Diachronic proximity vs. data sparsity in cross-lingual parser projection: a case study on Germanic

Maria Sukhareva, Christian Chiarcos

Angaben zur Veröffentlichung / Publication details:

Sukhareva, Maria, and Christian Chiarcos. 2014. "Diachronic proximity vs. data sparsity in cross-lingual parser projection: a case study on Germanic." In *Proceedings of VarDial: the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, August 23, 2014, Dublin, Ireland*, edited by Marcos Zampier, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, 11–20. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-5302.

Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic

Maria Sukhareva
Goethe University Frankfurt
sukharev@em.uni-frankfurt.de

Christian Chiarcos
Goethe University Frankfurt
chiarcos@em.uni-frankfurt.de

Abstract

For the study of historical language varieties, the sparsity of training data imposes immense problems on syntactic annotation and the development of NLP tools that automatize the process. In this paper, we explore strategies to compensate the lack of training data by including data from related varieties in a series of annotation projection experiments from English to four old Germanic languages: On dependency syntax projected from English to one or multiple language(s), we train a fragment-aware parser trained and apply it to the target language. For parser training, we consider small datasets from the target language as a baseline, and compare it with models trained on larger datasets from multiple varieties with different degrees of relatedness, thereby balancing sparsity and diachronic proximity.

Our experiments show

(a) that including related language data to training data in the target language can improve parsing performance,

(b) that a parser trained on data from two related languages (and none from the target language) can reach a performance that is statistically not significantly worse than that of a parser trained on the projections to the target language, and

(c) that both conclusions holds only among the three most closely related languages under consideration, but not necessarily the fourth.

The experiments motivate the compilation of a larger parallel corpus of historical Germanic varieties as a basis for subsequent studies.

1 Background and motivation

We describe an experiment on annotation projection (Yarowski and Ngai, 2001) between different Germanic languages, resp., their historical varieties, with the goal to assess to what extent sparsity of parallel data can be compensated by material from varieties related to the target variety, and studying the impact of diachronic proximity onto such applications.

Statistical NLP of historical language data involves general issues typical for low-resource languages (the lack of annotated corpora, data sparsity, etc.), but also very specific challenges such as lack of standardized orthography, unsystematized punctuation, and a considerable degree of morphological variation. At the same time, historical languages can be viewed as variants of their modern descendants rather than entirely independent languages, a situation comparable to low-resource languages for which a diachronically related major language exists. Technologies for the cross-lingual adaptation of NLP tools or training of NLP tools on multiple dialects or language stages are thus of practical relevance to not only historical linguistics, but also to modern low-resource languages.

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 11–20,
Dublin, Ireland, August 23 2014.

in this context, historical language allows to study the impact of the parameter of *diachronic relatedness*, as it can be adjusted relatively freely, e.g., by choosing dialects which common ancestor existed just a few generations before rather than languages separated for centuries. A focused study of the impact of diachronic relatedness on projected annotations requires sufficient amounts of parallel texts for major language stages, and comparable annotations as a gold standard for evaluation. In this regard, the Germanic languages provide us with a especially promising sandbox to develop such algorithms due to the abundance of annotated corpora and NLP tools of the modern Germanic languages, most noteably Modern English.

We employ annotation projection from EN to Middle English (ME), Old English (OE) and the less closely related Early Modern High German (DE) and Middle Icelandic (IS) for which we possess comparable annotations, and test the following hypotheses:

(H1) Adding data from related varieties **compensates the sparsity** of target language training data.

(H2) Data from related languages **compensates the lack** of target language training data.

(H3) The greater the **diachronic proximity**, the better the performance of (H1) and (H2).

We test these hypotheses in the following setup: (1) *Hyperlemmatization*: Different historical variants are normalized to a consistent standard, e.g., represented by a modern language (Bollmann et al., 2011). We emulate hyperlemmatization by English glosses automatically obtained through SMT. (2) *Projection*: We create training data for a fragment-aware dependency parser (Spreyer et al., 2010) using annotation projection from modern English. (3) *Combination and evaluation*: Parser modules are trained on different training data sets, and evaluated against existing gold annotations.

