Annotation Challenges for Reconstructing the Structural Elaboration of Middle Low German

Nina Seemann
Michaela Geierhos
Paderborn University
Heinz Nixdorf Institute
seemann@hni.upb.de
geierhos@hni.upb.de

Marie-Luis Merten
Doris Tophinke
Paderborn University
Department of German Linguistics and Comparative Literature
mlmerten@mail.upb.de
doris.tophinke@upb.de

Eyke Hüllermeier
Paderborn University
Department of Computer Science
eyke@upb.de

Abstract

In this paper, we present the annotation challenges we have encountered when working on a historical language that was undergoing elaboration processes. We especially focus on syntactic ambiguity and gradience in Middle Low German, which causes uncertainty to some extent. Since current annotation tools consider construction contexts and the dynamics of the grammaticalization only partially, we plan to extend CorA – a web-based annotation tool for historical and other non-standard language data – to capture elaboration phenomena and annotator unsureness. Moreover, we seek to interactively learn morphological as well as syntactic annotations.

1 Tracing Elaboration Processes

Language elaboration is a continuous development. According to Traugott and Trousdale (2013), it involves processes that change existing constructions with respect to formal or semantic aspects *(constructional change)* and processes in which new constructions emerge *(constructionalization)*. To distinguish between these processes and to be able to describe their dynamics, one needs a sophisticated inventory of descriptive categories. Those categories should capture formal and/or semantic micro-changes as well as ambiguity and vagueness. Assuming that constructions are "holistically" determined by interacting formal and semantic/functional characteristics, a single characteristic is always judged by its construction binding, i.e. in cooccurrence with other characteristics. Capturing (diachronic) semantic changes of constructions that are formally (yet) hardly visible or – vice versa – to recognize that formal characteristics have not (yet) the grammatical function that they will acquire in later language stages is only possible in a given context.

1.1 Related Annotation Work and Tools

Over the last years, historical texts gained a lot of interest from both computational and corpus linguistics. Due to the graphematic and grammatical variability, those texts are interesting and analytically challenging (Dipper et al., 2013b; Bollmann et al., 2014). Accordingly, many works emerged that deal with the annotation of historical texts. For German, Dipper et al. (2013a) introduced a tag set for historical language levels (called HiTS). Multiple projects developed different reference corpora for historical language levels that were annotated with or in analogy with HiTS. Recently, a reference corpus Middle Low German/Low Rhenish (1200–1650) was compiled in a collaboration between the German universities of Hamburg and Münster. Additionally, annotation tools ANNIS (Zeldes et al., 2009) and CorA (Bollmann et al., 2014) were introduced. In the context of literature studies, the heureCLÉA project used CATMA (Meister et al., 2017) to annotate their data in correspondence with strict annotation guidelines that cover aspects of uncertainty/ambiguity (Gius and Jacke, 2016).

1.2 Problem Definition

However, current annotation tools consider construction contexts and the dynamics of the grammaticalization only partially. Although these problems are well known (Dipper et al., 2013a; Bollmann et al., 2014), satisfying solutions are

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 40–45, Vancouver, BC, August 4, 2017.

still missing. For our purposes, an annotation tool should provide the following features/labels:

- *Part-of-speech (POS) tag ambiguity*: If a human annotator cannot clearly indicate which POS tag can be assigned on morphological or syntactic level, there must be an option to annotate which one is more likely. However, the annotator should not be forced to disambiguate – several (prioritized) interpretations should be allowed.
- *Syntactic gradience* (Aarts, 2007): Even in case of syntactic constructions, the above-mentioned grammatical indeterminacy will occur for construction tags.
- *Individual annotation order:* Instead of using the common text analysis pipeline with a strict order, where first POS tags and then constructions tags are assigned, our tool supports the cognitive annotation process. In some contexts, it is easier for annotators to start with (local) syntactic segmentation before POS tags can be disambiguated within its corresponding construction borders.
- *Annotator support*: We provide auto-suggestions (i.e. POS and/or construction tags and their corresponding uncertainty values stating a certain annotator unsureness) for unlabeled corpus data based on an interactive machine learning approach.
- *Annotator unsureness*: We need to express if human annotators are (un)sure about their annotation decision. In case of unsureness, we need additional comments which explain their incapability to make a clear decision.

