Combining Ontologies and Neural Networks for Analyzing Historical Language Varieties. A Case Study in Middle Low German

Maria Sukhareva **Christian Chiarcos**

{sukharev|chiarcos}@informatik.uni-frankfurt.de

Abstract

In this paper, we describe experiments on the morphosyntactic annotation of historical language varieties for the example of Middle Low German (MLG), the official language of the German Hanse during the Middle Ages and a dominant language around the Baltic Sea by the time. To our best knowledge, this is the first experiment in automatically producing morphosyntactic annotations for Middle Low German, and accordingly, no part-of-speech (POS) tagset is currently agreed upon.

In our experiment, we illustrate how ontology-based specifications of projected annotations can be employed to circumvent this issue: Instead of training and evaluating against a given tagset, we decompose it into individual features which are predicted independently by a neural network. By applying consistency constraints (axioms) from an ontology, then, the predicted feature probabilities are decoded into a sound (ontological) representation. Using these representations, we can finally bootstrap a POS tagset capturing only morphosyntactic features which could be reliably predicted. In this way, our approach is capable to optimize precision and recall of morphosyntactic annotations simultaneously with bootstrapping a tagset.

Keywords: morphosyntactic annotation, ontology-based annotation, feed-forward neural networks, Middle Low German

1. Background

For contemporary languages, most notably those from the Germania (English, German, Dutch, Swedish, etc.), a large amount of linguistic resources has been produced in the last decades, whereas historical and minor Germanic language varieties attracted attention only recently. With the increased interest in Digital Humanities and computational philology, several annotated corpora were released, e.g., for historical varieties of English (Kroch and Taylor, 2000, PCHE), Icelandic (Rögnvaldsson et al., 2012), High German (Linde and Mittmann, 2013; Light, 2013), or the oldest Germanic languages in general (Haug and Jøhndal, 2008, PROIEL). As for continental Germanic, however, it is remarkable that attention focused mostly on the oldest language strata (Old Saxon, Old High German, Gothic), whereas corpora and NLP tools for languages of the High Middle ages (esp., Middle High German and Middle Low German) are not (yet) publicly available.[1]

In this paper, we specifically focus on the morphosyntactic annotation of Middle Low German (MLG), the official language of the German Hanse during the Middle Ages and a dominant language around the Baltic Sea and in Scandinavia before its decline in the late 16th c. Despite its name, Low German is not closely related to standard (High) German, but a distinct language with its own literature and a written record of more than 1000 years (from the 9th c. *Heliand* to the rise of modern 'dialectal' literature since the 19th c.). Although we possess great amounts of historical texts for medieval (Middle) Low German, its manual annotation is extremely expensive, as it requires rare and specialized linguistic expertise. For other languages, where manually annotated resources are available, these are generally sparse, orthographically inconsistent and use different annotation schemes that hinder the direct application of the state-of-the-art statistical NLP tools (e.g., Old English as part of PROIEL and PCHE).

In order to solve the problem to adapt statistical NLP tools for contemporary languages to historical and/or low-resource varieties, both annotation projection (Tiedemann, 2014) and normalisation approaches (Pettersson et al., 2014). Yet, as far as Middle Low German is concerned, we present the first experiment in automatically producing morphosyntactic annotations for this particular language. Yet, albeit no part-of-speech (POS) tagset is currently agreed upon, information from related languages can be employed to approximate a part-of-speech (POS) tagset: Low German has a common origin with English, German and Dutch, and it evolved in close language contact with Dutch, German and Scandinavian languages, so that it shares many phonological, morphological and morphosyntactic traits with them. Accordingly, we employ the proximity to and influences from/upon these languages to bootstrap an annotation scheme for Middle Low German.

2. Approaches on Annotation Integration

Annotation projection and adaptation from multiple source languages generally suffers from inconsistencies between different annotation schemes employed for modern major languages like German, Dutch and English. Two primary approaches have been proposed to address this problem,

[1] Annotated data is to be expected from the German Reference Corpus projects. For Middle High German (http://referenzkorpus-mhd.uni-bonn.de) and Middle Low German (https://vs1.corpora.uni-hamburg.de/ren), however, no data has been released so far. Also, the Early Modern High German Reference Corpus data (http://www.ruhr-uni-bochum.de/wegera/ref) is not available yet. Independently from these efforts, a Corpus of Historical Low German is currently being created (http://www.chlg.ac.uk), but has not released annotated data either.

For Dutch, the situation is slightly better, and a POS-annotated corpus for Middle Dutch is available, albeit limited to 13th century material (http://tst-centrale.org/nl/producten/corpora/corpus-gysseling/6-59).

based on post-hoc mappings and ontology-based tagset decomposition, respectively.

2.1. Mapping-based Annotation Integration

The traditional approach to this problem is to map different existing schemes to a uniform meta tagset, as exemplified in EAGLES (Leech and Wilson, 1996), MULTEXT-East (Erjavec and Ide, 1998), and, more recently, in Petrov et al.'s original proposal for a Universal POS tagset (Petrov et al., 2012). It should be noted, though, that a post-hoc mapping of existing schemes is highly problematic as it aims to leverage independent terminological and analytical traditions developed by different communities for individual languages.