In our setting, we enforce data sparsity by using deliberately small training data sets. This is because we emulate the situation of less-documented languages that will be in the focus of subsequent experiments, namely, Old High German and Old Saxon, which are relatively poorly documented. We do hope, however, that scalable NLP solutions can be developed if we add background information from their descendants (Middle/Early Modern High German, Middle/Modern Low German), or closely related, and better documented varieties (Old English, Middle Dutch).

Hence, the goal of our experiment is not to develop state-of-the-art parsers, but to detect statistically significant differences in parsing performance. If these can be confirmed, this motivates creating a larger corpus of parallel texts in Germanic languages as a basis for subsequent studies and more advanced, projection-based technologies for older and under-resourced Germanic languages.

2 Languages and corpus data

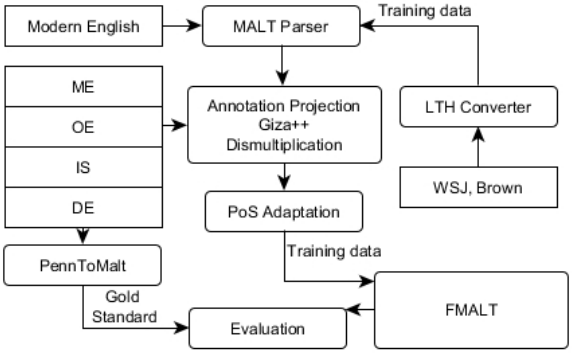
We use parallel biblical texts in Old English (OE), Middle English (ME), Middle Icelandic (IS) and Early Modern High German (DE). This selection is determined by the availability of syntactically annotated corpora with closely related annotation schemes. As these schemes are derived from the Penn TreeBank (PTB) bracketing guidelines (Taylor et al., 2003a), we decided to use Modern English (EN) as a source for the projections.

The Germanic languages derive from Proto-Germanic as a common ancestor. OE and Old High German separated in the 5th c. The antecessor of IS separated from this branch about 500 years earlier. Among Germanic languages, great differences emerged, but most languages developed similarly towards a loss of morphology and a more rigid syntax, a tendency particularly prevalent in EN.

As compared to this, OE had a relatively free OV word order, with grammatical roles conveyed through morphological markers. The OE case marking system distinguished four cases, but eventually collapsed during ME, resulting in a strict strict VO word order in EN (Trips, 2002; van Kemenade and Los, 2009; Cummings, 2010).

Unlike EN, DE preserved four cases, and a relatively free word order (Ebert, 1976). A characteristic of German are separable verb prefixes, leading to $1 : n$ mappings in the statistical alignment with EN.

Figure 1: Workflow



Unlike EN and DE, IS is a North Germanic language. It is assumed to be conservative, with relatively free word order with both OV and VO patterns and a rich morphology that leads to many $1:n$ alignments with EN, e.g., for suffixed definite articles; we thus expect special challenges for annotation projection under conditions with limited training data.

Different from the old languages, EN developed a rigid word order and a largely reduced morphology. A direct adaptation of an existing English parser to (hyperlemmatized) OE, IS or DE is thus not promising. Therefore, we employ an approach based on annotation projection.

The corpus data we used consists of parsed bible fragments from manually annotated corpora, mostly the gospels of Matthew (Mt), Mark (Mr), John (J) and Luke (L), from which we drew a test set of 147 sentences and a training set of 437 sentences for every language.

ME and OE The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)[1] and the York-Toronto-Helsinki Parsed Corpus of Old English Prose (Taylor et al., 2003b, YCOE) use a variant of the PTB annotation schema (Taylor et al., 2003a). YCOE contains the full West Saxon Gospel, but PPCME2 contains only a small fragment of a Wycliffite gospel of John, the ME data is thus complemented with parts of Genesis (G) and Numbers (N).

IS The Icelandic Parsed Historical Corpus (Rögnvaldsson et al., 2012, IcePaHC) is annotated following YCOE with slight modifications for specifics of IS. We use the gospel of John from Oddur Gottskálksson's New Testament, a direct translation from Luther.

DE The Parsed Corpus of Early New High German[2] contains three gospels from Luther's Septembertestament (1522). As an IcePaHC side-project, it adapts the IS annotation scheme.

EN For EN, we use the ESV Bible.[3] Due to a moderate number of archaisms, it is particularly well-suited for automated annotation.

