A Penn-style Treebank of Middle Low German

Hannah Booth, Anne Breitbarth, Aaron Ecay, Melissa Farasyn
Ghent University
Blandijnberg 2, B-9000 Ghent, Belgium
{hannah.booth, anne.breitbarth, melissa.farasyn}@ugent.be, aaronecay@gmail.com

Abstract

We outline the issues and decisions involved in creating a Penn-style treebank of Middle Low German (MLG, 1200-1650), which will form part of the Corpus of Historical Low German (CHLG). The attestation for MLG is rich, but the syntax of the language remains relatively understudied. The development of a syntactically annotated corpus for the language will facilitate future studies with a strong empirical basis, building on recent work which indicates that, syntactically, MLG occupies a position in its own right within West Germanic. In this paper, we describe the background for the corpus and the process by which texts were selected to be included. In particular, we focus on the decisions involved in the syntactic annotation of the corpus, specifically, the practical and linguistic reasons for adopting the Penn annotation scheme, the stages of the annotation process itself, and how we have adapted the Penn scheme for syntactic features specific to MLG. We also discuss the issue of data uncertainty, which is a major issue when building a corpus of an under-researched language stage like MLG, and some novel ways in which we capture this uncertainty in the annotation.

Keywords: annotation, historical treebank, Low German, parsed corpus, uncertainty

1. Introduction

The development of diachronic parsed corpora in recent decades has played a crucial role in facilitating quantitative studies of syntactic change and reproducible findings which have a strong empirical basis; see e.g. Pintzuk et al. (2017) for an overview. Historical parsed corpora now exist for a wide range of languages, including English (Taylor et al., 2003; Kroch and Taylor, 2000; Kroch et al., 2004), Icelandic (Wallenberg et al., 2011), French (Martineau et al., 2010) and Portuguese (Galves et al., 2017). One such resource currently under development is the Corpus of Historical Low German ('CHLG'), with support from the Hercules Foundation/FWO.[1] The CHLG is planned as a diachronic parsed corpus spanning Old Low German/Old Saxon (OLG, c.800-1050) and Middle Low German (MLG, c.1250-1600).[2] The OLG component is already available as a self-contained corpus, the HeliPaD (Walkden, 2015). The HeliPaD comprises 46,067 words from a single (the largest) OLG text, the *Heliand*, which is annotated according to the Penn standard for historical English (Santorini, 2010); for more details, see Walkden (2016).

In this paper, we report on the second part of the CHLG, the MLG component, which is currently under development, and will be searchable online at `www.chlg.ugent.be`. The attestation for MLG is rich, but the syntax of the language remains relatively understudied, in part due to the extant texts not being readily accessible, nor presented in a way to facilitate corpus-based syntactic studies. Contrary to earlier assumptions that MLG does not differ much syntactically from Middle/Early New High German (Saltveit, 1970; Rösler, 1997), a recent surge in research has brought to light a number of syntactic properties of MLG which demonstrate its position in its own right within West Germanic (Mähl, 2004; Mähl, 2014; Tophinke, 2009; Petrova, 2012; Wallmeier, 2015; Merten, 2015; Farasyn and Breitbarth, 2016; Farasyn, 2018; Breitbarth, to appear). While individual phenomena are now beginning to be understood, MLG syntax remains under-researched compared to other historical Germanic languages. The CHLG will enable deeper insights into the MLG syntactic system, as made elsewhere in Germanic where parsed corpora already exist. The corpus is a collaboration with the Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650), ('ReN') (ReN-Team, 2017).[3] Like the HeliPaD, the syntactic annotation follows the Penn standard for historical English. The syntactic annotation is the feature which sets the CHLG apart from other (only POS-tagged) digital corpora of historical German varieties such as the ReN, which is part of the wider Deutsch Diachron Digital (DDD) initiative, together with the Referenzkorpus Altdeutsch (750-1050) (Donhauser, 2015) and the Referenzkorpus Mittelhochdeutsch (1050-1350) (Petran et al., 2016).

Applying an annotation standard like the Penn scheme to MLG for the first time raises various issues, which are interesting both from a corpus-design as well as a strictly linguistic perspective. In this paper, we outline some of these issues and the steps we have taken in building the corpus. In Section 2., we discuss the relationship between the CHLG's MLG component and the ReN. In Section 3., we outline the texts which have been selected to be included in the corpus from the extensive MLG attestation. Section 4. discusses various aspects of the annotation, including the annotation process and novel principles for treating special syntactic phenomena of Middle Low German, as well as linguistic uncertainty. Section 5. concludes the paper.

[1] Grant number Hercules AUGE13/02 (1 July 2014–31 December 2015)/FWO G0F2614N (1 January 2016-present).

[2] MLG refers to a group of West Germanic dialects written in several scribal dialects in what is now northern Germany and the (north-)eastern Netherlands, which did not undergo the High German sound shift (i.e. *water, dorp, maken*). While LG dialects continued to be spoken, the written language was replaced by Early New High German during the 16th century (Peters, 2015).

[3] The ReN is available via the ANNIS-platform: `http://annis.corpora.uni-hamburg.de:8080/gui/ren`.

2. Collaboration with the ReN

2.1. The ReN

As mentioned, the MLG component of the CHLG is a collaboration with the ReN. The ReN contains approximately 1.45 million words across 145 texts. Each text is presented in a diplomatic transcription, and is lemmatised, POS-tagged and annotated for morphological information; for more details, see Barteld et al. (2017). The tagset used for the POS-tagging of the texts in the ReN is the Historische Niederdeutsch-Tagset (HiNTS) (Barteld et al., 2018), based on the Historische Tagset (HiTS) (Dipper et al., 2013), which is in turn based on the Stuttgart-Tübingen Tagset (STTS) (Schiller et al., 1999), the default tagset for German.