This has been acknowledged by current proposals for a universal tagset as part of the 'Universal Dependencies' (Nivre and others, 2015, UD)[2] which actively enforce reannotating existing corpora to facilitate conformance to language-independent specifications. But even Universal Tagset (further abbreviated UT) suffers from language-specific differences which – at the moment – remain unresolved (but are documented) in the definition. For example, 'adjective' is defined differently for individual languages with respect to the inclusion of participles, similarly adverbs with respect to transgressive, nouns with respect to infinitive and gerunds, verbs with respect to the classification of auxiliaries, participles, gerunds, infinitives and transgressives, particles with respect to verbal particles in German, etc. Other categories need to be redefined in an unconventional way in order to be extended to languages in which they are not grammaticalized (e.g., 'determiner' for Slavic quantifiers, numerals and attributive pronouns or 'adjective' for stative verbs in Chinese).

Three fundamental problems can be identified:

- language-specific extensions to a standardized meta-tagset are labor-intensive, if even possible (they require establishing a broad cross-linguistic consensus),
- categories (tags) in a (meta-)tagset are implicitly disjoint (thus requiring a cross-linguistic consensus on the resolution of language-specific interference phenomena between morphologic, syntactic and semantic factors involved in language-specific conceptualizations of the respective categories), and
- there are no formal means to express imprecise mappings to the meta-tagset, e.g., by defining a language-specific category as falling in the join (disjunction) of two language-indepenent meta-tags.

Since even approaches relying on manual annotation refinement did not produce consistent results yet, any attempt to *automatically* integrate annotations from different source annotations using a standardized meta-tagset will eventually lead to information loss and potential inconsistencies in the definition for the respective tags – as previously observed by Hughes et al. (1995) for English EAGLES specifications and Chiarcos and Erjavec (2011) for MULTEXT-East.

Even more severe problems arise with approaches ignorant against grammatical features involved in POS annotation. As such, Petrov et al.'s (2011) original proposal for a universal POS tagset was massively underspecified: It involved reductions from hundreds of tags (e.g., 294 tags for the Sinica/CoNLL07 treebank) to a total of of only 12 – later extended to 18.

Although this approach is still considered state of the art (Agić et al., 2015), it is extremely reductionist. It might have an application in technical applications, but without including additional specifications for morphosyntactic features (Zeman, 2008), this approach is bound to lose morphological and morphosyntactic information, making the resulting resource less valuable for studies in history, linguistics or philology. It is thus not desirable in a Digital Humanities context (as the primary locus of most research on historical language varieties). For the case of historical varieties of Germanic languages, this is even more problematic, as they tend to be morphologically richer than modern Germanic languages.

2.2. Ontology-based Annotation Integration

Ontology-based approaches, as described by Chiarcos (2010), Pareja-Lora and Aguado de Cea (2010), Hellmann et al. (2013), and Sukhareva and Chiarcos (2015) represent a promising, though understudied, alternative to mapping-based annotation integration, as they allow to circumvent the problems of mapping-based annotation integration:

- Without enforcing a fixed (and minimal) set of possible tags onto different languages, a broad set of terms can be organized in a hierarchical fashion, permitting varying degrees of granularity.
- Unlike (meta-)tagsets, this hierarchical organization is not a tree, but it employs a directed graph of subClassOf ($\sqsubseteq$) relations; it does not impose implicit disjointness constraints which need to be resolved in labor-intense discussions (e.g., a participle can be adjective and verb at the same time).
- As a representation formalism, OWL ontologies provide means to express and to constrain imprecise mappings using properties expressing different degrees of confidence and class operators like join ($\sqcup$), intersection ($\sqcap$) and negation ($\neg$) operators.

In previous work on ensemble combination (Chiarcos, 2010) and annotation integration (Sukhareva and Chiarcos, 2015) decomposed string-based annotations into individual features (attribute-value pairs, represented as RDF triples) using the *OLiA* architecture (Chiarcos, 2008):[3] Annotation schemes are modeled as independent ontologies ('Annotation Models') which are then linked to the overarching 'Reference Model' by means of declarative *rdfs:subClassOf* statements. Together, Annotation Models, Reference Model and their linking constitute the Ontologies of Linguistic Annotation (OLiA).

[2] http://universaldependencies.github.io/docs/u/pos/all.html

[3] http://purl.org/olia

As mentioned above, ontologies do not impose implicit disjointness constraints (like meta-tagsets where categories/tags are by definition disjoint), Disjointness applies only if defined. A second characteristic is that complex mappings are possible which use disjunction ($\sqcup$), intersection ($\sqcap$), and negation ($\neg$). A third characteristic is that the mapping is declarative, physically separated from the data and represented according to W3C standards with formalized semantics. On this basis, conceptual descriptions can be automatically *inferred* and *validated* using OWL2/DL reasoning capabilities. Finally, a fourth characteristic is that imprecise mapping can be represented and retrieved using alternative vocabularies to represent the linking (e.g., *skos:broader* instead of *rdfs:subClassOf*).

In terms of the first (and fourth) characteristic, an ontology-based is *lossless* in comparison to mapping-based annotation integration, in terms of the second, it is more expressive, and in terms of the thirds it introduces a clear separation between established concepts (annotation and OLiA Reference Model) and their interpretation (linking) – for which alternatives can be provided if necessary without breaking the formalism.

