The interplay between genre and syntax in a historical Low German corpus

Melissa Farasyn[1], George Walkden[2], Sheila Watts[3], Anne Breitbarth[1]

[1]Ghent University / [2]University of Konstanz / [3]University of Cambridge

In this chapter, we focus on the choice of different genres in the Middle Low German part of the tagged and parsed *Corpus of Historical Low German* and its implications for syntax. We discuss how the inclusion or exclusion of genres has an impact on the study and the discovery of syntactic phenomena in Middle Low German, such as null referential subjects, resumptive pronouns, relative particles and gaps in coordinations. This interplay between genre and syntax also influences parsing decisions. Furthermore, we look at the influence of (sparse) punctuation on (automatically) tagging the corpus itself, and how a closer study of genre-specific syntactic elements contributes to the improvement of the accuracy of automatic classifiers.

1. Introduction

In the last few years, syntactically annotated and parsed corpora, facilitating large-scale comparative and diachronic morphological and syntactic research, have become increasingly important. Such corpora make specifically designed queries reproducible and comparable amongst languages and language stages, due to widely used annotation and parsing standards. The corpus we focus on in this study is the Middle Low German (MLG) subcorpus of the *Corpus of Historical Low German* (CHLG).[1] The CHLG is a syntactically annotated and parsed corpus, which is currently under construction, and which will be specifically designed to enable further research into the syntax of Old Saxon (OS; also known as Old Low German) and Middle Low German.[2] Although there have been a number of recent studies focusing on the syntax of Middle Low German, leading to interesting insights (e.g. Breitbarth 2014; Farasyn & Breitbarth 2016a; Mähl 2014; Petrova 2012), this is still an under-researched field and worth further exploration. So far, it has been described as having its very own syntactic features, setting it apart from Middle Dutch in the (South)west and Middle High German (MHG) in the Southeast in many aspects. The language, although it was very important, and internationally used, especially during

[1] The *Corpus of Historical Low German* (CHLG) is currently funded by the Flemish Hercules foundation/FWO (grant AUGE 13/02 *A parsed Corpus of Historical Low German*).

[2] HeliPaD, the Old Saxon subpart of the CHLG, is already publicly available at http://www.chlg.ac.uk/helipad/corpus.html.

the heyday of the Hanseatic League, never became fully standardized and was never used as a national standard language. From the second half of the 16th century onwards, Early New High German (ENHG) became the language of writing in the area. As a consequence, Low German only continued to exist in the spoken dialects. However, thanks to its erstwhile importance, many written sources have been preserved. The vast majority of MLG texts are secular, ranging from chronicles and laws (city statutes, land property rights, inheritance rights) to charters and administrative texts (correspondences, bills, accounting books). Due to the number of texts, the extent of the area in which the language was spoken, and only partial standardization, compilers of the corpus of MLG need to consider many kinds of variation. Consequently, one of the goals in constructing the corpus is to recognize this variation and the effects it can have on the outcome of linguistic research, and, if possible, to give the researcher using the corpus the possibility to counter or manage these effects by building a corpus balanced for factors influencing variation, including for example the genre of the texts.

In this paper, we focus on the effects that the genre of the text can have on certain syntactic research outcomes, based on preliminary research conducted on the data from the MLG subcorpus of the *Corpus of Historical Low German*. We define genre in this paper as a specific type of text with a specific purpose. A distinction between text types can thus be made on the basis of the purpose of the text, which in the case of the *Corpus of*

Historical Low German expresses itself for instance in history writing (chronicles), defining rights and privileges for areas such as trade and heritage (charters), and defining rules for the community (legal texts/city laws).

In order to make the effects of genre in the corpus more tangible, this paper will present a number of case studies on MLG syntax within the generative framework, from a minimalist perspective. These indicate that syntactic phenomena can be strongly influenced by the genre of the text investigated. Before that, we describe the *Corpus of Historical Low German*, its purpose and the texts in the corpus in section 2. In section 3, we present the case studies, in which we look at the distribution of relative clauses with the comparative particle *alse*, null pronominal arguments, null resumptive pronouns in relative clauses, and double agreement in V2 clauses with and without inversion of subject and finite verb. In this way we discuss properties of specific genres such as charters, religious texts and chronicles. These case studies show that, in some instances, the syntactic observations can be significantly different in different genres, even leading to false claims about continuity or change from OS to MLG. They therefore demonstrate that the variety of genres in the CHLG is needed to counter these and similar effects. Section 3 starts with an example of how the study of a genre can help to improve the accuracy of an automatic classifier by looking at text-specific discourse markers.

2. A parsed corpus of Middle Low German

The MLG subcorpus of the *Corpus of Historical Low German* will be the first fully annotated and parsed corpus of MLG, covering the whole period in which the language was written (i.e. c. 1250–1600). In order to accomplish this goal, the corpus closely collaborates with the *Referenzkorpus Mittelniederdeutsch/Niederrheinisch* (ReN)[3] in sharing texts, transcription and part-of-speech and morphological mark-up. In contrast to ReN, the *Corpus of Historical Low German* also adds an extra layer of syntactic annotation. An important feature is that the *Corpus of Historical Low German*, like ReN, does not start from editions. All transcriptions are based on the original manuscripts; spelling, word breaks, syntactic structures etc. are thus original and not influenced by editorial decisions. The corpus currently consists of 722,000 words, which equals about 13 prose texts, text collections or books of charters. Based on manually developed gold standards, a tagger automatically assigning Part-of-Speech (POS) tags and additional morphological mark-up have been trained. These taggers use the HiNTS tagset, an adaptation of HiTS (Dipper et al. 2013), a fine-grained tagset especially adapted for Middle Low German, in order to make inter-interoperability with ReN possible. The syntactic annotation layer is currently under development. It will follow the Penn Treebank system, developed for the *Penn Parsed Corpora of*