2.2. Combining HiNTs with the Penn scheme

The POS-tagged texts from the ReN feed into the MLG component of the CHLG, where we add a Penn-style constituency-based syntactic annotation on top of the POS-tags, as we discuss in Section 4. This means that, unlike other Penn-style treebanks, the MLG corpus does not use the standard and rather broad tagset intended in the original Penn annotation scheme (Santorini, 1990), but instead the HiNTS tagset as used in the ReN, which involves a much more fine-grained set of distinctions. For instance, where the standard Penn tagset has just one POS-tag for all determiner-like categories ('D'), the HiNTS tagset distinguishes between 26 POS-tags for determiner-like categories, all of which have the basic label 'D' but differ in terms of the features shown in Table 1.

Feature	Possible value	Label
lexical identity	definite	D
	indefinite	I
	negative	NEG
	possessive	POS
	relative	REL
	wh	W
articleness	article-like	ART
	not article-like	–
position	pre-head	A
	post-head	N
	head	S
	head of nominal predicate	D

Table 1: Features of determiners encoded in the HiNTs tagset

Thus, a determiner which is definite, article-like and precedes its head noun, e.g. ***dat*** *land* 'that land', bears the POS-tag 'DDARTA'. A determiner which is negative, not article-like and the head of a nominal predicate, e.g. *dat ist* ***niemand*** 'that is no-one', bears the POS-tag 'DNEGD'.

Using the HiNTS tagset for the MLG component of the CHLG has two main practical advantages. Firstly, it means that we can use the readily available POS-tagged texts from the ReN, thus speeding up the corpus-building process. Secondly, it means that the corpus is in line with other historical corpora of German which employ a version of the HiTS tagset, and so will be easily accessible to researchers already familiar with these resources.

Nevertheless, using the HiNTS tagset also brings certain issues. Firstly, on a practical level, it means that the MLG component of the CHLG diverts from the OLG component (HeliPaD, see above), which instead employs the standard Penn tagset. Such discrepancies in annotation across what is intended as a single resource are not ideal. Furthermore, while the HiNTs tagset will be familiar to those working with corpora of historical German, its more fine-grained nature may pose a challenge for those used to working with the broader tagset in pre-existing Penn-style treebanks. For this reason, we leave open the future possibility to produce an alternative version of the corpus, with the POS-tags converted from HiNTS to the standard Penn tagset. This would allow a wider pool of researchers to easily use the MLG corpus.

The second issue arising from the HiNTs tagset is strictly linguistic in nature. As already demonstrated for determiner-like elements (see Table 1), the HiNTS tags actually encode some information about word order and constituency (e.g. pre-head/post-head). This is in line with the fact that the tagset was initially designed for corpora where there is no additional layer of syntactic annotation. Combining the HiNTs tags with the Penn constituency-based annotation thus results in some redundancy in terms of the encoding of syntactic information. Nevertheless, it makes little sense to lose altogether the rich information encoded in the HiNTs tags. For now, we retain these tags and leave open the possibility that they could be converted to the broader Penn tagset in future, as already mentioned.

3. Text selection

The aim of the CHLG is to produce a corpus of the MLG language which is as representative as possible. As such, there are various dimensions to consider, most notably the spatial (the texts should span a number of scribal dialects), the diachronic (the texts should represent various subperiods within the overall period) and the generic (the texts should represent a range of different text types). Unfortunately, when dealing with a historically attested language stage it is rarely possible to fully satisfy all of these ideals, due to the fact that the extant texts which have fortunately survived rarely offer a balanced picture. For the MLG component of the CHLG, best efforts have been made to design a corpus which is well balanced with respect to these three dimensions, taking into account (a) the available texts and (b) the fact that the annotation process is very time-consuming. Unfortunately, this means that the corpus cannot cover all of the scribal dialects represented in the ReN. However, the four scribal dialects which are included – Westphalian (WP), Eastphalian (EP), North Low German (NLS), and Eastelbian (EE) – are each robustly represented. An overview of the texts which are intended to be included in the corpus is given in Table 2, together with the (sometimes approximate) dating of each text, and the dialect and genre they represent.

Text	Dialect	Date	Genre
Arzneibuch	WP	1451-1500	science
Herforder		1375	law
Soest		1367	law
Spieghel		1444	religion
Braunschweig	EP	1301-1500	law/charter
Duderstadt		1279	law
Engelhus		1435	literature
Zeno		1401-1450	literature
Bremen	NLS	1300-1350	law/charter
Buxtehuder		1451-1500	religion
Griseldis		1502	literature
Oldenburg		1350-1500	law/charter
Willeken		1535	private letter
Flos	EE	1401-1450	literature
Greifswald		1451	law
Rostock		1580	law
Schwerin		1451-1500	law
Stralsund		1301-1500	law/charter

Table 2: MLG texts in the CHLG

4. Annotation

The ReN – with its relatively large scope, diplomatic transcriptions, lemmatisation, POS-tagging and rich morpholgical information – constitutes a valuable new resource for research on MLG. Indeed, the ReN already enables a certain level of syntactic investigation, by making searching within clause-spans possible (an innovation compared to other DDD-corpora) and by using the fine-grained HiNTS tagset which encodes some word order information, as discussed in Section 2. However, syntax extends beyond adjacency and manipulates constituents, not words alone. While the annotation of the clause-span in the ReN is very helpful, there is no efficient way to search for syntactic phenomena extending beyond adjacency and clause boundaries, e.g. verb placement w.r.t. other constituents, as in Petrova (2012) and Breitbarth (to appear). Therefore, a more sophisticated system which encodes linear, hierarchical, and functional relations between constituents is required to enable in-depth syntactic studies. The Penn annotation standard for historical English is one such system which we adopt for the syntactic annotation for a number of reasons, as we next explain.