3 Experimental setup

We study the projection of *dependency syntax*, as it is considered particularly suitable for free word-order languages like IS, OE and DE. The existing constituent annotations were thus converted with standard tools for PTB conversion. Figure 1 summarizes the experimental setup.

For **annotating EN**, we created dependency versions of WSJ and Brown sections of the PTB with the LTH Converter (Johansson and Nugues, 2007). We trained Malt 1.7.2 (Nivre, 2003), optimized its features with MaltOptimizer (Ballesteros and Nivre, 2012), and parsed the EN bible using the resulting feature model.

[1] http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html
[2] http://enhgcorpus.wikispaces.com
[3] http://esv.org

The ME, OE, DE and IS datasets were **word aligned** with EN using GIZA++ (Och and Ney, 2003). $1:n$ alignments were resolved to the most probable $1:1$ mapping. During **annotation projection**, we assume that the aligned words represent the respective heads for the remaining $n-1$ words. These dependent words are assigned the dependency relation `FRAG` to the word that got the highest score in the translation table. This solution solves, among others, the problem of separable verb prefixes in DE, for example, DE *ruffen* with prefix *an* would be aligned to English word *call*: As $P(\text{"}call\text{"}|\text{"}an\text{"}) < P(\text{"}call\text{"}|\text{"}ruffen\text{"})$, the syntactic information of *"call"* will be projected to *"ruffen"* and *"an"* will be its dependent labeled with "FRAG". The projected dependency trees were checked on well-formedness, sentences with cycles were dismissed from the data set.

We formed **training sets** containing 437 sentences for ME, OE, DE, IS. Monolingual data sets were combined into bi-, tri- or quadrilingual training data sets with a simple concatenation, thereby creating less sparse, but more heterogeneous training data sets. For every language, **test data** was taken from J, 174 sentences per language.

We used the projected dependencies to train fMalt (Spreyer et al., 2010), a fragment-aware dependency parser, in order to maximize the gain of information from incomplete projections.

In our setting, fMalt used two features, POS and hyperlemmas.

POS The tagsets of the historical corpora originate in PTB, but show incompatible adaptations to the native morphosyntax. Tagset extensions on grammatical case in OE, IS and DE were removed and language-specific extensions for auxiliaries and modal verbs were leveled, in favor of a common, but underspecified tagset for all four languages. As these generalized tags preserve information not found in EN, they were fed into the parser.

(hyper-)lemma Lexicalization is utterly important for the dependency parsing (Kawahara and Uchimoto, 2007), but to generalize over specifics of historical language varieties, hyperlemmatization needs to be performed. Similar to Zeman and Resnik (2008), we use projected English words as hyperlemmas and feed them into the parser. Hyperlemmatization against a closely related languages is acceptable as we can expect that the syntactic properties of words are likely to be similar.

The projected annotations were then **evaluated** against dependency annotations created analoguously to the EN annotations from manual PTB-style constituency syntax. As LTH works exclusively on PTB data, the historical corpora were converted with its antecessor Penn2Malt[4] using user-defined head-rules (Yamada and Matsumoto, 2003).

4 Evaluation results

	baseline	ΔUAS worst model				ΔUAS best model				ΔUAS
	UAS	+1		+2		+1		+2		+3
ME	.60	+DE	+.00$^{n.s.}$	+DE+IS	-.01$^{n.s.}$	+OE	+.01$^{n.s.}$	+OE+IS	+.01$^{n.s.}$	-.00$^{n.s.}$
OE	.31	+IS	-.00$^{n.s.}$	+DE+IS	-.02$^{n.s.}$	+DE	+.02$^{n.s.}$	+ME+DE	+.00$^{n.s.}$	+.02$^{n.s.}$
DE	.41	+OE	+.02$^{n.s.}$	+OE+IS	+.03*	+ME	+.04***	+ME+IS	+.03*	+.04**
IS	.32	+IS	-.02$^{n.s.}$	+DE+OE	-.02$^{n.s.}$	+ME	+.00$^{n.s.}$	+ME+DE	-.01$^{n.s.}$	-.04**