[3] https://vs1.corpora.uni-hamburg.de/ren/

Historical English (Mitchell et al. 1993), to make the corpus comparable with many existing historical corpora using this annotation system, such as for instance the *Icelandic Parsed Historical Corpus* (IcePaHC; Wallenberg et al. 2011), the *Tycho Brahe Parsed Corpus of Historical Portuguese* (Galves & Faria 2010) and the *Parsed Linguistic Atlas of Early Middle English* (LAEME; Truswell et al. 2016). However, as every language has its own specific syntactic properties, some decisions about the exact tags are corpus-specific, and must thus still be made for the *Corpus of Historical Low German*. These decisions include for instance the addition of extended labels such as locative, directional or temporal extensions for adverb phrases. The MLG subpart of the CHLG will follow the guidelines from the HeliPaD, the OS subpart of the corpus, where possible.[4] Middle Low German is the name of a group of related dialects spoken in northern Germany, for which the first written attestations date from the first half of the 13th century. These surface after a period of about 150 years in which only Latin had been written, making it difficult to make statements about continuity between Old Saxon (OS), the predecessor of MLG, and MLG itself. Recent syntactic analyses, e.g. on the distribution of (different types of) referential null subjects in OS and MLG (Walkden 2014, Farasyn & Breitbarth 2016a, respectively) seem to point at some degree of continuity, however, which further underlines the importance of syntactically annotated

[4] http://www.chlg.ac.uk/helipad/corpus.html; see Walkden (2016) for an overview.

corpora. MLG can be divided into two main periods (Peters 2003: 437). The first lasts until about 1350; in this period there were many different scribal dialects, each belonging to a separate smaller region and closer to the spoken language. When the Hanseatic League gained importance, i.e. between about 1350 and 1550, these scribal dialects were partly standardized. They incorporated features of the surrounding dialect or adapted to influential chanceries like the one of Lübeck, in order to facilitate interregional and international correspondence (e.g. in the Baltic region and in the East). MLG lost its role as the leading written language in the area after a transition to ENHG between c. 1550 and 1600 (Peters 1973).

The MLG subcorpus of the *Corpus of Historical Low German* is meant to be balanced concerning different parameters that tend to influence variation.[5] The intention is to keep the corpus as representative as possible for the language as a whole. As a first measure to reach this goal, the corpus only consists of non-translated texts. Diachronic variation is covered by including texts from the whole period in which MLG was written. Only texts which are clearly dated are included in the corpus. The corpus also tries to offer a balanced picture of diatopic variation by including localized texts from the main MLG dialectal areas: Westphalian, Eastphalian (including Elbe-Eastphalian) and North Low German, both from the Saxon

[5] Because of the sparse attestation of OS – the vast majority being religious epic texts in alliterative verse – the OS part of the *Corpus of Historical Low German* obviously can never be balanced for genre, for instance.

heartland (*Altland*, lit. 'old land') and the areas colonized from the 11th century onwards (*Neuland*, lit. 'new land'). Texts from areas that do not belong to present-day Germany, such as the Baltic and Low Prussian areas, are not included in the corpus.[6] Texts answering all these criteria belong to the key text types in the language: Numerous charters and legal documents have been selected alongside narrative texts including chronicles, religious and medical prose texts. Table 1 gives an overview of the texts that have been included in the corpus so far.

[6] They are, however, included in ReN.

Table 1: Texts included in the MLG subpart of the *Corpus of Historical Low German* so far, by scribal language, genre, period and place of origin.

Place	**Scribal language**	**Genre**	**Period**	**Name**	**Number of word tokens**
Braunschweig	Eastphalian (Altland)	charters	13th–15th c.	*Urkundenbuch Braunschweig*	ca. 81000
Herford	Westphalian (Altland)	legal texts	1375	*Herforder Rechtsbuch*	16227
Lübeck	North Low Saxon (Neuland)	charters	13th–15th c.	*Urkundenbuch Lübeck*	ca. 179000
Magdeburg	Elbe-Eastphalian (Neuland)	charters	13th–15th c.	*Magdeburger Urkundenbuch*	ca. 39000
Magdeburg	Elbe-Eastphalian (Neuland)	medical prose	1483	*Promptuarium medicinae*	ca. 128000
Münster	Westphalian (Altland)	religious prose	1444	*Spieghel der leyen*	24505
Münster	Westphalian (Altland)	religious prose	1480	*Dat myrren bundeken*	ca. 91000
Münster	Westphalian (Altland)	charters	14th–15th c.	*Urkundenbuch Münster*	ca. 95500
Oldenburg	North Low Saxon (Altland)	charters	14th–15th c.	*Oldenburger Urkunden*	28241
Oldenburg	North Low Saxon (Altland)	legal texts	1336	*Oldenburger Sachsenspiegel*	24377
Rüthen	Westphalian (Altland)	legal texts	3 parts: c. 1300, c. 1350, 1460–1500	*Statuarrecht Rüthen*	6804
Soest	Westphalian (Altland)	legal texts	c. 1367	*Soester Schrae*	8241

The text selection criteria for the corpus are intended to offer a representative picture of the language written from 1250 until 1600. However, the user of the corpus needs to keep in mind that the corpus data represent scribal languages, and not the spoken dialects. In these scribal languages, the writers/scribes did not try to represent the local dialect, and the difference between spoken and written language might well have been considerable (Fedders 1988).

3. Syntactic variation and the role of genre in the corpus

In-depth studies of the syntax of MLG are still only sparsely available, and only a small number of comparative studies on the effect of the text genre on syntactic phenomena in MLG have been performed. Two of these rare examples are Dreessen & Ihden (2015), for the effect of genre on the placement of the verb in subordinate clauses, and Farasyn et al. (2016) on different aspects of MLG syntax in psalms translated from Latin into MLG compared to these syntactic phenomena in authentic MLG text material. In the first case study presented below, we focus on how investigating genre-specific phenomena can inform the construction of the corpus, and the automatic tools developed for this purpose, such as the part-of-speech tagger or the parser. As an example of this, we will specifically focus on the study of discourse particles. With the other case studies, we want to show

how the inclusion of different genres in the corpus can lead to new and diverging results when studying syntax based on the corpus.