4.1. Choice of annotation scheme

The Penn annotation standard for historical English is constituency-based, which sets it apart from dependency-based annotation schemes, e.g. Universal Dependencies (Nivre et al., 2016). Dependency-based schemes have also been used for historical corpora, e.g. in the Pragmatic Resources in Old Indo-European Languages (PROIEL) family of treebanks (Haug and Jøhndal, 2008). Neither system is superior to the other; as Taylor (2020) points out, the same questions can be investigated with either annotation type.

Nevertheless, there are several practical reasons why the constituency-based Penn annotation standard is a good choice for the CHLG. Firstly, we are able to take advantage of the work that has gone into defining Penn-style annotation schemes for previous corpora, including those for closely related Germanic languages, such as for OLG (Walkden, 2015) and for Early New High German (Light, 2011). Secondly, computational tools developed for Penn-style corpora, including the CorpusSearch query program (Randall, 2005) and the syntactic annotation GUI Annotald (Beck et al., 2015) can be used without modification. Thirdly, the background familiarity and expertise held by researchers working with Penn-style corpora will apply in a straightforward way to the CHLG, ensuring that the information it contains will be immediately accessible to a community of historical syntacticians – especially (but by no means exclusively) those working on historical Germanic varieties.

Moreover, there are also strictly linguistic motivations for employing the Penn annotation scheme over other schemes used for parsed corpora of German, such as TüBa-D/Z (Telljohann et al., 2006) or TIGER (Albert et al., 2003). TüBa-D/Z, for instance, has a separate annotation level for topological fields, achieving the primary ordering of constituents. An example is shown in Figure 1 (from Telljohann et al. 2015:29), where VF stands for prefield ('Vorfeld'), LK for left sentential bracket ('linke Klammer'), MF for middle field, and VC for verbal complex.

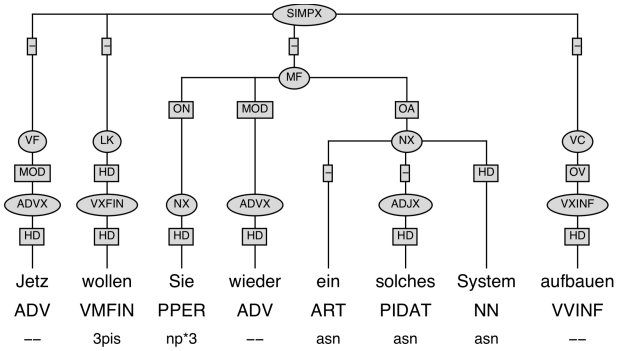

Figure 1: Example of the TüBa-D/Z annotation scheme

The TüBa-D/Z scheme thus relies on being able to clearly identify the verbal brace structure of a particular sentence. However, this is not always possible for MLG. While all studies to date agree that MLG was in principle a verb-second language with a head-final VP (Rösler, 1997; Petrova, 2012; Mähl, 2014; Wallmeier, 2015; Dreessen and Ihden, 2015), there are various word orders which deviate from this. Firstly, non-verbal XPs can be interspersed with verbal parts of the cluster, giving rise to so-called 'Distanzstellung' ('distance positioning') (Mähl, 2014), e.g. (1) (from Mähl, 2014:93).

(1) wente **dat** ik di mit my **mach** in mynes
until COMP I you with me may in my.GEN
vaders lant **voren**
father.GEN land lead
'until I may lead you with me into my father's land'

Secondly, the extraposition of adjuncts and arguments from the middle field is characteristic of MLG. In cases where all elements of the middle field are extraposed, this can lead to 'Kontaktstellung' ('contact positioning') of the verbs (Mähl, 2014), e.g. (2) (from Mähl, 2014:83).

(2) Vnde ik **hebbe gegeuen** [deme huse dines
and I have given the house your.GEN
vaders] [alle dat offer der
father.GEN all the sacrifice the.GEN
kindere van Ysrahel].
children.GEN of Israel
'And I have given to your father's house the whole sacrifice of the children of Israel.'

Given these issues concerning the verbal brace, we have opted for the Penn annotation scheme which, unlike TüBa-D/Z, remains neutral with respect to topological positions. A key feature of the TIGER scheme, used for instance for the Mercurius Treebank of ENHG newspapers (Demske, 2007), is the variable use of a VP-node, depending on the complexity of the verb form. In sentences where the verbal complex contains more than two (non-finite) verbs, this results in the nesting of multiple VPs, e.g. Figure 2 (Albert et al., 2003, 50). This inevitably makes search queries more complex as, hierarchically, verbs may be more or fewer nodes removed from other constituents. This is another motivation to use the Penn annotation scheme, which does not annotate VP-nodes as a general principle. As finite and non-finite verbs are immediate daughters of their containing clause (IP), just like arguments, precedence and dominance relations can be straightforwardly used to find verbs relative to other (sister) constituents of the containing clause.

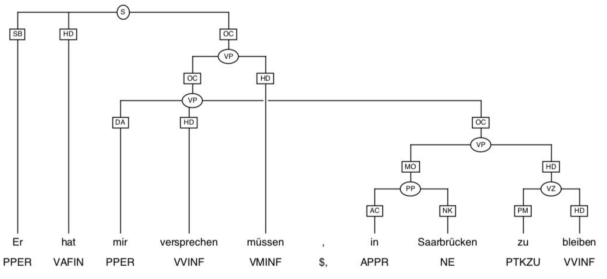

Figure 2: Nested VPs in the TIGER annotation scheme

4.2. Annotation process

The annotation process is composed of interleaving phrases of automated (computer) annotation and manual (human) annotation. This approach is designed to maximise on the natural strengths and weakness of computers versus humans. The process is conducted using Annotald (Beck et al., 2015), a GUI developed specifically for the syntactic annotation of Penn-style treebanks. The pipeline can be summarised in terms of five stages:

1. Automatic rule-based parser: basic constituents resolved at a single (IP-)level.
2. Conversion of the POS-tagged files to Annotald input.
3. First manual pass: erroneous constituents corrected; arguments labelled for grammatical function; empty categories inserted.
4. Automatic rule-based validation: warnings activated at points which deviate from the annotation guidelines.
5. Second manual pass: remaining errors and inconsistencies corrected.