(a) trained on **target and** related language(s)

	baseline	ΔUAS worst model				ΔUAS best model				ΔUAS
	UAS	1		2		1		2		3
ME	.60	OE	-.09***	DE-IS	-.01$^{n.s.}$	IS	-.05***	IS+OE	-.02$^{n.s.}$	-.02$^{n.s.}$
OE	.31	DE	-.03*	ME-DE	-.01$^{n.s.}$	ME	-.02$^{n.s.}$	ME+IS	-.01$^{n.s.}$	-.00$^{n.s.}$
DE	.41	OE	-.01$^{n.s.}$	OE-IS	+.02$^{n.s.}$	IS	+.02$^{n.s.}$	IS+ME	+.05***	+.04**
IS	.32	OE	-.07***	DE-OE	-.02$^{n.s.}$	ME	-.06***	ME+DE	-.02$^{n.s.}$	-.04**

(b) trained on related language(s) **alone**

Table 1: Performance of best- and worst-performing parsing models (UAS diff. vs. baseline with χ^2: * $p < .05$, ** $p < .01$, *** $p < .005$)

We evaluate the unlabeled attachment score (Collins et al., 1999, UAS), i.e., the proportion of tokens in a sentence (without punctuation) that are assigned the correct head, on test sets of 174 sentences in each language.

[4]http://stp.lingfil.uu.se/ nivre/research/Penn2Malt.html

As a **baseline** for the evaluation we take the performance of the parser trained solely on the target language data. As shown in Tab. 1 (second col.), the UAS scores mirror both the diachronic relatedness (ME>DE>IS), as well as the relative loss of morphology (ME>DE>IS/OE), indicating that diachronic relatedness may not be the only factor licensing the applicability of the annotation projection scenario (H3). It is also important, though, to keep in mind that the OE and IS translations of the Bible had considerable influence of Latin syntax, whereas DE and ME translations aimed for a language easy to understand.

Table 1a gives the best and worst results for the unlabeled attachment score for the parser trained on target and related language(s) **(H1)**. With the exception of DE, we observed no significant differences in UAS scores relative to the baseline. DE may benefit from ME because of its more flexible syntax (thus closer to ME [and OE] than to Modern English), and from IS because of Luther's direct influence on the IS bible. That ME did not mutually benefit from German may be due to the good quality of ME annotation projections (resulting from its proximity to EN). Parsers trained on trilingual and quadrilingual sets exhibited no improvement over the bilingual sets. Taken together, we found *no positive effect* of using additional training data from language stages diachronically separated for more than 500 years (e.g., OE/ME), but also, we did *not find* a negative effect among the West Germanic languages. If additional training material is carefully chosen among particularly closely related varieties, however, the DE effect can be replicated, and then, including related language data to training data in the target language can improve parsing performance.

While in our setting, training data from related languages may (but does not have to) improve a parser training if training data for the target language is available, it may very well be employed fruitfully *if no training data for the target language is available* **(H2)**: Table 1b shows that, unsurprisingly, parsers trained only on one related language had the lowest performance in the experiment, so using multiple train languages seems to compensate language-specific idiosyncrasies. The best-performing parsing models trained on *two* or more related languages achieved a performance not significantly worse (if not better) than models being trained on target language data. This effect extends to all languages except for IS and indicates that a *careful choice* of additional training data from related varieties may facilitate annotation projection. Equally important (and valid across all languages) is that *none* of the models trained on one language outperformed any of the model trained on two languages. Using training data from two related languages doesn't seem to hurt performance in our setting. Adding a third language did not yield systematic improvements, the scores for trilingual models are in the range of the bilingual models.

Again, DE is exceptionally good, benefitting from being a direct source of the IS translation as well as structurally comparable to ME. In both settings, the worst-performing language is IS, with a significant drop in annotation projection quality with Western Germanic material added, indicating that diachronic distance between Northern and Western Germanic languages limits the applicability of **(H2)**, thereby supporting **(H3)**.

Taken together, our results indicate

1. a significant positive effect for the Western Germanic languages (ME, OE, DE) for **(H2)**, and
2. a significant negative effect for Western and Northern Germanic languages (IS) for **(H2)**

As a tentative hypothesis, one may speculate that languages separated for 1000 years (OE-IS) or more are too remote from each other to provide helpful background information, but that languages separated within the last 750 years (ME-DE) or less are still sufficiently close. This novel assumption may provide a guideline for future efforts to project annotations among related languages, and is thus of immense practical relevance for developing future NLP tools for historical and less-resourced language varieties. Ultimately, one may formulate rules of best practice like the following:

- If no syntactic annotations for a target language are available, annotation projection among closely related languages may be a solution. Even with limited amounts of parallel data, diachronic distances of more than 500 years can be successfully bridged (EN/ME, baseline).