3.1 Discourse markers

In order to train a high-performing part-of-speech tagger, it is crucial to have rich textual information. A part-of-speech tagger always relies in the first place on a set of custom features related to the token, such as word length, first *n* letters, last *n* letters, capitalization and punctuation for tagging tokens in a natural language. Based on these features and robust machine learning algorithms, the tagger learns how tokens can be divided into different categories: in other words, it learns which labels should be assigned to the token. One of the challenges of constructing automated tools to tag a historical corpus is dealing with (a lack of) punctuation, as it is often hard to see where sentences start or stop without having such information. For the construction of an automatic tagger or parser, it is however highly useful if information regarding coherent chunks of information is already (partly) encoded in the training data. In order to indicate finite clauses in the *Corpus of Historical Low German*, clause boundaries need to be inserted manually in the transcription or in the POS tagging phase of the text. These get included as additional features for training the parser. Larger chunks of information are harder to find. That is the reason why several corpus-specific features on which the POS tagger relies were included in the

Corpus of Historical Low German.

The POS tagger was trained on gold standard data from three legal texts, which means that they all belong to the same text-genre. In a later stage, the features will be adapted to be more applicable on other genres as well, as first out-of-domain tests have shown that the accuracy of the POS-tagger drops about 10% when applied on *Spieghel der Leyen*, a religious prose text. An outstanding feature of the legal texts on which the tagger was trained is the use of discourse markers. A recurrent property of charters, for example, which are highly formulaic, is that new parts of the text are always introduced with the same word, *vortmer* ('furthermore'), as can be seen in example (1).

(1) *UOrtmer . js eyner vrowen ere man doyt . wel sey dan nemen eynen anderen man . heuet sey mer kindere dan eyn . so sal sey nemen den derden deyl alles des ghudes . heuet sey nicht mehr dan eyn kint . so nemet sey den haluen deyl .*

Furthermore is a.GEN woman.GEN her husband dead want she then take an other husband has she more children than one so will she take the third part all the.GEN good.GEN has she not more than one child so take she the half part

'Furthermore. If a woman's husband died and she wants to take another man, if she has more children than one, then she will take

the third part of all goods. If she does not have more than one child, she takes half of it.'

(Soest, *Soester Schrae*, 1367)

This word was thus included as a corpus-specific feature, i.e. as a discourse marker, indicating the start of a new sentence or the start of a new chunk of information. The other corpus-specific features which were added to the corpus were brackets and paratext.[7] Testing all possible features and feature combinations separately to see which one(s) perform(s) better is computationally very intensive. A solution was to work with genetic algorithms, i.e. algorithms that are based on the idea of natural selection, which make it possible to look for optimal features or feature combinations much more efficiently. Koleva et al. (2017) gives a more extensive description of how this has been done. When applying genetic algorithms on the gold standard data,[8] the outcome shows that most features (i.e. bigrams (sequences of two tokens), trigrams (sequences of three tokens) and lowercase) related to the token are consistently needed to obtain a well-

[7] Brackets are either part of the original manuscript or decisions made by the transcribers, according to the guidelines in Barteld et al. 2014. Curly brackets are used to resolve abbreviations, square brackets for text that is unreadable/hard to read. Manuscripts usually contain round brackets to indicate non-specified continuation of the text.

[8] The gold standard consists of two texts manually POS-tagged by two annotators, and one by one annotator. In order to reach full inter-annotator agreement on the doubly annotated texts, the double annotations were compared. For some cases, such as idioms or multi-word expressions, there was no tagging consistency between annotators, so consensual decisions were taken.

performing tagger. Corpus-specific features do play a role in all texts, although to a lesser extent: in all experiments on legal texts, *vortmer* is selected in between 50% and 70% of all cases, which means that the classifier partly relies on the information about *vortmer* in between 50 and 70% of all cases to be able to predict the right POS tag for the token. In *Soester Schrae* and the *Herforder Rechtsbuch*, the discourse marker was found in more than 70% of all cases selected.

In later experiments, in which the tagger will be optimized to be applied on out-of-domain texts, i.e. on texts from different cities/periods/genres, further discourse markers indicating new chunks of information can be added as a feature to optimize the tagger. For other genres, the role of other structuring elements that seem to function as chunk-introducing discourse markers in MLG still needs to be investigated more carefully. The word *vnde* (‘and’) for instance seems to have a more discourse structuring role as introducer of new informative chunks as well, rather than being a conjunction (Farasyn & Breitbarth 2016a). Example (2) shows how *Vn(de) se wolden* clearly is not used as a second conjunct, but rather introduces a new piece of information. This function is also emphasized by the capitalization of the *V*.

(2) *Des quemen des bioscopes ammetlude van Mynden in dat erfhus in eyn ghe-heghet richte un(de) anclaghede(n) mit eren vorspreken dessulven Johannes herwede unde sin erve, went he des stichtes vulschuldighe eghene man w(er)e un(de) horde in dat ammet to Hul-Horst.* ***Un(de)*** *se wolden ene vorbosmen un(de) vortughen, alze des ammetes recht is*

this.GEN came the bishop.GEN officials from Minden in the inheritance.house in a limited court and filed.suit with their spokesmen the.same.GEN John.GEN armour and his inheritance for he the.GEN convent.GEN serf were and belonged in the authority of Hüllhorst and they wanted him to.claim.as.serf and testify as the.GEN authority's right is

'Because of that, the officials of the bishop of Minden came in the house of the decedent in a limited court and filed suit with their spokesmen about the armour of this John and his inheritance, as he was (supposedly) a serf of the convent and belonged to the authority of Hüllhorst. **And** they wanted to claim him as a serf, as is the right of the authority'

(Herford, *Herforder Rechtsbuch*, 1375)

3.2 Null pronominal arguments

Much syntactic research that has been done on the basis of the *Corpus of Historical Low German* under construction concerns (referential) null pronominal arguments. In this section, we focus on four types of (referential) null arguments or structures containing these arguments, which have turned out to have a connection with the genre of the texts they occur in: referential null subjects, pronominal gaps in *alse*-clauses, (null) resumptives in non-restrictive relative clauses with a first or second person head and pronominal gaps in asymmetric coordinations.