At the first stage, the automatic rule-based parser mainly focuses on linear structure, relying on (non-)preceding or (non-)succeeding parts-of-speech in order to form constituents. A series of basic phrase-structure rules are applied which generate flat structures with no recursion, e.g. (3).

(3) NP → Det Adj* Noun

During this intermediate step towards a fully parsed text, the shallow parser assigns a constituent label to each token, indicating the type of constituent it belongs to, as well as its position in the constituent. In (4), constituent labels starting with B indicate the first word of the constituent, the ones with I the other words in the constituent. The O tag is used when a token does not belong to any constituent.

(4) In → B-PP
dem → B-NP
beginne → I-NP
was → B-VP
dat → B-NP
wort → I-NP
. → 0

At the second stage, the output of the POS-tagger and the automatic rule-based parser is converted to a labelled bracketing structure which can be read by Annotald. The labelled brackets include an intermediate tag for each token, which consists of the HiNTs POS-tag and the constituent label assigned by the shallow parser, as well as the original MLG token. As main clause boundaries are marked during the transcription stage of the project, the input of the CHLG is immediately divided up into matrix clauses (labelled IP-MAT).

(5)
```
( (IP-MAT  (APPR.Dat@B-PP@ In)
(DDARTA.Neut.Dat.Sg@B-NP@ deme)
(NA.Neut.Dat.Sg@I-NP@ beginne)
(VVFIN.Irr.3.Sg.Past.Ind@B-VP@ was)
(DDARTA.Neut.Nom.Sg@B-NP@ dat)
(NA.Neut.Nom.Sg@I-NP@ wort)
(!!ED!!@O@ $.$)))
```

The full parse of e.g. the first NP in (5) after consequent (semi)manual parsing in Annotald is (6).[4]

(6)
```
(NP (DDARTA (META (CASE dat)
                  (GENDER neut)
                  (LEMMA dē)
                  (NUMBER sg))
            (ORTHO deme))
    (NA (META (CASE dat)
              (GENDER neut)
              (LEMMA begin)
              (NUMBER sg))
        (ORTHO beginne))))
```

[4] In what follows, we omit the `(META ...)` tags from the parsed structures, for readability.

Stage four, the automatic validation, takes the form of a series of python subroutines which examine trees in a corpus file and flag constituents which violate basic annotation principles, many of which are analogous to wellformedness constraints in the syntactic literature, e.g. endocentricity (every phrase must have a head), the subject condition (every clause must have a subject) and selectional restrictions (e.g. prepositions should have exactly one nominal or clausal complement).

4.3. Adapting the Penn scheme for MLG

The Penn annotation scheme is not designed to reflect a particular theoretical analysis but rather to make data easily operationalisable for the annotator, i.e. minimising the number of subjective decisions, while also facilitating efficient searching for the corpus-user. Following this principle, we have adapted the Penn scheme to handle special features particular to MLG syntax. Some of these decisions are outlined in this section.

4.3.1. Pronominal adverbs

Like its High German counterpart, MLG exhibits pronominal adverbs which consist of an uninflected pro-particle and a preposition (e.g. *dar-umme*, *dar-von*) and express anaphoric or cataphoric reference. Pronominal adverbs occur very frequently in MLG texts and are commonly discontinuous. We treat them as PPs, headed by the preposition. The pro-particle has its own POS-tag in the HiNTs scheme ('PAVKO') and projects an AdvP which is the sister of the head P. A continuous example is shown in (7).

(7)
```
(IP-MAT (PP (ADVP (PAVKO Dar))
            (PAVAP (vmme)))
        (VVFIN is)
        (NP-SBJ-1 (DPDS dat))
        (ADJP-PRD (ADJD nutte))
        (CP-THT-1 ...))
```
'therefore it is useful that...' (Engelhus)

Discontinuous pronominal adverbs are treated following the same principles, with the discontinuity captured using the standard Penn treatment for movement (insertion of an *ICH* which is coindexed with the 'moved' constituent), e.g. (8).

(8)
```
(IP-MAT (ADVP-1 (PAVKO dar))
        (VAFIN heuit)
        (NP-SBJ (DDARTA de)
                (NA uogit))
        (NP-OB2 (DNEGA nen)
                (NA recht))
        (PP (ADVP *ICH*-1)
            (PAVAP an))
```
'the representative has no right to that' (Duderstadt)

4.3.2. Multi-word adpositions

Multi-word adpositions are another feature of MLG, e.g. *wente an*, 'until'. In cases where both elements are clearly prepositional in nature, we treat both elements as prepositions which head the PP, e.g. (9). The same principle applies for discontinuous cases, e.g. (10). Clearly, on a theoretical level this violates endocentricity, but this treatment offers a simple approach which is easy for the annotator to implement and facilitates efficient searching.

(9)
```
(PP (APPR wente)
    (APPR an)
    (NP (NP-POS (NA godes)
                (NA ghebort)))
```
'until God's birth' (Engelhus)

(10)
```
(PP (APPR bet)
    (ADVP (AVD bauen))
    (APPO an))
```
'up until the top' (Buxtehuder)

In cases where a multi-word adposition is emerging via grammaticalisation of the combination P and N, e.g. *van halven* 'by means of' or *van wegen* 'because of', we follow the ReN POS-tagging decision which treats the second element conservatively as a noun (NA), i.e. does not assume that the grammaticalisation process is fully completed.