Using OLiA Reference Model concepts to populate output vectors of neural networks, Sukhareva and Chiarcos (2015) showed a simple feed-forward architecture can be trained to predict triple probabilities. By applying using constraints from the Reference Model, these can be decoded into ontological descriptions. In addition, they showed how such neural networks can be successfully trained against heterogeneous annotations (by means of their OLiA representations), and that these descriptions are richer than the original annotations (preserving maximum degree of granularity of the source schemes), while performance remains stable in terms of precision (in comparison to single-corpus trained neural networks) and yields state-of-the-art performance if compared with string-based POS taggers.

Here, we adopt this approach to integrate multilingual source annotations in application to a single corpus of a historical language variety. We show that by exploiting OLiA,

- there is no need to limit the annotations to coarse-grained universal POS tags,
- neural networks in combination with pre-trained word embeddings allows us to achieve results comparable to Agić et al. (2015), but with higher granularity, and finally,
- triple decomposition allows us to untangle robustly predicted features and less robustly predicted features, so that we can bootstrap a prototypical annotation scheme specifically designed for Middle Low German.

3. Prerequisites

3.1. MLG Corpus and Gold Annotation

Both annotation projection and tool adaptation require a certain amount of parallel data to be trained upon. Such approaches presuppose the existence of at least a small amount of parallel corpora. For most historical varieties of the Germanic languages, such data is available from partial translations, retellings or excerpts of the Bible (Chiarcos et al., 2014). Sukhareva and Chiarcos (2014) used such data to train fragmental dependency parsers on multilingual annotation projections to Old English, Middle Icelandic, and Early Modern High German. Later, Agić et al. (2015) used annotation projections from multiple Bibles aggregated by majority voting to train POS taggers for several target languages, including modern English, German, Danish and Icelandic.

For our MLG experiments, we used a 15th c. Gospel of John digitized by the Middle Low German Reference Corpus.[4] With only 19000 tokens (2442 verses) in total, this dataset is extremely sparse.

As the corpus did not have any POS annotation, we manually annotated 1000 tokens to have a test set for further evaluation. This annotation was produced by two annotators working jointly, using a slightly adapted version of the German standard tagset STTS (Schiller et al., 1999), further referred to as 'MLG gold'. The MLG gold annotations of both annotators were represented as OLiA triples. For training, the remaining 18000 tokens were employed.

3.2. Normalisation and Annotation of the Training Set

As we consider the amount of MLG training data barely sufficient for direct annotation projection, we opted for a combined multilingual projection-adaptation approach, instead. We employ machine translation to normalize (or, more precisely, hyperlemmatize) the original Low German to modern German, English and Dutch; the normalised train set is then annotated with the Stanford tagger (Toutanova et al., 2003) trained on modern language corpora, and these annotations are then transferred to the original MLG text.

Our task is to normalise (hyperlemmatize) every word in the test and training sets, but any word-based MT system available will lack sufficient coverage on the sparse amount of MLG training data. Accordingly, we employ character-based machine translation (CBMT), known to be particularly useful when applied to related languages as in our setting (Nakov and Tiedemann, 2012).

We combine two state-of-the-art character alignment tools (*pialign*, Neubig et al., 2012; *m2m-aligner*, Jiampojamarn et al., 2007) and one word alignment tool (*GIZA++*, Och and Ney, 2003).

For training the different modules,

1. the parallel data is word-aligned using the GIZA++ IBM-2 model to English, German and Dutch, respectively,
2. we extract 1:1 alignments by resolving multi-words alignment using lexical translation probabilities,
3. we limit the raw word list to word pairs $\langle src, tgt \rangle$ with relative Levenshtein distance[5] $\lambda(src, tgt) \leq 0.4$ (to train CBMT on etymological cognates, only), and

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

[5] I.e., Levenshtein distance divided by the length of both words, $|src| + |tgt|$.

4. we train pialign and m2m-aligner on the resulting word list.

For actual normalization, GIZA++, pialign and m2m are consulted in parallel. Be w an MLG word to be normalized to any of the target languages, w_p the pialign-predicted target word, w_m the m2m-predicted target word, $p(\cdot|w)$ the GIZA++ translation probability, and $\lambda(x, y)$ relative Levenshtein distance. The following disambiguation heuristics apply:

if $w_p = w_m$, return w_p, **else**

if $p(w_p|w) \geq p(w_m|w)$ and $p(w_p|w) > 0$, return w_p, **else**

if $p(w_m|w) > 0$, return w_m, **else**

set w'_m and w'_p to those target language words with minimal Levenshtein distance from w_m and w_p

if $\lambda(w'_p, w_p) \geq \lambda(w'_m, w_m)$, return w'_p,

else return w'_m

The normalized text is then annotated by the Stanford Tagger trained on the Penn Treebank (Marcus et al., 1993, PTB), the German NEGRA corpus (Skut et al., 1997, STTS) and the Dutch Alpino corpus (Bouma et al., 2001), respectively. These annotations are transferred to the original MLG text, these represent the sole annotation available for the training data and the basis for the experiments described below.

4. Experiment 1: Annotation Integration

In order to compare OLiA-based annotation integration with mapping-based annotation integration, we compare the results of the direct mapping from individual tagsets to OLiA triples with a mapping mediated by Petrov et al.'s (2011) original proposal for UT.