3.2.1 Referential null subjects

The presence of referential null subjects (RNS) in MLG is a syntactic phenomenon that is strongly related to text type/genre. A sentence containing an RNS is interpreted as if it contains a referential subject, although this subject is not expressed. Based on results of the distribution of RNS in a corpus of twenty MLG texts, both of the MLG subpart of the *Corpus of Historical Low German* and of ReN, Farasyn & Breitbarth (2016b) show that MLG was a partial null subject language, i.e. that the language allowed null subjects under certain conditions (cf. Walkden 2014; Holmberg 2010). One of these conditions, which is a common property of many partial null subject languages, is the preference of the RNS to occur in the main clause. This is the case in clause-initial topic position (3) as well as

in V2 clauses with another topic (4) and in V1 conditional clauses (5) (and other V1 clauses, e.g. interrogatives).[9] RNS occur about three times more often in main clauses (3.3%) than in subordinate clauses (0.9%). Besides clause type, person seems to be an additional important conditioning factor for the occurrence of RNS in MLG, as RNS have a preference to appear in the 2nd and the 3rd person in MLG. This is however a relaxation of the strong preference for 3rd person in predecessor Old Saxon (OS) and related Old High German (OHG).

(3) *Vnde* [pro] *hebbe dyne kyndere beyde seer wol bewaren*
lathen vnde nycht ghedodeth
and [I] have your children both very well protect let and NEG killed
'And I have let both your children be very well protected and have not killed them'

(Hamburg, *Griseldis*, 1502)

---

[9] We can be fairly certain that the gap precedes the finite verb in (3), as there is evidence that MLG did not have inversion in second conjuncts (following *vnde* 'and'). First, we find examples (both from the same text as (3); *Griseldis*) with overt pronominal (i) and full nominal subjects (ii).

(i) *Vnde sze hudden syck nycht vor de bo̊sze anlaghe des vaders*
and they protect REFL NEG for the evil accusation of.the father
'and they do not protect themselves from the evil accusation of the father'

(ii) *vnde de vrowe slot de dore na to*
and the lady locked the door after shut
'and the lady locked the door shut after [them]')

Second, evidence from double agreement (see Section 3.2.4 below) also indicates that there is no inversion in second conjuncts in MLG.

(4) *v(m)me vns to verlose(n) heuest* [pro] *willen anneme(n) vnse kranch(ei)t* […]

for us to redeem have [you] want.IPP on-take our disease

'in order to redeem us, you have wanted to take on our disease'

(Münster, *Dat myrren bundeken*, 1480)

(5) *heuet* [pro] *ene ane bůrghe ghelaten so mach hey dat selue doyn*

has [he/one] him without bailsman left so may he that himself do

'if he/one left him_i without a bailsman, he_i may do that himself.'

(Soest, *Soester Schrae*, 1367; Farasyn & Breitbarth 2016b)

In addition to the internal linguistic properties of null subjects in MLG, the extra-linguistic factors period (of writing) and scribal language turn out to be strong predictors of the expression of null referential subjects: RNS become more common between 1450 and 1550 and are more often used in the Eastphalian dialect, whereas they are almost absent in earlier texts or texts from the area of Lübeck. The strongest extra-linguistic factor predicting the presence of RNS in MLG, however, is genre/text type, as can be seen in Table 1, in which the results of a multiple logistic regression performed in Rbrul are displayed (cf. Farasyn & Breitbarth 2016b).

Table 2. Influence of the factor genre on the expression of a referential pronominal subject as null (cf. Farasyn & Breitbarth 2016b)

genre	log odds	odds	factor weight	N	% RNS
chronicle	1.475	0.075	0.814	425	7.53
letter	0.428	0.028	0.605	216	2.78
religious	0.208	0.022	0.552	1249	2.24
literature	0.074	0.020	0.518	1882	1.97
legal	-0.403	0.012	0.400	1709	1.30
charters	-1.782	0.003	0.144	320	0.69

Table 2 shows that, of the six different genres that have been evaluated, RNS in MLG are much more common in chronicles than in any other genre. It is probably the narrative character of chronicles that leads to this text genre displaying a remarkable 7.53% of RNS, whereas on average RNS only make up 2.12% of all pronominal subjects in the corpus (cf. Farasyn & Breitbarth 2016a). The study of genre in relation to the expression of null subjects shows that genre additionally is a very strong predictor of the type of RNS (SpecCP or the position after C) that is found in the text. An example is the Saxon chronicle (which is currently not in the *Corpus of Historical Low German*, but is part of ReN), in which 31 out of 33 are RNS

occurring in SpecCP, as can be seen in example 6.[10] In this example, we first encounter three cases of regular conjunction reduction. The last two gaps seem to be cases of conjunction reduction, but the subjects do not refer to the subject of the preceding referent (i.e. God, in 'he gave her to Adam as his wife'), as the first gap rather refers to Adam ('He was meant to live forever') and the second one to God ('He forbade him [Adam] to eat fruit from a certain tree'). The example therefore also shows MLG's tendency to use discourse antecedents which are introduced more implicitly, as the antecedent is neither structurally parallel nor c-commanding the gap.