- If no syntactic annotations for a target language are available, a parser trained on hyperlemmatized corpora in two languages may yield a performance comparable to a parser trained on small amounts of target data. A parser trained on hyperlemmatized monolingual data may be significantly worse (H2).

- The sparsity of parallel text to conduct annotation projection and train a (hyperlemmatized) parser can only be compensated by adding parallel data from *one* related language if these are closely diachronically related (with a separation being less than, say, 500 years ago) *and* at a similar developmental stage (DE/ME, H1). Adding data from multiple, equally remote languages does not necessarily improve the results further.

At the current state, such recommendations would be premature, they require deeper investigation, but with the confirmation of **(H2)** and **(H3)**, we can now motivate larger-scale efforts to compile a massive parallel corpus of historical Germanic language varieties as a basis for subsequent studies. Initial steps towards this goal are described in the following section.

5 Towards a massive parallel corpus of historical Germanic languages

With the long-term goal to systematically assess the impact of the factor of diachronic proximity, we focus on annotation projection among the Germanic languages as test field. The Germanic languages represent a particularly well-resourced, well-documented and well-studied language family which development during the last 1800 years is not only well-explored, but also documented with great amounts of (parallel) data, ranging from the 4th century Gothic bible over a wealth of Bible translations since the middle ages to the modern age of communication with its abundance of textual resources for even marginal varieties. Motivated from our experiment, we thus began to compile a parallel corpus of historical and dialectal Germanic language varieties. Primary source data for a massive parallel corpus of historical varieties of any European language is mostly to be drawn from the Bible and related literature. The Bible is the single most translated book in the world and available in a vast majority of world languages. It is also often the case that there are several biblical translation existing for a language. Bible data also represents the majority of parallel data available for historical Germanic languages, and for the case of OS and OHG, gospel harmonies represent even the majority of data currently known. Beyond this, the corpus includes Bible excerpts and paraphrases from all Germanic languages and their major historical stages.

Tab. 2 gives an overview over the current status of the Parallel Bible Corpus. At the moment, 271 texts with about 38.4M tokens have been processed, converted from their original format and verse-aligned according to their original markup or with a lexicon-supported geometric sentence aligner (Tóth et al., 2008). In the table, 'text' means any document ranging from a small excerpt such as the Lord's Prayer (despite their marginal size valuable to develop algorithms for normalization/[hyper]lemmatization) over gospel harmonies and paraphrases to the entire bible that has been successfully aligned with Bible verses. The compiled corpus, excerpts and fragments for all Germanic languages marked up with IDs for verses, chapters and books. For data representation, we employed an XML version of the CES-scheme developed by Resnik et al. (1997). Having outgrown the scale of Resnik's earlier project by far, we are currently in transition to TEI P5.

As it is compiled from different sources, the corpus cannot be released under a free or an academic license. It contains material without explicit copyright statement, with proprietary content (e.g., from existing corpora), or available for personal use only. Instead, we plan to share the extraction and conversion scripts we used. For the experiments we aim to prepare, we focus on primary data, the texts in this collection are not annotated. Where annotations are available from other corpora or can be produced with existing tools, however, these annotated versions will be aligned with the Bibles and included in subsequent experiments.

	after 1900	1800-1900	1600-1800	1400-1600	1100-1400	before 1100
						West Germanic
English	2	2	2	6	3 (+2)	1
Pidgin/Creol	2					
Scots	(6)			(1)		
Frisian	2 (+8)	(12)				
Dutch	4		1	5		(1)
L. Franconian	(47)	(21)				
Afrikaans	3					
German	3	1	(19)	1 (+4)	1 (+1)	1
dialects	3 (+2)					
Yiddish	1					
Low German	3 (+18)	(66)		(2)		1
Plautdietsch	2					
Danish	1				**North & East Germanic**	
Swedish	3			(3)	(1)	
Bokmål	2					
Nynorsk	2					
Icelandic		1		1		
Faroese	1					
Norn			(2)			
Gothic						1
tokens	21.8M	3.2M	2.7M	9.2M	1.2M	0.2M

Table 2: Verse-aligned texts in the Germanic parallel Bible corpus (parentheses indicate marginal fragments with less than 50,000 tokens)

6 Summary and outlook

This paper describes a motivational experiment on annotation projection, or more precisely, strategies to compensate data sparsity (the lack of parallel data) with material from related, but heterogeneous varieties to facilitate cross-language parser adaptation for low-resource historical languages. We used a fragment-aware dependency parser trained on annotation projections from ESV Bible to four historical languages.