In order to put our results in relation to UT/UD-based projection results reported by Agić et al. (2015), we also provide UT-based accuracy scores in Tab. 1. UT annotation were obtained from the PTB, STTS and Alpino annotations converted to the universal tags using the mappings provided by Petrov.[6] The existing STTS mapping was adapted for MLG gold annotations. Table 1 reports UT results for the test set, with direct mapping of single-source annotations to UT, UT disambiguation using majority vote on three source annotations and UT disambiguation using MACE (Hovy et al., 2013),[7] a more elaborated expectation maximisation method originally developed for estimating the reliability of crowdsourced annotations.

The accuracy of the tags predicted by MACE is slightly better than the accuracy of the tags predicted by majority vote. Nevertheless, the combination of the three predicted tags did not result in any improvement over the best monolingual tagger. The German source annotations have considerably higher accuracy than English and Dutch. This was rather predictable because of the similarities of MLG gold and STTS. The low accuracy of Dutch annotations is due to the peculiarities of the tagset. Alpino is more coarse-grained, e.g., it does not have a separate label for pronouns and thus tags pronouns as nouns. Although Dutch has the closest ties with MLG from the languages under consideration, the overall performance using Dutch annotations is low throughout all the experiments in this paper.

DE	EN	NL	MV	MACE
0.71	0.59	0.46	0.67	0.69

Table 1: UT tagging accuracy on normalised data and their combination by majority vote (MV) and MACE

The result of the 71% tagging accuracy in Tab. 1 seems decent for a low resource language such as MLG for which no training data for supervised POS tagging is available. One may compare Agić et al. (2015) who achieved accuracy scores between 70.8% and 72.2%, albeit with *substantially* larger amounts of parallel training data.[8] But there is more to this story: The limited granularity of the UT approach renders its results insufficient for the designated users of *any* NLP tool for Middle Low German. Natural Language Processing of historical corpora should in the first place be orientated towards realistic use cases, coming in this case mostly from digital humanities, historical linguistics and philology. The POS underspecification enforced by the Universal Tagset prohibits detailed linguistic analysis of the output, but preserving more grammatical information from the original annotation schemes would lead to a deeper understanding of how the language developed and what specific grammatical characteristics it shares with related languages.

We thus estimate the information loss arising from mapping to the Universal Tagset, and it can be avoided by using OLiA. For PTB, STTS, and Alpino, we used existing OLiA Linking Models, for MLG gold, the STTS linking was adapted, and in addition, we created a mapping from OLiA to the Universal Tagset. All the tags from the source annotations and UT can thus be rendered as OLiA triples. For example, the UT tag *Verb* is mapped to the RDF triple *X rdf:type olia:Verb*. In order to guarantee a fair comparison with UT-based approaches, we only consider triples assigning word classes (subclasses of olia:MorphosyntacticCategory, roughly corresponding to UD POS tags), but ignore triples with olia:MorphosyntacticFeature assignments (roughly corresponding to morphosyntatic features in UD). As all triples considered here are thus *rdf:type* (*a*) assignments, statements are abbreviated by the OLiA Reference Model concept assigned, here *olia:Verb*. Table 2 shows the OLiA type assignments for tags assigned to a past participle by the different schemes. Only PTB, STTS and MLG gold are sufficiently fine-grained to have a distinct tag for a past participle; the Universal Tagset and Alpino do not distinguish the past participles from other verb forms.

Table 3 summarizes the result of this evaluation. The first

[6] https://github.com/slavpetrov/universal-pos-tags

[7] http://www.isi.edu/publications/licensed-sw/mace

[8] We use only 19K tokens of parallel MLG, they had entire New Testaments (∼150K tokens) or full Bibles (∼800K tokens) at their disposal, i.e., a ratio of 1:8, resp. 1:42.

tag	type assignments (triples)	tagset
VVPP	*olia:Verb, olia:NonFiniteVerb, olia:Participle, olia:PastParticiple*	MLG gold
VVPP	*olia:Verb, olia:NonFiniteVerb, olia:Participle, olia:PastParticiple*	STTS
Verb	*olia:Verb*	Alpino
VBN	*olia:Verb, olia:NonFiniteVerb, olia:Participle, olia:PastParticiple*	PTB
Verb	*olia:Verb]*	UT

Table 2: POS tags assigned to past participles and their mapping to OLiA type assignments (triples) for Universal Tagset (UT), PTB, STTS, MLG gold and Alpino

	UT			direct		
	p	r	f	p	r	f
DE	**0.78**	0.41	0.53	0.75	**0.70**	**0.72**
EN	0.64	0.32	0.42	0.63	0.48	0.54
NL	0.50	0.21	0.30	0.50	0.23	0.32
DE-EN	0.70	0.36	0.48	0.70	0.56	0.62
DE-NL	0.56	0.25	0.35	0.61	0.41	0.49
EN-NL	0.47	0.21	0.29	0.52	0.30	0.38
DE-EN-NL	0.49	0.21	0.30	0.68	0.50	0.58

Table 3: Precision, recall and f-score for OLiA triples mediated by the Universal Tagset (UT) and directly linked from the original tags (STTS, PTB, Alpino)

three rows show the evaluation for single-source (monolingually tagged) annotations mapped to OLiA triples using the intermediate UT representation (*UT* column), resp. a (*direct*mapping (*direct* column). As predicted, the recall of the universal tags for all the normalisations is significantly lower than the recall of direct mapping from the original tagsets. The only exception is the Dutch which can be explained, though, by the coarse-grained Alpino tagset.

