245–248. Association for Computational Linguistics.

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.

Kristina Spranger and Manuel Kountz. 2007. Efficient ambiguity-handling using underspecified representations. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications*. Gunter Narr Verlag, Tübingen.

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

Amir Zeldes, Florian Zipser, and Arne Neumann. 2013. PAULA XML Documentation: Format version 1.1. Technical Report Version: P1.1.2013.1.21a, Humboldt-Universität zu Berlin, Berlin.