(11)
```
(PP (APPR von)
    (NP (DPOSA orer)
        (ADJA eyghen)
        (NA sunde)
    (NA wegen))
```
'because of her own sin' (Engelhus)

For all such cases, we carefully document the constructions involved so they can be easily recovered if one is interested in this particular type of grammaticalisation. In this way, we leave the analysis as to how grammaticalised such constructions are to the corpus-user.

4.3.3. The CP-layer

In the Penn annotation scheme, clauses are by default annotated as IPs. An additional CP-layer is only postulated for finite subordinate clauses (complement, adverbial, degree etc.) and clauses with *wh*-movement (questions, relative clauses). Moreover, all CPs are generally required to have a complementiser present in the annotation. In the absence of an overt complementiser, an empty category (C 0) is thus inserted.

In the MLG component of the CHLG, we have elected not to insert empty complementisers as a general rule, but rather in a restricted set of three contexts. The guiding principle we follow is that each and every CP must have a daughter which licenses it (either a complementiser in C or a *wh*-phrase in SpecCP). We thus insert a null complementiser in V1 conditionals, with an additional extended label (C-**V1** 0), e.g. (12).

(12)
```
(CP-ADV (C-V1 0)
        (IP-SUB (VVFIN Were)
                (NP-SBJ-1 (PPER t))
                (ADVP (AVKO ok))
                (CP-THT-1 (KOUS dat)
                          ...)))
```
'if it were the case that...' (Bremen)

The second context where we insert a null complementiser is in asyndetic dependent V2 clauses (irrelevance conditionals, e.g. (13) and exceptive clauses, e.g. (14)), where

the main marker of dependency is subjunctive marking on the finite verb. These clauses generally have conditional semantics but, atypically, are not introduced by an overt complementiser or the finite verb in first position. Again, the specific type of null complementiser inserted in such contexts bears its own extended label (C-**SUBJ** 0).

(13)
```
(CP-ADV (C-SUBJ 0)
        (IP-SUB (NP-SBJ (PPER he))
                (VVFIN gewinne)
                (KON eder)
                (VVFIN verlese)))
```
'whether he win or lose' (Braunschweig)

(14)
```
(CP-ADV (C-SUBJ 0)
        (IP-SUB (NP-SBJ (PPER he))
                (PTKNEG ne)
                (VVFIN hebbe)
                 (NP-OB1 (NA prouend))
))
```
'if he doesn't have a living...' (Braunschweig)

The third context is with *verba dicendi* which take a *that*-complement (CP-THT), where a null C alternates with *dat* ('that'), e.g. (15). In such instances, the null complementiser has a bare label (C 0).

(15)
```
(IP-MAT (ADVP (AVKO Ok))
        (VVFIN sede)
        (NP-SBJ (DPIS me))
        (CP-THT (C 0)
                (IP-SUB ...)))
```
'and one says that...' (Engelhus)

This system was designed to make the annotation more robust. In the standard Penn scheme, it is difficult to automatically verify the correctness of the annotation of null complementisers. For instance, if a CP contains no C, that could be because it is a V1 conditional or because a null C was mistakenly not inserted. With these revised guidelines, automatic verification should be more straightforward. Moreover, the use of the extended labels -V1 and -SUBJ enables the researcher to easily isolate different types of null complementiser contexts, compared to the general Penn scheme, which does not encode such distinctions.

4.3.4. Left-dislocation and resumption

MLG legal texts in particular exhibit various high-frequency features which pose a challenge for the Penn annotation scheme. One such issue concerns left-dislocation and resumption structures. The Penn scheme works on the basis that clauses usually have one left-dislocated and resumptive pair, with only rare exceptions. As such, left-dislocated constituents and their corresponding resumptive elements are not co-indexed. However, clauses with more than one left-dislocated and resumptive pair are relatively common in MLG legal texts. In such cases, we use co-indexation to capture the various pairings, e.g. (16).

(16)
```
(IP-MAT (NP-LFD-1 (FM Ieu))
        (ADVP-LFD-2 (KOUS do)
                    (CP-ADV ...))
        (ADVP-RSP-2 (AVD do))
        (VVFIN ghewan)
        (NP-SBJ-RSP-1 (PPER he))
        (NP-OB1 (NE Saruth)))
```
'Ieu, when (...), then he got Saruth' (Engelhus)

Another problematic structure not uncommon in MLG legal texts is where there is more than one left-dislocated constituent of the same category and a single resumptive element which could in principle pair with either one. We adopt an innovation for such structures, inserting a null resumptive element (*-RSP 0) to correspond with the additional left-dislocated constituent(s), e.g. (17). As a default principle, the null resumptive element is inserted immediately following the first left-dislocated constituent. The insertion of a null resumptive element in such cases means that each left-dislocated constituent has a corresponding resumptive element, and the coindexation principle can be applied as above. We extend this innovation to instances where there are two or more left-dislocated constituents, of which only the latter is overtly resumed at clause-level, e.g. (18). Here, we insert a null resumptive element directly after the first left-dislocated constituent, and coindex the pairs as outlined above.

By employing a null resumptive element in the contexts outlined above, we depart from the standard Penn scheme, which only marks a constituent as left-dislocated if it has a corresponding overt resumptive element at IP-level. However, this means that instances like (17) and (18) are not easy to isolate for researchers interested in the development of e.g. clausal integration in the language, which has attracted attention in the literature on MLG syntax (Tophinke, 2009; Breitbarth, to appear). With this innovation, we aim to make the full range of structures relevant to studies of left-dislocation and resumption easily discoverable, as ever, with no prior judgement on a particular analysis.