Our results indicate a lexicalized fragment-aware parser trained on a small amount of annotation projections can yield good results on closely related languages. In a situation of the absence of training data for the target language (or, for example, in the situation where there is no parallel corpora for the target language), a hyperlemmatized parser trained on (projected) annotations from two or more related languages is likely to outperform a parser trained on a single related language.

We achieved statistically significant differences in parser performance trained on (a) target language data, and (b) target language and data from related varieties, resp. (c) data from related varieties only. These indicate that closely related languages (say, with a common ancestor about 750 years ago, such as DE and ME) have some potential to compensate sparsity of parallel data in the target variety, wheres this potential does not seem to exist for more remotely related languages (say, with a common ancestor more than 1000 years ago such as OE and IS).

The experimental results revealed that the parser performance can, indeed, be improved by means of including a related language to the training data, but we had a significant effect for only one language under consideration, indicating that the diachronic proximity of the languages considered was possibly too large, and thereby motivating subsequent experiments, and in particular, the creation of a larger parallel corpus of historical Germanic language varieties. We described initial steps in the compilation of this corpus.

Our experiment raises a number of open issues that are to be pursued in subsequent studies:

1. Our setup has a clear bias towards English (in the annotation schemes used and the source annotations), and parser performance was strongly affected by the syntactic difference between the target language and Modern English from which the syntactic dependencies were projected, indicating the relevance of diachronic relatedness as well as the developmental state of a related language. Subsequent experiments will hence address the inclusion of richer morphological features, projection from other languages and evaluation against syntactic annotations according to other schemes not derived from the Penn Treebank, as currently available, for example, for Old High German, Old Norse, and Gothic.

2. The *hyperlemmatization* in our approach was achieved through alignment/SMT, and a similar lexically-oriented approach has been suggested by (Zeman and Resnik, 2008). Alternative strategies more suitable for scenarios with limited amounts of training data may include the use of orthographical normalization techniques (Bollmann et al., 2011) or substring-based machine translation (Neubig et al., 2012) and are also subject to on-going research. We assume that SMT-based hyperlemmatization introduces more noise than these strategies, so that it is harder to achieve statistically significant results. Our findings are thus likely to remain valid regardless of the hyperlemmatization strategy. This hypothesis is, however, yet to be confirmed in subsequent studies.

3. Our experiment mostly deals with data translated from (or at least informed by) the Latin Vulgate. Our data may be biased by translation strategies which evolved over time, from very literal translations (actually, glossings) of Latin texts in the early middle ages to Reformation-time translations aiming to grasp the intended meaning rather than to preserve the original formulation. A focus on classical languages is, however, inherent to the parallel material in our domain. A representative investigation of annotation projection techniques thus requires the consideration of quasi-parallel data along with parallel data. This can be found in the great wealth of medieval religious literature, with Bible paraphrases, gospel harmonies, sermons and homilies as well as poetic and prose adaptations of biblical motives. The parallel corpus of Germanic languages thus needs to be extended accordingly.

4. One may wonder how the annotation projection approach performs in comparison to direct applications of modern language NLP tools to normalized historical data language (Scheible et al., 2011). While it is unlikely that such an approach could scale beyond closely related varieties, successful experiments on the annotation of normalized historical language have been reported, although mostly focused on token-level annotations (POS, lemma, morphology) of language stages which syntax does not greatly deviate from modern rules (Rayson et al., 2007; Pennacchiotti and Zanzotto, 2008; Kestemont et al., 2010; Bollmann, 2013). For the annotation of more remotely related varieties with more drastic differences in word order rigidity or morphology as considered here, however, projection techniques are more promising as they have been successfully applied to unrelated languages, as well, but still benefit from diachronic proximity, cf. Meyer (2011) for the projection-based morphological analysis of Modern and Old Russian.