We settled on the aforementioned features within our interdisciplinary group, especially valuing the expertise in Middle Low German from the group's linguists. In order to significantly reconstruct the temporal and spatial dynamics of elaboration processes, it is essential to analyze bigger corpora. Methodologically, it is not possible to do that on the basis of single occurrences. Moreover, on the basis of statistically significant evidence provided by our tool, the difference between the construction change and the emergence of new constructions can be detected much easier. To face the above-mentioned features, we will extend CorA (Bollmann et al., 2014) instead of starting from scratch.

1.3 Scope of Study

Although there are some annotation tools dealing with historical German, they do not perfectly fit our purposes. For one, it is clear that we are dealing with highly ambiguous data (due to the language shift) and this can result in uncertainty. Next, having a large corpus means a lot of manual annotation. As this is very time intensive work, we want to develop a parser that learns over time to suggest possible annotations for each token and even constituents. Finally, as language elaboration involves the change and emergence of constructions, we need syntactic/grammatical annotations for such constructions. None of the features mentioned in Section 1.2 is currently supported by any annotation tool.

The outline of the paper is as follows. In Section 2, we give an overview of our project goals, data, and scope before we show in Section 3 what kind of uncertainties we are facing. Finally, we will draw our conclusions and present some future work in Section 4.

2 Interactive Grammar Analysis

2.1 Project Aims

As already stated above, we investigate the (structural) elaboration of Middle Low German (MLG) from the 13th century to the written language shift (16th/17th century). During this period, MLG lost its dominant position as a supraregional written language accompanied with the growing influence of (written) Early New High German (ENHG). The study makes an important contribution to the reconstruction of grammatical developments in written MLG, which are hitherto examined only to some extent. Our overall aim is to verify the assumption that written ENHG instantiates – despite an ENHG lexis – a MLG syntax/grammar. Particularly, we have four questions that we intend to answer: (1) Which kind of elaboration processes do occur? (2) How far does the elaboration go? (3) How fast does each elaboration process establish itself? (4) Are spatial points of origin identifiable?

2.2 Key Facts about our Corpus

Our empirical base is a corpus that consists of legal texts from the 13th to the 17th century.

2.2.1 Characteristics

There are several reasons why we focus on *urban legal statutes*: We assume that processes of lan-

guage elaboration can be investigated especially in legal texts. They have to construe complex (legal) issues understandable independently of contextual information, so that elaborated linguistic structures capable of such a construal must be developed. These legal issues specifically occur in the form of *conditional relations*; consequently, we are able to examine changes concerning the linguistic construction of conditionality during the investigation period. Furthermore, legal statutes are locatable and datable, with the result that developmental dynamics of elaboration processes can be spatio-temporally reconstructed.

2.2.2 Distribution and Size

The corpus is divided into two parts: *MLG* consists of Middle Low German texts from 1227 to 1650 and covers about 1.2 million tokens. *ENHG* contains a selection of the first ENHG texts (400,000 tokens) arising in the Low German language area after the written language shift.

2.2.3 Text Sources

Another important aspect of our corpus is the use of primary materials for transcriptions and not editions. We want the transcriptions to be as diplomatic as possible, i.e. we keep text structuring elements like numbering, paragraphs, rubrifications, initials, and similar. In a broader sense, changes in the layout of a written text can be seen as elaboration phenomena (Krämer et al., 2012). It is conceivable that structural elements were introduced to separate or highlight textual and/or grammatical units. Keeping this information enables us to examine whether the language shift was accompanied by grammaticalization of interpunction.