Rows 4 to 7 show the results of the combined OLiA triple annotations. The combination of the annotations was done in a similar manner as the majority voting in Tab. 1: For each word each triple was scored by its frequency. Based on OLiA Reference Model axioms, paths were formed from the assigned triples, i.e., type assignments following the $\sqsubseteq$ axis in the ontology. Paths are scored by the sum of the triple scores divided by the length of the path. The final path would be the one with the highest score.

As in UT evaluation, none of the combined annotations outperformed the best monolingual German OLiA mapping, again reflecting the proximity of MLG gold and STTS. For all combinations, OLiA recall exceeds UT recall, and for most, OLiA precision exceeds UT precision, as well. The key result is thus that direct OLiA mapping preserves information more reliably than possible if mediated by UT tags. We conduct another experiment in order to assess the reproducibility of OLiA annotations and describe the subsequent bootstrapping of an MLG-specific tagset.

5. Experiment 2: Annotating MLG

For annotating Middle Low German, we train a neural classifier to predict the entire range of OLiA triples found in the training data simultaneously. The output layer is decoded using OLiA, and subsequently used for bootstrapping an annotation scheme for MLG.

5.1. Configuring and Encoding the Neural Network

Following Sukhareva and Chiarcos (2015), we employ a simple feed-forward backpropagation neural network:

1. Input neurons that correspond to the concatenated word embeddings of the word under investigation, its predecessor and its successor.
2. One hidden layer with the tanh activation function. The number of neurons in the hidden layer is heuristically set to the average length of input and output layers, thus, a natural geometric (pyramidal) design.
3. A layer of output neurons that represent OLiA `MorphosyntacticCategory`s, again with tanh normalization. The activations of these neurons represent the output vector.

As input vectors for the neural networks, we adopt pretrained 100-dimensional German word embeddings,[9] which are assigned to Middle Low German words through their German normalization. We concatenate the embeddings of the preceding word, the actual word and the following word, the length of the input layer is thus 300 neurons. The length of the output layer varies depending on the number of the OLiA triples produced by the respective annotation schemes. Reflecting differences in granularity, the monolingual network trained on Dutch Alpino annotations had the least number of output neurons (57) while the output layer of the neural network trained on the trilingual data had the maximum length of 68 neurons.
The output layer was encoded as follows: Every possible attribute-value combination (e.g., an RDF triple X *rdf:type olia:Noun* for a token X) corresponds to a node in the output layer. If the ontological representation of a tag – say

[9] `https://www.ukp.tu-darmstadt.de/research/ukp-in-challenges/germeval-2014`

	p	r	f	g
DE	0.79	0.65	0.71	59
EN	0.75	0.42	0.54	60
NL	0.73	0.20	0.31	57
DE-EN-NL	0.80	0.63	0.71	68

Table 4: Performance of an ontology-based neural network on Middle Low German in terms of precision, recall, f-score and granularity (i.e., the number of different triple types/MorphosyntacticCategory classes predicted)

NN in German STTS – includes the corresponding triple, the value of this neuron is set to 1, otherwise the value is set to 0.

We experimented with three single-source (monolingual) networks and one multi-source (trilingual, DE-EN-NL) network. For training the trilingual network, we conducted majority vote over the three monolingual OLiA annotations (cf. Sect. 4.) and used the resulting set of triples to populate the output layer. Though the trilingual combination of the OLiA annotations is inferior to the German OLiA annotations in terms of precision and recall, the trilingual network can generate maximally fine-grained annotations as it is capable to preserve/reproduce all OLiA triples that could possibly be produced by any of the monolingual schemes.

While analysing the results of the first experiment, we observed that the straightforward majority voting over the triples was problematic in cases when an annotation scheme has triples that are not covered by other schemes. For example, the Alpino annotation scheme does not make a distinction between *ProperNoun* and *CommonNoun* while both PTB and STTS schemes do. In the cases when the same word was tagged as a proper noun through one of the schemes and as a common noun through another one, the majority voting would just choose a shorter path on which most of the taggers agree and in this particular example the word would be annotated as a *Noun* underspecifying whether it is a common noun or a proper noun. To overcome this problem, the value of the output neurons which correspond to triples that could potentially prolong the given path were set to 0.5.

In total, three monolingual and one trilingual network were trained in 100 iterations. Training was conducted on 18000 words, with German embeddings assigned via the normalization to German, and annotations for German, English and Dutch, respectively.

5.2. Decoding and Evaluating the Neural Network

For decoding neural networks, we employ the best-performing pruning strategy reported by Sukhareva and Chiarcos (2015), i.e., returning the maximally probable path (a sequence of OLiA subclasses along the $\sqsubseteq$ axis) applicable to the word under consideration.

Table 4 shows precision, recall and f-score for the paths extracted predicted in this way by three monolingual and one trilingual neural networks. Remarkably, the f-scores of the monolingual neural networks are comparable with the scores obtained in the first experiment by directly mapping POS tags to OLiA representations (Tab. 3). All monolingual neural networks produce the output with significantly higher precision than the direct mapping, albeit with a drop in recall.