The goal of our experiment was not to achieve state-of-the-art performance, but to show whether background material from related languages with different degrees of diachronic distance can help to compensate data sparsity, in this case with an experiment on annotation projection. This hypothesis could be confirmed and we found effects that – even on the minimal amounts of data considered for this study – indicated statistically significant improvements.

It is thus to be expected that even greater improvements can be achieved by considering more closely related pairs of languages, with greater amounts of data. The further exploration of this hypothesis is the driving force behind our efforts to compile a massive corpus of parallel and quasi-parallel texts for all major varieties of synchronic and historical Germanic languages. Algorithms successfully tested in this context can be expected to be applicable to other scenarios in which, e.g., well-researched modern languages may be employed to facilitate the creation of NLP tools for less-ressourced, related languages. Our efforts are thus not specific to historical languages.

As the diachronic development and the diversification of the Germanic languages is well-documented in this body of data, and the linguistic processes involved are well-researched, this data set represents an extraordinarily valuable resource for philological and comparitve studies as well as Natural Language Processing. In particular, we are interested in developing algorithms that explore and exploit the variable degree of diachronic relatedness found between the languages in our sample. At the same time, we cooperate with researchers from philology, historical and comparative linguistics, which research on intertextuality, diachronic lexicology, phonology, morphology and syntax we aim to support with NLP tools developed on the basis of this body of parallel text.

References

Miguel Ballesteros and Joakim Nivre. 2012. Maltoptimizer: A system for maltparser optimization. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage (LaTeCH-2011)*, pages 34–42, Hissar, Bulgaria, September.

Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 11–18, Sofia, Bulgaria, August. Association for Computational Linguistics.

Michael Collins, Lance Ramshaw, Jan Hajič, and Christoph Tillmann. 1999. A statistical parser for czech. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 505–512. Association for Computational Linguistics.

Michael Cummings. 2010. *An Introduction to the Grammar of Old English: A Systemic Functional Approach*. Functional Linguistics. Equinox Publishing Limited.

Robert P. Ebert. 1976. *Infinitival complement constructions in Early New High German*. Linguistische Arbeiten. De Gruyter.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia, May 25-26.

Daisuke Kawahara and Kiyotaka Uchimoto. 2007. Minimally lexicalized dependency parsing. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 205–208. Association for Computational Linguistics.

Mike Kestemont, Walter Daelemans, and Guy De Pauw. 2010. Weigh your words: Memory-based lemma-retrieval for Middle Dutch literary texts. In *CLIN 2010. Computational linguistics in the Netherlands 20*, Utrecht, The Netherlands, May.

Roland Meyer. 2011. New wine in old wineskins? Tagging Old Russian via annotation projection from modern translations. *Russian linguistics*, 35(2):267–281.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–174, Jeju Island, Korea, July. Association for Computational Linguistics.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Marco Pennacchiotti and Fabio Massimo Zanzotto. 2008. Natural language processing across time: An empirical investigation on italian. In *Advances in Natural Language Processing*, pages 371–382. Springer.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of the 4th Corpus Linguistics Conference (CL-2007)*, Birmingham, UK.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1997. Creating a parallel corpus from the book of 2000 tongues. In *Proc. of the Text Encoding Initiative 10th Anniversary User Conference (TEI-10)*.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurdhsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *LREC*, pages 1977–1984.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH-2011)*, pages 19–23, Portland, OR, USA, June.

Kathrin Spreyer, Lilja Ovrelid, and Jonas Kuhn. 2010. Training parsers on partial trees: A cross-language comparison. In *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, Valletta, Malta, May.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003a. The Penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003b. The york-toronto-helsinki parsed corpus of old english prose. *University of York*.

Krisztina Tóth, Richárd Farkas, and András Kocsor. 2008. Sentence alignment of hungarian-english parallel corpora using a hybrid algorithm. *Acta Cybern.*, 18(3):463–478, January.

C. Trips. 2002. *From OV to VO in Early Middle English*. Linguistics today. John Benjamins Pub.

A. van Kemenade and B. Los. 2009. *The Handbook of the History of English*. Blackwell Handbooks in Linguistics. John Wiley & Sons.

Hiroyasu Yamada and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of IWPT. Vol. 3. 2003.*

David Yarowski and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL 2001*, pages 200–207.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–41, Hyderabad, India, Jan.