4.3.5. Summary of adaptations

A summary of the adaptations discussed, together with their frequency in the total texts annotated to date, is provided in Table 3. Null resumptives represent a later adaptation and are still being inserted.

Adaptation	Relevant Label	Count
Pronominal adverb	PAVAP	1639
Multi-word adposition	sister APPR/APPOs	33
Empty C (CP-THT)	C 0	71
V1 Conditional	C-V1 0	700
Subjunctive dependency	C-SUBJ 0	118

Table 3: MLG-specific adaptations and their frequency

4.4. Annotating linguistic uncertainty

A major issue which arises when annotating MLG concerns uncertainty. While uncertainty can arise with virtually any type of linguistic data, this is particularly relevant for a historical language stage like MLG. Unlike synchronic studies, in diachrony we do not have direct access to speaker knowledge, and must deduce that knowledge from the written texts available (van Kemenade and Los, 2014). Moreover, MLG poses challenges of this type in particular, since

(17)
```
(IP-MAT (CP-ADV-LFD-1 (C-V1 0)
                      (IP-SUB (VVFIN Is)
                              (NP-SBJ-2 (DPDS dat))
                              (ADVP (AVKO also))
                              (CP-THT-2 (KOUS dat) ...)
        (ADVP-RSP-1 0)
        (CP-ADV-LFD-3 (C-V1 0)
                      (IP-SUB (VVFIN steruet)
                              (NP-SBJ (DDARTA der)
                                      (DPIS eyn)
        (ADVP-RSP-3 (AVKO so)
        (VMFIN sal)
...)
```
'if it is also the case that (...), if one dies, then shall...' (Rüthen)

(18)
```
(IP-MAT (NP-LFD-1 (CP-FRL (WNP-2 (PTKG S)
                                  (DWA welich)
                                  (NA voget))
                          (IP-SUB (NP-SBJ *T*-2)
                                  (NP-OB1 (DIARTA enen)
                                          (NA richtere))
                                  (VVFIN set)
                                  (PP (APPR an)
                                      (NP (DPOSA sine)
                                          (NA stat)))))
        (NP-RSP-1 0)
        (NP-LFD-3 (CP-FRL (WNP-4 (PTKG s)
                                 (DPWS waz))
                          (IP-SUB (NP-SBJ *T*-4)
                                  (PP (APPR vor)
                                      (NP (DPDS dheme)))
                                  (VVPP gelent)
                                  (VAFIN wert))))
        (NP-SBJ-RSP-3 (DPDS dat))
        (VMFIN sal)
...)
```
'whichever representative sets a judge in his place, whatever is foeffed before that, that shall...' (Braunschweig)

the syntax of the language remains relatively under-studied. The under-researched status of MLG syntax is of course the prime motivation behind the development of the corpus and so it would be unwise to implement arbitrary decisions in instances of uncertainty, which might cloud later research and lead to unreliable or even misleading findings.

Despite the prevalence of data uncertainty in historical linguistic data, there is no standard framework in which different types of uncertainty can be captured and harnessed to provide a better understanding of the data itself. Nevertheless, in recent years the issue has begun to attract attention, and there are now some sparse resources which provide some treatment of uncertainty. One such resource which is relevant to our project is that developed by Merten and Seemann (2018) (see also Seemann et al. (2017)), a novel interface which enables the annotator to mark various types of uncertainty. This was developed specifically for the investigation of language elaboration processes in MLG (e.g. the development of complex sentence types and other complex structures). Since language elaboration is a continuous process, it involves gradualness, gradience and ambiguity, and thus various types of uncertainty. Essentially, their innovation is that annotators can assign two different POS-tags for a single lexeme, thus resulting in two different annotations. Annotators can then categorise the relevant point of uncertainty as one of three types:

1. more likely than: the lexeme is assigned two categories A and B, and category A is judged more likely ('fitting') as an analysis than category B.
2. ambiguous: used for items which are ambiguous due to context and allow for more than one possibility.
3. unsure: the annotator is unsure what the POS-tag should be due to incomplete knowledge (e.g. meeting a structure for the first time)

Of 150,000 already annotated tokens, 1,601 were annotated as uncertain in some way, and among these the most frequent type of uncertainty was 'unsure'.

Due to the nature of their overall annotation scheme – which extends to POS-tags but not to hierarchical phrase-structure – the scheme by Merten and Seemann (2018) captures uncertainty at the level of the lexeme. While they

allow for complex functional words emerging throughout the period to be grouped as a single token (e.g. the subordinator *na-dem*, 'after-that' > 'after'), they do not extend their uncertainty treatment to other more complex syntactic structures. Our Penn-style annotation throws up additional uncertainty issues which extend beyond lexical items and adjacency, and which the Penn scheme as it stands cannot capture. Here, we outline how we treat one such case of uncertainty, concerning clause type. We acknowledge that this is just one way in which linguistic ambiguity in the MLG system challenges the corpus annotation, and leave treatment of further instances for future work.

Diagnosing a clause as either main (IP-MAT) or subordinate (IP-SUB) can be problematic in MLG texts, for a number of reasons. Firstly, sentential punctuation and capitalisation are often absent, or used in a way which does not systematically distinguish main and subordinate clauses. Secondly, MLG exhibits greater variation in verb position than e.g. Present-day German, and so verb position alone cannot be used to distinguish between main versus subordinate clauses, since the latter are not consistently verb-final (Mähl, 2014). Thirdly, a number of adverbial subordinators in the language are formally identical to sentential adverbs (e.g. *also, dar*), while others are formally identical to coordinators (e.g. *wente*), see (19), taken from Härd (2000).

(19) vnde ik sach et . vnde betugede et . [**wente** dit is and I saw it and attested it WENTE this is godes sone]
god.GEN son
'and I saw it and attested it **because/that/but** this is god's son' (Buxteh. Ev.)