[10] See Farasyn & Breitbarth (2016a) for a distinction between two types of RNS in MLG, viz. null topics in SpecCP and null clitics in Wackernagel position. Note that our treatment of this position differs from that of Wackernagel (1892) in assuming it to be a syntactically-defined (rather than prosodically-defined) position high in the clause structure, but below C: see Lenerz (1977), Anagnostopoulou (2008) and many others for this interpretation.

(6) *Vnd in der ersten stunde des dages mackede got*$_{i}$ *Adame*$_{j}$ *van der erde na synem likenisse vnd* [Ø$_{i}$] *gaff ome gewalt over fee ouer voggel ouer fissche vnd* [Ø$_{i}$] *sande one*$_{j}$ *in dat Paradis dar mackede he Eua van Adames ribbe Jn der dridden stunde des dages die wile dat he*$_{j}$ *sleyp vnd* [Ø$_{i}$] *gaff eua adame*$_{j}$ *to wiue vnd* [pro$_{j}$] *scholde ewich leuen vnde* [pro$_{i}$] *vorbot one frucht an eynem bome to eten*

and in the first hour of-the day made god Adam from the earth to his image and gave him power over mammals over birds over fish and sent him in the paradise there made he Eve from Adam's rib in the third hour of the day the while that he slept and gave Eve Adam to wife and [he] should forever live and [he] forbade him fruit from one tree to eat

'And in the first hour of the day, God created Adam from earth in his image, gave him power over mammals, birds and fish and sent him to paradise. There, he made Eve from Adam's rib in the third hour of the day, while he was asleep, and gave her to Adam as his wife. [He] was meant to live forever and [he] forbade him to eat fruit from a certain tree.'

(*Cronecken der Sassen*, 1492)

The fact that narrative texts display a higher amount of RNS in SpecCP is probably due to a form of emerging discourse drop, as this is the position in which topics occur. This is highly visible in texts with a narrative character

like these chronicles. This case study consequently underlines that the inclusion of different genres is highly important for the study of null subjects: a corpus entirely based on the most common type of texts in MLG, i.e. charters and legal documents, would definitely create a misleading image of the distribution of null subjects in this language. Thanks to the study of narrative texts, however, we can conclude that the type of null subjects in SpecCP show that MLG is in the transition to a topic-drop language of the modern V2-Germanic type, although it displays a certain continuity with Old Saxon in its preference for clause type and person.

3.2.2 Pronominal gaps in alse-*clauses*

In their earlier research on null subjects in MLG, Farasyn & Breitbarth (2016a) report on pronominal gaps in adverbial clauses introduced by the comparative particle *alse* '(just) as'. The gap is located right after *alse*, and hence is in the Wackernagel position, as are the second type of referential null subjects described in the previous subsection. Further evaluation of these cases shows that these sentences behave like relative clauses modifying the whole preceding situation. In example (7), for example, the *alse*-clause refers to the whole action of claiming someone as a serf and testifying.

(7) *Un(de) se wolden ene vorbosmen un(de) vortughen, alze des ammetes recht is*

and they wanted him claim.as.serf and testify, as [_] the.GEN authority's right is

'And they wanted to claim him as a serf and testify, as [it] is the authority's right'

(= '… which is the authority's right')

(Herford, *Herforder Rechtsbuch*, 1375)

The use of *(al)so/als(o)* as (approaching) a relative particle has already been described for MHG and ENHG (Paul [25]2007:405/426; Ebert et al. 1993:447/479). Indeed, in our MLG data, we find cases like (8), where *alse* is clearly used as a relative particle. Instead of referring to a whole preceding situation like (7), (8) has an object gap, and an overt subject (Farasyn & Breitbarth 2016a).

(8) *van wegen eynes huszes alse de obg(ena)nte Jacob van luebeke dem vorb(enomed)en Bernd papke(n) verkofft hadde vp passchen lest vorleden tobetale(n)de.*

Because of a house as the above-mentioned Jacob of Lübeck the.DAT aforementioned Bernd Papken sold had on Easter last past to pay

'because of a house, which the above-mentioned Jacob of

Lübeck had sold to the aforementioned Bernd Papken, to be paid this past Easter'

(Lübeck, *Urkunde 1500b*, 20/01/1500)

In cases like (7), there is no relative or resumptive pronoun overtly realising the subject of the adverbial/relative clause. Farasyn & Breitbarth (2016a) therefore propose to assume a null resumptive pronoun in the Wackernagel position in such cases, which also agrees with the verb. This analysis is further corroborated by null first and second person resumptives in MLG relative clauses, which we discuss in the next subsection.

In the context of the focus of the current article, it is furthermore crucial to note that relative(-like) *alse*-clauses with a subject gap/null resumptive pronoun are almost exclusively found in charters, a highly formalized genre. This again emphasizes the importance of including different genres in a historical corpus. Additionally, not finding these structures in certain genres may be taken as an indication of the fact that the construction is not representative of spoken MLG, but of more formal writing. This contrasts sharply with RNS, for instance, which tend to show up much more frequently in narrative texts, a genre that tends to be closer to the spoken language.

3.2.3 Null resumptives in non-restrictive relative clauses

A third syntactic phenomenon for which the inclusion of different genres in

a representative MLG corpus is indispensable is the variation between overt and null resumptives in non-restrictive relative clauses (NRRCs) with a first or second person head in MLG. In order to understand the possible structures in MLG, a comparison with present-day German (PDG) is useful. In PDG, NRRCs with a first or second person head can in principle take three different forms, though they are not equally acceptable (Ito & Mester 2009, Trutkowski & Weiß 2016).[11] In the first type, there is agreement between the 1st or 2nd person antecedent (head) in the main clause and the verb in the relative clause, which shows a 1st or 2nd person ending (9a). This type of agreement pattern will thus be called head agreement (HA) in what follows. In the second type, agreement seems to be established between the (3rd person) relative pronoun and the finite verb in the relative clause (9b), leading to the term relative pronoun agreement (RPA). The third type looks like the first type in that it has 1st or 2nd person agreement on the verb in the relative clause, but there is an additional resumptive pronoun (9c). This structure can therefore be called resumptive pronoun agreement (ResPA).