The monolingual results also confirm the findings from experiment 1 in that the performance varies depending on the linguistic similarity between the languages and tagsets. The best results in terms of precision are achieved on the German normalisation. The most obvious explanation is that MLG gold was particularly close to German STTS, and that German-based annotations thus had an a-priori advantage in high overlap with the MLG gold triples and as well as the inevitable similarity of the linguistic analysis. As observed before, the Dutch recall is low because of underspecifications and differences in the linguistic classification employed in Alpino. The English-based annotation is richer than Alpino tags but not as rich as STTS. Naturally, PTB annotations also differ in linguistic analysis and their OLiA representations have less overlap with MLG gold triples. Thus, the recall drop is not surprising for the English normalisation. An additional factor may be that English is linguistically more remote from continental Western Germanic languages due to its heavy influence from French.

The performance of the trilingual network is more interesting, as it outperformed the trilingual combination of the direct mapping by a high margin: Almost 0.13 increase of the f-score, and the best precision out of all the settings. This is an important finding showing the potential of the combination of neural networks and ontologies for encoding and decoding them.

The last column in Tab. 4 shows that OLiA is capable of preserving and accumulating source-specific differentiations, thereby leading to a higher granularity than any of the source tagsets. The ontology-based approach thus releaves us from resorting to coarse-grained universal POS tags. Yet, ontology-based annotations are still a step away from conventional tagsets. The following section thus describes how a tagset and string-based annotations can be bootstrapped from the results of experiment 2.

6. Towards an Annotation Scheme for MLG

Ontology-based morphosyntactic annotation provides us with the opportunity to assess the performance of a tagger on triples (i.e., concepts and features) rather than on opaque tags. This actually allows us to reverse the classical process of tagset development. Rather than starting with a theoretically motivated model and iterative cycles of tagger training and tag set refinement, we start with a feature-based classifier and propose aggregate tags based on the features successfully recognized. In this way, we reduce the need for iterative cycles of tagger training and tagset optimization.

We illustrate this for our MLG data and bootstrap a tagset by mapping successfully classified features to tags. As measurement of successful classification, we apply the f-score for individual triples, requiring that every prospective tag has a well-defined minimal f-score. In addition, we employ a threshold operating on paths rather than triples to guarantee consistent results.

For decoding the neural network annotations of a given MLG word, we experimented with a pruning strategy operating on the activation of triples. The general tendency is

(minimal) path selection threshold	minimal f-score (per triple) used for triple selection: trilingual NN 0.40	trilingual NN 0.60	trilingual NN 0.80	German NN 0.40	German NN 0.60	German NN 0.80
0.00	29	20	9	28	20	7
0.05	31	20	9	28	21	6
0.10	32	20	9	28	21	7
0.15	32	20	9	28	21	7
0.20	32	20	9	28	21	7
0.25	32	20	9	28	21	7
0.30	31	20	9	28	21	7
0.35	31	20	9	28	21	7
0.40	31	20	9	29	21	7
0.45	31	20	9	29	21	7
0.50	31	20	9	29	21	7
0.55	31	20	9	29	21	7
0.60	31	20	9	29	21	7
0.65	31	20	9	29	21	7
0.70	31	20	9	29	21	7
0.75	31	20	9	29	21	7
0.80	31	20	9	29	21	7
0.85	29	20	9	29	21	7
0.90	29	20	9	29	21	7
0.95	30	20	9	27	21	7

Table 5: Selecting high-accuracy triples using path selection threshold and triple f-score threshold, trilingual and German-only neural networks

that superclass triples also have a high score if the triple is scored high by the neural network. Deviations from this pattern are taken as an indication of spurious triples, as are triples with scores close to 0. Thus, we tested whether the system could benefit from admitting only paths whose triples' activation exceedes a minimal path selection threshold.

Table 5 illustrates the interaction of the path selection threshold and the triple f-score in terms of the triples that can still be distinguished, again limited to instances of MorphosyntacticCategory (and ignoring MorphosyntacticFeature triples). In general, the path selection threshold does not seem to have a strong impact, indicating that predictions by the neural network are generally compliant with the path. While the numbers indicate that the trilingual NN produces more (reliable) triples than the German NN, this is not generally the case (e.g., for triple f-score 0.6). However, our evaluation is biased towards features from the German annotation as we use the German-based MLG gold annotations for bootstrapping the annotation scheme. Indeed, using (OLiA representations of) projected annotations would be an interesting alternative to be further explored in the future. Although this will add projection noise, thus increase the error rate and thus lead to less reliable scores, the bootstrapped tagset is likely to reflect the increase in granularity obtained from multilingual training.

For bootstrapping an annotation scheme, we chose a configuration with path selection score 0.10 and triple f-score 0.40 from the trilingual NN. As subclasses of MorphosyntacticCategory are organized in a hierarchical fashion, we can directly map paths to a positional tagset: Top-level categories are represented by the first two characters of the tag, their children by the third character, etc.

The following simplifications apply:

- If an OLiA category is assigned exactly as many individuals as one of its subclasses, this class is skipped.
- If a category contains subclasses, but not all of its instances are assigned to one of these subclasses, represent these unassigned individuals by this class.
- PronounOrDeterminer is skipped in favor of its subclasses Pronoun and Determiner.
- In order to enforce a tree structure, AttributivePronoun (i.e., the intersection of Determiner and Pronoun) is excluded from Pronoun and classified as Determiner.