Thus MLG texts present clauses which cannot be handled by the Penn annotation scheme, in which every finite clause must be annotated as either IP-MAT or IP-SUB.

An additional complication is the way in which items like *wente* are treated in the HiNTs POS-tagging scheme. Word order diagnostics distinguish between three POS-tags for clause-introducing items: KON for V2, KOUS for V-later than V2, and KO* if the word order is undiagnosable, i.e. in cases where the clause contains only two constituents. The ReN team are careful to point out that the labels KON and KOUS do not necessarily reflect an analysis as coordinating or subordinating (Schröder et al., 2017); they are just a way to capture word order differences across clauses. This is another example of syntactic information being encoded in the HiNTS POS-tags.

How should clauses introduced by e.g. *wente* be annotated in CHLG, which adds a hierarchical syntactic annotation on top of the HiNTs POS-tags? The principle which underpins our annotation decisions is to consider an eventual potential user of the corpus, and the types of research questions they would be interested in. For example, it is quite plausible that a linguist interested in MLG syntax would want to investigate word order patterns across main and subordinate clauses. In such cases, a decision made at the corpus-design stage to annotate ambiguous clauses like (19) – which are relatively common in MLG texts – as either IP-MAT or IP-SUB would affect the findings of such an investigation. As such, it does not make sense to use word order diagnostics alone to operationalise between IP-MAT and IP-SUB, as this would disguise precisely the type of variation which a linguist may be interested in.

Our solution is to employ a novel label, IP-X, alongside the standard IP-MAT and IP-SUB. IP-X is employed in two different contexts where *wente* introduces a finite clause. Firstly, clauses like (19) – which are unambiguously V2 and where *wente* is tagged KON – are tagged as IP-X. Secondly, IP-X is also used for clauses introduced by *wente*, which contain only two constituents and thus are ambiguous in terms of verb placement, e.g. (20) (KO* contexts in terms of HiNTs POS-tags).

(20) ... **wente** he kam
WENTE he came
'... **because/that/but** he came'

In both cases, the IP-X label is designed to capture the uncertain status of the clause as either main or subordinate. Since the two different contexts are distinguished by the POS-tags from ReN (KON vs. KO*), it is still possible for eventual users of the corpus to isolate these from one another. Clauses introduced by *wente* which have clearly V-later order, and where *wente* is tagged as KOUS, are straightforwardly treated as IP-SUB.

Our introduction of the IP-X label thus allows us to capture the uncertain status of clauses like (19) and (20), makes such examples easy to isolate and leaves the decision as to their status open for the corpus-user to decide.

5. Conclusion

In this paper, we have presented the MLG component of the parsed Corpus of Historical Low German (CHLG), which is currently under construction. The main aim of this corpus is to make MLG texts searchable for syntactic structures that go beyond mere adjacency or co-occurrence of POS-tags within the boundaries of one clause. In particular, given the broad acceptance of the Penn annotation scheme within the historical syntax research community, the CHLG is designed to be interoperable with other historical corpora from this family. Applying the Penn-system to MLG texts required certain language-specific adaptations to the scheme, which we discussed above. In all cases, the aim was to take consistent and robust decisions, which might also serve as a model for other Penn-style treebanks. Furthermore, we hope that the first steps we have taken towards the annotation of linguistic uncertainty as discussed here will serve as a point of departure for future efforts to capture data uncertainty in diachronic language resources, which we believe would significantly benefit future corpus-based studies of syntactic change.

6. Acknowledgements

We would like to express our gratitude to the creators of the ReN, in particular Ingrid Schröder, Robert Peters, and Norbert Nagel, for granting us access to many of the texts to include in the CHLG, and for agreeing to the collaboration between the two projects. We are also indebted to Fabian Barteld, Katharina Dreessen and Sarah Ihden of the ReN-project for very helpfully making this collaboration a practical reality.

7. Bibliographical References

Barteld, F., Dreessen, K., Ihden, S., and Schröder, I. (2017). Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) – Korpusdesign, Korpuserstellung und Korpusnutzung. *Mitteilungen des Deutschen Germanistenverbandes*, 64(3):226–241.

Barteld, F., Ihden, S., Dreessen, K., and Schröder, I. (2018). HiNTS: A Tagset for Middle Low German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 3940–3945.

Breitbarth, A. (to appear). Degrees of integration: Resumption after left-peripheral conditional clauses in Middle Low German. In Karen De Clercq, et al., editors, *And then there were three. The syntax of V3 adverbial resumption in Germanic and in Romance: a comparative perspective*. Oxford University Press.

Demske, U. (2007). Das MERCURIUS-Projekt: Eine Baumbank für das Frühneuhochdeutsche. *Sprachkorpora: Datenmengen und Erkenntnisfortschritt*, pages 91–104.

Dipper, S., Donhauser, K., Klein, T., Linde, S., Müller, S., and Wegera, K.-P. (2013). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28(1):85–137.

Donhauser, K. (2015). Das Referenzkorpus Altdeutsch. Das Konzept, die Realisierung und die neuen Möglichkeiten. In Jost Gippert et al., editors, *Historical Corpora. Challenges and Perspectives*, pages 35–49. Narr.

Dreessen, K. and Ihden, S. (2015). Korpuslinguistische Studien zur mittelniederdeutschen Syntax. *Jahrbuch für Germanistische Sprachgeschichte*, 6(1):249–275.

Farasyn, M. and Breitbarth, A. (2016). Nullsubjekte im Mittelniederdeutschen. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 138(4):524–559.

Farasyn, M. (2018). *Fitting in or standing out? Subject agreement phenomena in Middle Low German*. Ph.D. thesis, Ghent University.

Härd, J. E. (2000). Syntax des Mittelniederdeutschen. In Werner Besch, et al., editors, *Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, volume 2, pages 1456–1463. de Gruyter.