(9) a. *Du, der mein Bruder bist, …* (HA)
you, REL my brother are.2.SG
'You, who are my brother'

[11] Although Ito & Mester label the first structure as ungrammatical in PDG, Trutkowski & Weiß (2016) point out, based on a magnitude estimation experiment, that the three structures are all used, but have a varying level of acceptability.

b. *Du, der mein Bruder ist, …* (RPA)

You, REL my brother are.3.SG

c. *Du, der du mein Bruder bist, …* (ResPA)

You, REL RESP.2.SG my brother are.2.SG

In MLG, the only agreement patterns that can be found are HA (10a) and ResPA (10b); no examples of RP have been found so far. HA is a puzzle insofar as noun-verb-agreement is normally clause-bound, and as the head noun is not necessarily a subject, but may be the object, a prepositional object, or a possessive pronoun in the matrix clause. In (10a) an example of a non-restrictive relative clause modifying an object (*dy*) is given, Example (10b) illustrates how the clause modifies the complement (*dy*) of a preposition (*van*) and (10c) shows a clause modifying the possessive *dyner*.

(10) a. *vp dat ick dy* ***de*** *dat ouerste gud bist v(m)me myne eghene traecheit vn(de) vnuulherdicheit nicht en mote verlesen* (HA)

so that I you REL the highest good are.2.SG for my own slowness and unpersistency NEG NEG must.1.SG lose

'so that I mustn't lose you, who are my highest good, because of my own slowness and lack of persistency'

(Münster, *Dat myrren bundeken*, 1480)

b. *meer warhen sal ick van dy vlein **de** du allerwegen*
Jegenwordich byst … (ResPA)
but where.to shall.1.SG I from you flee REL you
everywhere present are.2.SG
'but where will I flee from you, who are present
everywhere...'
(Münster, *Ey(n) Jnnige clage to gode*, 1480)

c. *v(er)beide(n)de de behoerlike tijd **dyner** gheboerten de [ ]*
na dyner godheit ghine tijd en heuest noch iare
biding the appropriate time your.GEN birth REL [ ] after
your divinity no time NEG have.2.SG nor years
'biding the time appropriate for your birth, who has no time
nor years due to your divinity'
(Münster, *Dat myrren bundeken*, 1480)

It is noteworthy that every NRRC is introduced by *de*. While the relative pronoun in PDG (*der, die, das*) spells out gender and number features, this is not the case in MLG for *de*: it is invariable, even though the respective heads can be marked for the number and gender features in question (Farasyn 2017a). The fact that relative pronoun agreement is not attested in MLG raises the question whether the clause-introducing *de* is a relative pronoun at all, or whether it is an invariant relative particle (a syntactic head

in C) instead. As a particle never contains φ-features (i.e. person, gender and number) and therefore cannot agree with the verb, the assumption of a relative particle could explain why 3rd person agreement is impossible. However, even though *de* can be found in free relative clauses as a relative particle in the Eastphalian dialects, additional evidence leads to the assumption that this is not the case in NRRCs. In Example (11), for instance, there are three non-restrictive relative clauses. All contain *de*, followed by either another *de*, or *dar*. We can take the second *de* in these combinations – if present – to be a relative particle *de*, located in C, which alternates with other particles in NRRCs, like *dar*. Semantically, the heads of all three relative clauses in (11) are 2nd person (plural); in the third, *gy* 'you' is even overt.[12] Therefore, Farasyn (2017a) claims that the left-peripheral invariable *de* that is always present is a relative pronoun, though underspecified for number, gender and person, i.e. for features relevant for agreement with the verb inside the relative clause.

[12] In the first two clauses, the 2PL interpretation of *alle* is forced by the pronouns *iu* 'you.ACC' and *iuwes* 'you.GEN/POSS'.

(11) *Vrowet iu in deme heren alle* ***de de*** *enes guden leuendes mit*
ruwen be=gynnet vn(de) bewiset vtwendich de vroude iuwes
herten alle ***de dar*** *vort treden in enem guden leuende*
vn(de) beromet iu der ewighen ere alle gy ***de=de*** *rechtes*
herten sint ane straffinghe iuwer samwitticheit
rejoice you in the lord all REL REL a good life with remorse begins
and prove outwardly the joy of-your heart all REL REL forth go in a
good life and glory in the eternal glory all you REL=REL right heart
are without punishment of-your conscience
'Rejoice in the lord, all who begin a good life with remorse,
and outwardly show the joy of your heart, all who progress in
a good life, and glory in eternal glory, all of you who are of
the right heart without a guilty conscience'.

(Wolfenbüttel, Eastphalian psalms, 15th century)

Given that MLG shows either HN agreement or ResP agreement, and given further that *de* is invariant, and thus either a relative particle or an underspecified (non-agreeing) relative pronoun, Farasyn (2017a) argues that agreement is always established via a resumptive pronoun inside the relative clause. In cases that look like head noun agreement, Farasyn argues for a phonetically null resumptive in the Wackernagel position. This assumption is supported by the fact that the Wackernagel position, i.e. the position right after C, also contains other empty pronouns, which has been described

above for null referential pronouns as well as for *alse*-clauses. The underspecified relative pronoun makes it possible to establish an agreement chain through relations of Checking and Matching, as it mediates between the head in the matrix clause and the (null) resumptive in the relative clause.