In this way, we arrive at a tagset with 26 tags (reflecting 32 different triples):

```
AV    Adverb
CO    Conjunction
COc   CoordinatingConjunction
COs   SubordinatingConjunction
DT    Determiner (incl. AttributivePronoun)
DTa   (definite) Article
DTd   DemonstrativeDeterminer
DTi   IndefiniteDeterminer
DTp   PossessiveDeterminer
NN    Noun
NNc   CommonNoun
NNp   ProperNoun
NU    Numeral (skipped: Quantifier)
NUc   CardinalNumber
PN    Pronoun (skipped: PronounOrDeterminer;
      excl. AttributivePronoun)
PN$   PossessivePronoun
PNp   PersReflPronoun
PNpp  PersonalPronoun
PP    Preposition (skipped: Adposition)
PU    SentenceFinalPunctuation (skipped: Punctuation,
      MainPunctuation)
VE    Verb
VEf   FiniteVerb (i.e., finite main verb)
VEm   ModalVerb (skipped: AuxiliaryVerb)
VEn   NonFiniteVerb
VEnp  Participle
O     other MorphosyntacticCategory
```

While this is certainly preliminary, it represents a first assessment of the morphosyntactic differentiations that are to be expected for MLG, and more fine-grained than UD or UT POS tags, as illustrated in the example below:

MLG text	English	MLG gold	MLG tag	OLiA path
Des	the	D	DTa	Determiner, Article
anderen	other	AA	O	(MorphosyntacticCategory)
dages	day	NN	NNc	Noun, CommonNoun
sach	saw	VVFIN	VE	Verb
johannes	John	NE	NNp	Noun, ProperNoun
ihesum	Jesus	NE	NNp	Noun, ProperNoun
to	to	PR	PP	Adposition, Preposition
sick	him(self)	REF	PNp	Pronoun, PersReflPronoun
komende	coming	VVPR	VEn	Verb, NonFiniteVerb

The other day, John saw Jesus coming to him.

It should be noted that although this tagset is still less fine-grained than German STTS or MLG gold, we can guarantee a minimal f-score for every individual tag, and thus, a high (and balanced) degree of accuracy for the tagset in its entirety. Yet, this comes at a price, and certain categories are not assigned individual tags. From the 'canonical' tags, this includes adjectives whose differentiation from Participles is

resolved differently in PTB and STTS. This adds to noise in the OLiA representation, so that these could not be extracted with a sufficiently reliable f-score from the trilingual NN, thus rendered as "other MorphosyntacticCategory". Similarly, reflexive pronouns are tagged as (other) PersReflPronoun (PNp) only, and thus identified by their superconcept rather by a designated class. As for another apparent gap in the annotation scheme, one may wonder about the concept StrictAuxiliary ($\sqsubset$ AuxiliaryVerb), a sibling concept to ModalVerb. Although it can be easily recognized on its surface forms *hebben* 'to have' and *sint/wesen* 'to be', it is not extractable from STTS, PTB or Alpino: Alpino doesn't subclassify Verb at all, whereas STTS and PTB define the corresponding class as being determined by its lexical form rather than the syntactic context. Accordingly, the OLiA linking represents these verbs with a disjunction between of (strict) auxiliary verb and finite/nonfinite (main) verb. This representation is lossless, but as it cannot be resolved into a distinct path (i.e., a *conjunction* of terms along the $\sqsubseteq$ axis), its resolution depends on the availability of a source annotation with a non-disjunctive representation. Such an annotation is available, for example, from syntactic parses, yet, beyond scope of our current experiment. Alternatively, this represents a prime example illustrating the potential of manual refinement of automated annotations.

Indeed, this preliminary tagset – being relatively reliable but underspecified – represents a basis to generate additional training data in a semiautomated fashion, e.g., focusing on the subclassification of O (other MorphosyntacticCategory) tags in an Active Learning fashion. On this data, the process of tagset extrapolation can then be reiterated until we arrive at a sufficiently refined model. It should be noted, though, that this possible reiteration differs fundamentally from classical tagset design as it is not necessary to perform any reannotation of the original gold data due to tag set revision. Instead, we are able to employ target annotations of different granularity – if represented as OLiA triples.

7. Summary and Discussion

This paper described experiments on the morphosyntactic annotation of Middle Low German (MLG) operating on CBMT-based normalization to three modern related languages in order to apply existing POS taggers to the historical language variety.

With three divergent annotations now available for a total of 19000 MLG tokens, we studied the potential of ontology-based annotation integration using the OLiA ontologies. Experiment 1 compared ontology-based annotation integration with a state-of-the-art approach based on mapping to 'universal' POS tags (UT). We showed that our normalization pipeline, applied to UT, yields results comparable to those of Agić et al. (2015), but also, that the UT approach leads to considerable loss of information that can be preserved and reliably extracted from the source annotations.