Haug, D. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Mähl, S. (2004). *Studien zum mittelniederdeutschen Adverb*. Böhlau.

Mähl, S. (2014). *Mehrgliedrige Verbalkomplexe im Mittelniederdeutschen: ein Beitrag zu einer historischen Syntax des Deutschen*. Böhlau.

Merten, M.-L. and Seemann, N. (2018). Analyzing constructional change: Linguistic annotation and sources of uncertainty. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 819–825. ACM.

Merten, M.-L. (2015). Sprachausbau des Mittelniederdeutschen im Kontext rechtssprachlicher Praktiken. Konstruktionsgrammatik meets Kulturanalyse. *Niederdeutsches Jahrbuch*, 138:27–51.

Peters, R. (2015). Zur sprachgeschichte des norddeutschen raumes. In Markus Hundt et al., editors, *Deutsch im Norden*, pages 18–35. Cambridge University Press.

Petran, F., Bollmann, M., Dipper, S., and Klein, T. (2016). ReM: A reference corpus of Middle High German – corpus compilation, annotation, and access. *Journal for Language Technology and Computational Linguistic*, 31:1–15.

Petrova, S. (2012). Multiple XP-fronting in Middle Low German root clauses. *The Journal of Comparative Germanic Linguistics*, 15(2):157–188.

Pintzuk, S., Taylor, A., and Warner, A. (2017). Corpora and quantitative methods. In Adam Ledgeway et al., editors, *The Cambridge Handbook of Historical Syntax*, pages 218–240. Cambridge University Press.

Randall, B. (2005). CorpusSearch2 User's Guide. Philadelphia: Dept. of Linguistics, University of Pennsylvania. http://corpussearch.sourceforge.net.

Rösler, I. (1997). *Satz, Text, Sprachhandeln: Syntaktische Normen der mittelniederdeutschen Sprache und ihre soziofunktionalen Determinanten*. Universitätsverlag Winter.

Saltveit, L. (1970). Befehlsausdrücke in mittelniederdeutschen Bibelübersetzungen. In Dietrich Hoffmann, editor, *Gedenkschrift für William Foerste*, pages 278–289. Böhlau.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). Technical Report, University of Pennsylvania Department of Computer & Information Science.

Santorini, B. (2010). Annotation manual for the Penn Historical Corpora and the PCEEC. Department of Linguistics, University of Pennsylvania. https://www.ling.upenn.edu/hist-corpora/annotation/index.html.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Universität Stuttgart: Institut für maschinelle Sprachverarbeitung. http://www.sfs.uni-tuebingen.de/resources/stts-1999.

Schröder, I., Barteld, F., Dreessen, K., and Ihden, S. (2017). Historische Sprachdaten als Herausforderung für die manuelle und automatische Annotation: Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650). *Niederdeutsches Jahrbuch*, 140:43–57.

Seemann, N., Merten, M.-L., Geierhos, M., Tophinke, D., and Hüllermeier, E. (2017). Annotation challenges for reconstructing the structural elaboration of middle low German. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 40–45, Vancouver, Canada, August. Association for Computational Linguistics.

Taylor, A. (2020). Treebanks in historical syntax. *Annual Review of Linguistics*, 6(1):195–212.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., and Beck, K. (2006). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany*.

Tophinke, D. (2009). Vom Vorlesetext zum Lesetext: Zur Syntax mittelniederdeutscher Rechtsverordnungen im Spätmittelalter. In Angelika Linke et al., editors, *Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt*, pages 161–186. Niemeyer.

van Kemenade, A. and Los, B. (2014). Using historical texts. In Robert J. Podesva et al., editors, *Research methods in linguistics*, pages 216–232. Cambridge University Press, Cambridge.

Walkden, G. (2016). The HeliPaD: a parsed corpus of Old Saxon. *International Journal of Corpus Linguistics*, 21:559–571.

Wallmeier, N. (2015). Konditionale Adverbialsätze und konkurrierende Konstruktionen in mittelniederdeutschen Rechtstexten. *Niederdeutsches Jahrbuch*, 138:7–26.

8. Language Resource References

Albert, S., Anderssen, J., Bader, R., Becker, S., Bracht, T., Brants, S., Brants, T., Demberg, V., Dipper, S., Eisenberg, P., et al. (2003). TIGER Annotationsschema. Universität des Saarlandes and Universität Stuttgart and Universität Potsdam.

Beck, J., Ecay, A., and Ingason, A. K. (2015). Annotald. version 1.3. 7.

Galves, C., Andrade, A. L. d., and Faria, P. (2017). Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/ tycho/corpus/texts/psd.zip.

Kroch, A. and Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. Second edition. http://www.ling.upenn.edu/hist-corpora/.

Kroch, A., Santorini, B., and Delfs, L. (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. First edition. http://www.ling.upenn.edu/hist-corpora/.

Light, C. (2011). Parsed Corpus of Early New High German (Luther's Sepembertestament 1522). Version 0.5.

Martineau, F., Hirschbühler, P., Kroch, A., and Morin, Y. C. (2010). Corpus MCVF annoté syntaxiquement. Ottawa: University of Ottawa. http://www.arts.uottawa.ca/voies/corpus_pg_en.html.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

ReN-Team. (2017). Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.3. Publication date 2017-06-15. http://hdl.handle.net/11022/0000-0006-473B-9.

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). York-Toronto-Helsinki Parsed Corpus of Old English Prose. http://www-users.york.ac.uk/ lang22/YCOE/YcoeHome.htm.

Walkden, G. (2015). HeliPaD: the Heliand Parsed Database. Version 0.9. http://www.chlg. ac.uk/helipad/.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC), version 0.9. http://linguist.is/icelandic_treebank.