NRRCs with 1st and 2nd person heads are only sparsely attested in the texts of the *Corpus of Historical Low German*, in particular with null resumptives/apparent HA. Table 3 shows through multiple logistic regression analysis in Rbrul that such clauses are almost entirely restricted to religious prose texts.[13]

Table 3. Influence of the factor genre on NRRCs with 1st and 2nd person heads and HA

genre	log odds	odds	factor weight	N
religious	7.133	0.013	0.814	2221
letter	6.212	0.004	0.605	273
literature	4.995	0.001	0.552	3216
legal	3.724	0.000	0.518	2158
chronicle	-10.900	0.000	0.400	1002
charters	-11.165	0.000	0.144	1360

[13] N refers to the total number of finite clauses in the current (sub)corpus, the odds ratio indicates the presence of NRRCs with a 1st or 2nd person head (e.g., 0.013 ~ 1.3%).

This means again that it is absolutely necessary to add all kinds of texts to the dataset to be incorporated into the corpus. It also means that religious texts are of high importance to decide on the possible labels and (empty) categories that have to be included in the syntactic mark-up of MLG texts, as they raise the question of whether fixed slots have to be reserved for (null) resumptives, just like it is common to reserve for instance slots for traces in the syntactic annotation layer, in order to make it possible for the researcher using the corpus to search for exactly these clauses.

3.2.4 Pronominal gaps in asymmetric coordinations

Like many other languages, the subject of a second conjunct in a coordination construction is almost always omitted if it is shared with and overtly expressed in the first conjunct. Such conjunction reduction can be seen in example (12).

(12) *dey sal deme Rayde wedden eyn half půnt ande [ ] sal vte deme gherichte ewelike wesen vorwyset*
that-one will the council pay a half pound and [ ] will out the court eternally be outlawed
'he will pay the council half a pound and [ ] will forever be outlawed'

(Soest, *Schrae im Statutenbuch*, 1367)

When conjunction reduction takes place in coordinations with subject-verb inversion in the first conjunct, something remarkable happens. As can be seen in Example (13), the first clause has subject-verb inversion because of the presence of the adverb *Vortmer*. The subject in the second conjunct is omitted, as it is clear from the content that both clauses share the same referent. This does however raise the question where the gap in the second clause should be located. In Example (12) for instance, it is reasonable to assume that both clauses are parallel, since there is no inversion: the overt subject in the first conjunct as well as the subject gap in the second are located in front of the verb. In (13a), however, we do not know if *Vortmer* also induces inversion in the second conjunct, with the subject gap following the verb, as in (13b), or if does not, with an expected gap preceding the verb, as in (13c).

(13) a. *Vortmer,* ***bidde*** *wi vnde* ***manen*** *alle guode*
lude, Houeman, vnde husman
Dat se alle mit eneme schrichte volghen…
furthermore, pray we and demand all good people,
nobleman, and peasant that they all with a complaint follow
'Furthermore we pray and demand from every good
man, nobleman and peasant, that they all sue with a
complaint...'

b. *Vortmer , **bidde** wi vnde **manen** [wi] alle guode lude […]*

c. *Vortmer , **bidde** wi vnde [wi] **manen** alle guode lude […]*

(Lübeck, *Urkundenbuch Lübeck*, 1334)

Examples of coordinations with first person plural subjects like (13), which are particularly found in the opening formulas of charters, are invaluable for solving this puzzle. Note that the ending of the verb in the first conjunct is different from the one in the second conjunct. This is due to the fact that MLG has different agreement morphology on the verb in the 1st and 2nd person plural depending on the position of the pronoun relative to the verb. If the pronoun precedes the verb, the normal ending of the *Einheitsplural* ('unity plural'), which is *-n* or *-t* in MLG depending on period and region/scribal dialect, is used. However, if the pronoun follows the verb, the ending is different: in most cases, the *-n* or the *-t* is dropped. As the *-n* in Example (13) is dropped in the first clause, but not in the second one, this means that the gap in the second clause must be located before the verb and that these coordinations are therefore asymmetric (Farasyn 2017b). This conclusion is supported by research into asymmetric coordinations in High German (Reich 2009).

This case study shows once more how the study of a particular genre can have implications for what we can know about MLG word order. As

dated and localized historical documents containing 1st and 2nd person plural are very hard to find, it shows us that the inclusion of charters in the corpus is paramount for shedding light on the nature of these structures. Once again, it is the in-depth syntactic study of a particular genre that solves a (syntactic) puzzle that could not have been solved without including this specific genre in the dataset. Furthermore, this study demonstrates the importance of the examination of a specific genre for parsing decisions, as in this case, the decision on where the gap in the clause must be located.

4. Summary and outlook

This paper examined the role of genre in the construction of a historical corpus of Middle Low German, incorporating all kinds of variation. A case study on the role of discourse markers made clear how future in-depth study of genre-specific discourse-markers might lead to improvement of the accuracy of the POS-tagger when applied to texts from genres that were not included in the gold standard training data. Multiple syntactic case studies on the role of genre for null referential arguments showed amongst others how genre significantly influences the amount of referential null subjects and how formulaic structures in the legal genre can lead to the discovery of *alse* as a relative particle. The study of religious texts led in its turn to the discovery of a (null) resumptive inside the non-restrictive relative clause

and delivered insights into the word order of (asymmetric) conjuncts in MLG. All of these insights can be used to adapt and extend the labels in the layer of syntactic annotation.

References

Anagnostopoulou, Elena. 2008. Notes on the Person Case Constraint in Germanic (with special reference to German). In *Agreement Restrictions*, Roberta D'Alessandro, Susann Fischer & Gunnar Hrafn Hrafnbjargarson (eds), 15–48. Berlin: Mouton de Gruyter.

Barteld, Fabian, Dreessen, Katharina, Ihden, Sarah, Schröder, Ingrid, Glawe, Meike, Kleymann, Verena, Nagel, Norbert, Peters, Robert & Schilling, Elmar. 2014. Transkriptionshandbuch des Projekts Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200–1650) (Version 1. August 2014, Auszug). https://vs1.corpora.uni-hamburg.de/ren/media/pdf/transkriptionshb1.pdf

Breitbarth, Anne. 2014. *The History of Low German Negation*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199687282.001.0001

Dipper, Stefanie, Donhauser, Karin, Klein, Thomas, Linde, Sonja, Müller, Stefan & Wegera, Klaus-Peter. 2013. HiTS: Ein Tagset für historische Sprachstufen des Deutschen. *JLCL* 28(1): 85–137.