Experiment 2 then employed neural networks trained on monolingual and trilingual annotations for our MLG training corpus. Remarkably, precision and recall remained stable as compared to the setting in experiment 1, the trilingual network even outperformed the trilingual ontology-based annotation integration. We found that MLG annotations can be most reliably predicted using either a German monolingual neural network or the trilingual network, which performed at the same level. Obviously, the neural network was able to outperform the simple majority vote performed in experiment 1.

Finally, we described how a rudimentary annotation scheme for MLG can be bootstrapped using this neural architecture. It should be noted that this approach on establishing a tagset for a novel language reverses the process normally applied for the purpose: Instead of training and evaluating against a given tagset, we decomponse it into individual features which are predicted independently by a neural network. By applying consistency constraints (axioms) from an ontology, then, the predicted feature probabilities are decoded into a sound (ontological) representation. Using these representations, we can finally bootstrap a POS tagset capturing only morphosyntactic features which could be reliably predicted. In this way, our approach is capable to optimize precision and recall of morphosyntactic annotations simultaneously with bootstrapping a tagset.

It is insightful to put our results in relation to those of Agić et al. (2015). Our normalization-based annotation pipeline performs comparable to their projection-based approach in terms of UT/UD tag prediction. At the same time, we used only a fraction of the training data available to them. In that sense, our results are surprisingly good, and this can only be explained by the linguistic proximity of the languages involved which allowed us to use a normalization-based annotation pipeline on the basis of character-based machine translation. An important difference, however, is that the granularity of morphosyntactic analyses predicted by our system is substantially higher than the 18 tags distinguished in the UD version of the universal tagset they employed, so that our results are more informative: While we meet the state of the art, the granularity of our analyses is higher as it is *non-reductionist* and yields an ontologically sound result.

Our experiments primarily illustrated the potential to generalize over heterogeneous annotations using neural networks in combination with the Ontologies of Linguistic Annotation. Along the way, we developed the first POS tagger for Middle Low German.

8. References

Agić, v., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China, July. Association for Computational Linguistics.

Bouma, G., Van Noord, G., and Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

Chiarcos, C. and Erjavec, T. (2011). Owl/dl formalization of the multext-east morphosyntactic specifications. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 11–20. Association for Computational Linguistics.

Chiarcos, C., Sukhareva, M., Mittmann, R., Price, T., Detmold, G., and Chobotsky, J. (2014). New technologies for old germanic. resources and research on parallel bibles in older continental western germanic. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 22–31, Gothenburg, Sweden, April. Association for Computational Linguistics.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Chiarcos, C. (2010). Towards robust multi-tool tagging. An OWL/DL-based approach. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 659–670, Uppsala, Sweden.

Erjavec, T. and Ide, N. (1998). The MULTEXT-East Corpus. In *Proc. of LREC-1998*.

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

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating nlp using linked data. In *The Semantic Web–ISWC 2013*, pages 98–113. Springer.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June. Association for Computational Linguistics.

Hughes, J., Souter, D., and Atwell, E. (1995). Automatic extraction of tagset mappings from parallel annotated corpora. In *Proceedings of the ACL-SIGDAT Workshop From Text to Tags: Issues in Multilingual Language Analysis*, pages 10 – 17, Dublin, Ireland.

Jiampojamarn, S., Kondrak, G., and Sherif, T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Kroch, A. and Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM.

Leech, G. and Wilson, A. (1996). EAGLES guidelines: Recommendations for the morphosyntactic annotation of corpora.

Light, C. (2013). Parsed Corpus of Early New High German (PCENHG), v. 0.5. University of Pennsylvania, `http://enhgcorpus.wikispaces.com/`.

Linde, S. and Mittmann, R. (2013). Old German Reference Corpus. Digitizing the knowledge of the 19th century. In Bennett, P., Durrell, M., Scheible, S., and Whitt, R. J., editors, *New Methods in Historical Corpus Linguistics = Korpuslinguistik und interdiziplinre Perspektiven auf Sprache – Corpus linguistics and Interdisciplinary perspectives on language (CLIP)*, volume 3 of *Korpuslinguistik und interdiziplinre Perspektiven auf Sprache – Corpus linguistics and Interdisciplinary perspectives on language (CLIP)*, Tübingen. Narr.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *ACL (2)*, pages 301–305. The Association for Computer Linguistics.

Neubig, G., Watanabe, T., Mori, S., and Kawahara, T. (2012). Machine translation without words through substring alignment. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2012)*, pages 165–174, Jeju Island, Korea, July.

Nivre, J. et al. (2015). Universal dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pareja-Lora, A. and Aguado de Cea, G. (2010). Ontology-based interoperation of linguistic tools for an improved lemma annotation in Spanish. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC-2010)*, Valetta, Malta, May.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey.

Pettersson, E., Megyesi, B., and Nivre, J. (2014). A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden, April. Association for Computational Linguistics.

Rögnvaldsson, E., Ingason, A. K., Sigurdsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, May.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universitäten Stuttgart und Tübingen.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 88–95.

Sukhareva, M. and Chiarcos, C. (2014). Diachronic proximity vs. data sparsity in cross-lingual parser projection. a case study on germanic. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 11–20, Dublin, Ireland, August. Association for Computational Linguistics and

Dublin City University.

Sukhareva, M. and Chiarcos, C. (2015). An ontology-based approach to automatic part-of-speech tagging using heterogeneously annotated corpora. In *Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data*, Hissar, Bulgaria, September.

Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.