2.3 Human-in-the-Loop Annotation Support

It is common practice for annotation tools to provide annotation suggestions (e.g. Cunningham et al. (2011); Stenetorp et al. (2012); Yimam et al. (2014); Bollmann et al. (2014); Bögel et al. (2015)). We also plan to develop an interactive procedure that combines machine learning and expert feedback (Holzinger, 2016) to solve one of the most central problems of existing annotation tools for historical texts. Due to the historical dynamics of grammar, we cannot use existing parsing and tagging system since those require static (a priori defined) syntactic rules and grammatical categories. We want to discover an evolving, dynamic grammar by using rule-based text analysis techniques and machine learning methods (Hüllermeier, 2011). This enables us to reconstruct the language elaboration in an evidence-based way, which is a novelty.

3 Annotation Uncertainties

None of the current annotation tools cover the aspect of *uncertainty* as we do. Of course, this aspect is known and projects have strategies to cover it. In the end, it always comes down to the annotators being forced to decide for exactly one annotation. Instead, our tool allows for multiple annotations with an option to state which one is more likely. A similar idea by Jurgens (2013) allowed weighted multiple word sense annotations, and his results show improvements in the task of word sense disambiguation. In the following, we will explain our morphological (3.1) and syntactic (3.2) uncertainty.

3.1 POS Tag Ambiguity

The sentence in example (1) shows an excerpt for a specific state of the language shift process. The function word group *na deme dat* consists formally of a $preposition_1$ + $reanalyzed\ pronoun_2$ + $primary\ subjunction_3$ but expresses a functional unit $[complex\ subjunction]_4$. This is an obvious interpretation, as the unit *na deme dat* establishes a (temporal) relation between two entities construed as processes. From a cognitive grammar point of view, this functional characteristic is the crucial subjunctional criterion. Besides that, its further grammaticalization – which can be reconstructed based on our texts – suggests such an analysis as complex secondary subjunction (early state of a so called subjunctionalization). In the following stadium, as part of a formal erosion powered by frequency effects, we can observe the reduction of the primary subjunction *dat* and the univerbation of *na* and *dem(e)* as shown in example (2). Furthermore, as a result of desemantication, a causal relation is encoded. In our tool, users can annotate each member of the function word group of example (1) with its respective POS tag as well as assigning a POS tag to the whole group. Additionally, they can state which analysis is more likely.

When annotating historical texts, one could be uncertain which POS tag to assign due to missing context information or other circumstances. This is quite normal, but we are facing uncertainty due to our goal to analyze a language that is undergo-

(1) *We sik erue gůdes vnderwint · oder an sprikt · [**na**$_1$ **deme**$_2$ / **dat**$_3$]$_4$ it*
Who himself hereditary goods takes possession of · or claims · after this / that it
im vordelet is vor gherichte · Dat is en vredebrake
him denied is before court · This is a breach of peace

'Who takes possession of hereditary goods or claims them after it was denied through a court order to do so: This is a breach of peace.' (Goslar, 1350)

(2) [***Nademe***]$_4$ *yt ein groht / und erschrecklyk Laster / unde Sůnde ys den Nahmen des*
After it a great / and terrible vice / and sin is the name of
Allmächtigen Gades tho miszbruken. §.1. So scho+elen vo+erdann dejennigen / de ...
Almighty God to misuse. §.1. So shall henceforth all those / who ...

'Thus, it is a great and terrible vice as well as a sin to misuse the name of the Almighty God. Henceforth, all those who ...' (Dithmarschen, 1567)

ing a shift. In our case, we will have to face issues where it is hard to tell if a token is still a member of category A or already a member of category B. So we should add a degree of how certain an annotation is and give annotators the possibility to exactly state why the annotation is uncertain in this case. This provides a great level of transparency and may lead to new insights.

3.2 Syntactic Gradience

Our aim is the analysis of constructions in the area of language elaboration (Maas, 2010; Tophinke, 2012; Merten, 2015). Therefore, we focus on constructions that model conditional relations of circumstances or have a conditional interpretation. Additionally, we capture all characteristics that cooccur for each construction given the temporal aspect that proved to be typical for this form. Furthermore, we investigate gradient structures and describe constructional changes in their gradual nature in a detailed way. We are interested in changes concerning form and/or meaning/function with respect to the textual perspective: from texts meant to be read out to texts that were designed for reading for oneself (Tophinke, 2009).