Dreessen, Katharina & Ihden, Sarah. 2015. Korpuslinguistische Studien zur mittelniederdeutschen Syntax. *Jahrbuch für Germanistische*

Sprachgeschichte 6(1): 249–275. https://doi.org/10.1515/jbgsg-2015-0016

Ebert, Robert Peter, Reichmann, Oskar, Solms, Hans-Joachim & Wegera, Klaus-Peter. 1993. *Frühneuhochdeutsche Grammatik*. Tübingen: Niemeyer. https://doi.org/10.1515/9783110920130

Farasyn, Melissa. 2017a. Kongruenzmuster in mittelniederdeutschen Relativsätzen: Eine Pilotstudie. In *Aktuelle Tendenzen in der Variationslinguistik (Kleine und regionale Sprachen)*, Line-Marie Hohenstein, Kathrin Weber, Heike Wermer, Meike Glawe & Stephanie Sauermilch (eds), 67–90. Hildesheim: Georg Olms.

Farasyn, Melissa. 2017b. Deletion in the verbal paradigm in Middle Low German: An interface phenomenon. Ms, Ghent University.

Farasyn, Melissa & Breitbarth, Anne. 2016a. Nullsubjekte im Mittelniederdeutschen. *Beiträge zur Geschichte der Deutschen Sprache und Literatur* 138(4): 524–559. https://doi.org/10.1515/bgsl-2016-0040

Farasyn, Melissa & Breitbarth, Anne. 2016b. Null Subjects in Middle Low German. Ms, Ghent University.

Farasyn, Melissa, Witzenhausen, Elisabeth & Breitbarth, Anne. 2016. Anmerkungen zur mittelniederdeutschen Syntax in Psalmenübersetzungen des (14. und) 15. Jahrhunderts. Ms, Ghent

University.

Fedders, Wolfgang. 1988. Zur Erhebung historischer Schreibsprachdaten aus der Textsorte ‘Urkunde’. *Niederdeutsches Wort* 28: 61–74.

Galves, Charlotte, Andrade, Aroldo Leal de, & Faria, Pablo. 2017. Tycho Brahe Parsed Corpus of Historical Portuguese. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip.

Holmberg, Anders. 2010. Null subject parameters. In *Parametric Variation: Null Subjects in Minimalist Theory*, Theresa Biberauer, Anders Holmberg, Ian Roberts & Michelle Sheehan (eds), 88–124. Cambridge: CUP.

Ito, Junko & Mester, Armin. 2000. Ich, der ich sechzig bin: An agreement puzzle. In *Jorge Hankamer WebFest*, Sandy Chung, Jim McCloskey & Nathan Sanders (eds). http://ling.ucsc.edu/Jorge/ito_mester.html

Koleva, Mariya, Farasyn, Melissa, Desmet, Bart, Breitbarth, Anne & Hoste, Veronique. 2017. An automatic part-of-speech tagger for Middle Low German. *International Journal of Corpus Linguistics* 22(1): 108–141. https://doi.org/10.1075/ijcl.22.1.05kol

Lenerz, Jürgen. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.

Mähl, Stefan. 2014. *Mehrgliedrige Verbalkomplexe im Mittelniederdeutschen. Ein Beitrag zu einer historischen Syntax des*

Deutschen. Köln: Böhlau.

Paul, Hermann. [25]2007. *Mittelhochdeutsche Grammatik*, neu bearb. von Thomas Klein, Hans-Joachim Solms & Klaus-Peter Wegera. Mit einer Syntax von Ingeborg Schöbler, neu bearb. von Heinz-Peter Prell. Tübingen: Niemeyer.

Peters, Robert. 1973. Mittelniederdeutsche Sprache. *Niederdeutsch. Sprache und Literatur* 1: 66–129.

Peters, Robert. 2003. Variation und Ausgleich in den mittelniederdeutschen Schreibsprachen. In *The Dawn of the Written Vernacular in Western Europe*, Michèle Goyens & Werner Verbeke (eds), 427–440. Leuven: Leuven University Press.

Petrova, Svetlana. 2012. Multiple XP-fronting in Middle Low German root clauses. *The Journal of Comparative Germanic Linguistics* 15(2): 157–188. https://doi.org/10.1007/s10828-012-9050-y

Reich, Ingo. 2009. *'Asymmetrische Koordination' im Deutschen*. Tübingen: Stauffenburg.

Truswell, Robert, Alcorn, Rhona & Donaldson, James. 2016. A parsed linguistic atlas of Early Middle English. Paper presented at the first Angus McIntosh Centre symposium, June 10, 2016. Slides available at http://robtruswell.com/assets/pdfs/AMC_talk.pdf

Trutkowski, Ewa & Weiß, Helmut. 2016. When personal pronouns

compete with relative pronouns. In *The Impact of Pronominal Form on Interpretation*, Patrick Grosz & Pritty Patel-Grosz (eds), 135–166. Berlin: De Gruyter.

Wackernagel, Jacob. 1892. Über ein Gesetz der indogermanischen Wortstellung. *Indogermanische Forschungen* 1: 333–436. https://doi.org/10.1515/9783110242430.333

Walkden, George. 2014. *Syntactic Reconstruction and Proto-Germanic*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780198712299.001.0001

Walkden, George. 2016. The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics* 21(4): 559–571. https://doi.org/10.1075/ijcl.21.4.05wal

Wallenberg, Joel C., Ingason, Anton Karl, Sigurðsson, Einar Freyr & Rögnvaldsson, Eiríkur. 2011. Icelandic Parsed Historical Corpus (IcePaHC), Version 0.9. http://www.linguist.is/icelandic_treebank