In previous work, we already identified some constructions that proved to be relevant to evidently reconstruct the language shift. These syntactic constructions are of varying complexity that can be either called *phrases* (i.e. nominal phrase, prepositional phrase, ...) or *transphrases* (i.e. complex sentences). An interesting constructionalization is the evolution of subjunctional constructions. In our earliest texts, those have the form "[situational context] [specification of situation]" and differ a lot from (literate) subjunctional constructions as we know them today. Although they are formally marked as subjunctional entities (initial subjunction markers like *of, (so) wanne, weret also dat*[1] + verb final position), they are not integrated into a (supposed) matrix structure. This is illustrated in Example (3) through bracketing. From a syntactic perspective, the relation between the syntagma introduced by the subjunction and the following entity is much more loose than we know it from typical subjunctional constructions of present times. But focusing on semantic-functional aspects, one has to emphasize their specific functionality: They are typical structures of so called *space building* (Merten, 2016) which is highly linked to/functionalized with respect to the reception of these (older) texts. As they were meant to be read out, these orate structures ensured – to a certain extent – an easier access for potential listeners (Szczepaniak, 2015). As a result of the ongoing syntactic elaboration processes, the form turned into "[[subjunctional sentence] matrix structure]" where the complete sentence is a subordinative conditional construction that exhibits the nowadays common *if-then* structure. We show an excerpt from our texts in Example (4).

4 Conclusion and Future Work

We showed that there are still missing features in current annotation tools and how we plan to provide them. Our CorA-based annotation tool will be able to handle uncertainties and allow syntactic annotations. Additionally, it will have a feature that provides annotation suggestions to support the human annotator in his/her work.

[1] All three markers translate to *if*. The literal meaning of *weret also dat* is *were it so that = if*.

(3) [*So wanne enen manne ein pant gheset wert.*] [*it si erue that eme ane*
So when a man a pawn given was. it is inheritance that him without
sinen danch wert gheset. ofte ein kisten pant. that scal he up beden to theme
his knowledge was given. Or a mobile pawn. This shall he up weigh to the
nagesten thinghe.]
next thing.

'If a man is offered a pawn and it is - unknowingly to him - an inheritance that was offered or a mobile pawn. This he should upweigh to the next thing. (Stade, 1279)

(4) [[*WAnnehr einer syne Sake dorch Tůgen wahr maken und bewysen wil /] schal*
If one his case through witnesses true make and prove want / shall
he de Tůgen im Rechten nahmkůndig maken]
he the witnesses in law name make (Dithmarschen, 1667)

'If one wants to introduce witnesses to prove his case, he must legally name those witnesses.'

In the future, we plan to also integrate an information retrieval system that allows one to search for certain grammatical/syntactical constructions. As we are interested in temporal-spatial aspects, limiting the search to a specific time span and/or a specific region should be possible. In the long run, it is planned that the tool displays the search results on a (dynamic) map over time.

All tools created in the course of this project will be made available. We will report our progress and news on the project website: http://www.uni-paderborn.de/forschungsprojekte/intergramm/

Acknowledgments

The authors thank the reviewers for their helpful comments. Furthermore, we want to express our gratitude to the *ReN project* for providing us with data and *Marcel Bollmann* for his transfer of CorA. All authors are financially supported by the German Research Foundation (DFG).

References

Bas Aarts. 2007. *Syntactic Gradience. The Nature of Grammatical Indeterminacy*. Oxford University Press, New York.

Tomas Bögel, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris, and Jannik Strötgen. 2015. Collaborative Text Annotation meets Machine Learning: heureCLÉA, a Digital Heuristic Narrative. *DHCommons Journal* 1. http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heureclé-digital-heuristic.

Marcel Bollmann, Florian Petran, Stefanie Dipper, and Julia Krasselt. 2014. CorA: A web-based annotation tool for historical and other non-standard language data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. pages 86–90.

Hamish Cunningham, Diana Maynard, Kalina Botcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrel, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (version 6)*. University of Sheffield Department for Computer Science.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013a. HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics* 28:85–137.

Stefanie Dipper, Anke Lüdeling, and Marc Reznicek. 2013b. NoSta-D: A Corpus of German Non-Standard Varieties. In Marcos Zampieri and Sascha Diwersy, editors, *Non-standard Data Sources in Corpus-based Research*, Shaker Verlag, Aachen, pages 69–76.

Landrecht Dithmarschen. 1667. Print from 1667. https://books.google.de/books?id=t88pAAAAYAAJ&pg=PR6&source=gbs_selected_pages&cad=2#v=onepage&q&f=false. Online; accessed March 24, 2017.

Evelyn Gius and Janina Jacke. 2016. Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets. http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf.

Andreas Holzinger. 2016. Interactive Machine Learning (iML). *Informatik-Spektrum* 39(1):64–68.

Eyke Hüllermeier. 2011. Fuzzy machine learning and data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(4):269–283.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, pages 556–562.

Gustav Korlén. 1950. Das Stader Stadtrecht vom Jahre 1279. In Erik Rooth, editor, *Lunder Germanistische Forschungen*, Håkan Ohlssons Boktryckery, Lund, volume 22, pages 23–117.

Sybille Krämer, Eva Cancik-Kirschbaum, and Rainer Totzke. 2012. *Schriftbildlichkeit*. de Gruyter, Berlin.

Maik Lehmberg. 2013. *Der Goslaer Ratskodex - Das Stadtrecht um 1350: Edition, Übersetzung und begleitende Beiträge*. Verlag für Regionalgeschichte, Bielefeld.

Utz Maas. 2010. Einleitung / Literat und orat. Grundbegriffe der Analyse geschriebener und gesprochener Sprache. In *Grazer linguistische Studien*, Institut für Sprachwissenschaft, Universität Graz, volume 73, pages 5–150.

Jan Christoph Meister, Evelyn Gius, Janina Jacke, Marco Petris, and Malte Meister. 2017. CATMA 5.0. http://catma.de/. Tool homepage; accessed June 07, 2017.

Marie-Luis Merten. 2015. Sprachausbau im Kontext rechtssprachlicher Praktiken des Mittelniederdeutschen. Konstruktionsgrammatik meets Kulturanalyse. In Verein für Niederdeutsche Sprachforschung, editor, *Niederdeutsches Jahrbuch*, Wachholtz Verlag, volume 138, pages 27–51.

Marie-Luis Merten. 2016. *Literater Sprachausbau kognitiv-funktional. Funktionswort-Konstruktionen in der historischen Rechtsschriftlichkeit*. Ph.D. thesis, Paderborn University. Unpublished.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '12, pages 102–107.

Renata Szczepaniak. 2015. Syntaktische Einheitenbildung – typologisch und diachron betrachtet. In Christa Dürscheid and Jan Georg Schneider, editors, *Handbuch Satz, Äußerung, Schema (Handbücher Sprachwissen 4)*, de Gruyter, Berlin, New York, volume 3, pages 104–124.

Doris Tophinke. 2009. Vom Vorlesetext zum Lesetext: Zur Syntax mittelniederdeutscher Rechtsverordnungen im Spätmittelalter. In Angelika Linke and Helmuth Feilke, editors, *Oberfläche und Perfomanz. Untersuchungen zur Sprache als dynamischer Gestalt*, Niemeyer, Tübingen, pages 161–183.

Doris Tophinke. 2012. Syntaktischer Ausbau im Mittelniederdeutschen. Theoretisch-methodische Überlegungen und kursorische Analysen. In Angelika Linke and Helmuth Feilke, editors, *Niederdeutsches Wort*, Niemeyer, Tübingen, volume 52, pages 19–46.

Elizabeth C. Traugott and Graeme Trousdale. 2013. *Constructionalization and constructional changes*. Oxford University Press, Oxford.

Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 91–96.

Amir Zeldes, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. Annis: A search tool for multi-layered annotated corpora. In *Proceedings of Corpus Linguistics*. Liverpool, UK.