New Technologies for Old Germanic. Resources and Research on Parallel Bibles in Older Continental Western Germanic

Christian Chiarcos, Maria Sukhareva, Roland Mittmann,
Timothy Price, Jens Chobotsky, and **Gaye Detmold**
Goethe University Frankfurt, Germany
{lastname}@em.uni-frankfurt.de

Abstract

We provide an overview of on-going efforts to facilitate the study of older Germanic languages currently pursued at the Goethe-University Frankfurt, Germany.

We describe created resources, such as a parallel corpus of Germanic Bibles and a morphosyntactically annotated corpus of Old High German (OHG) and Old Saxon, a lexicon of OHG in XML and a multilingual etymological database. We discuss NLP algorithms operating on this data, and their relevance for research in the Humanities.

RDF and Linked Data represent new and promising aspects in our research, currently applied to establish cross-references between etymological dictionaries, infer new information from their symmetric closure and to formalize linguistic annotations in a corpus and grammatical categories in a lexicon in an interoperable way.

1 Background

We describe on-going efforts at the Goethe University Frankfurt on the study of older Continental Western Germanic languages, in particular, Old High German (OHG, ancestor of German), Old Saxon (OS, ancestor of Low German) and (to a lesser extent) Old Low Franconian (OLF, ancestor of Dutch) and their relation to Old English (OE), Gothic, German and other Germanic languages as well as the relation of OHG and OS religious texts to their Latin sources. This line of research is conducted in the context of two larger efforts, the Old German Reference Corpus and the LOEWE cluster "Digital Humanities", in collaboration with the Applied Computational Linguistics group at the Goethe-Universitt Frankfurt.

The Old German Reference Corpus is a DFG-funded project that emerged from the Deutsch Diachron Digital (DDD) initiative, conducted in cooperation between HU Berlin, U Frankfurt and U Jena, and aims to provide a morphosyntactically annotated, exhaustive reference corpus of Old High German and Old Saxon. The LOEWE cluster "Digital Humanities",[1] funded through a programm of the State of Hessen, is a collaboration between U Frankfurt, TU Darmstadt and Freies Deutsches Hochstift Frankfurt aiming to develop methodologies and infrastructures to facilitate information-technological support of research in the humanities.

The collaboration between the humanities and NLP described here is guided by different, though converging interests: For the **humanities**, the language resources, annotations, alignment and tools created in collaboration with NLP researchers represent novel instruments complementing traditional philological approaches, e.g., to investigate emergence and decay of syntactic patterns.

From an **NLP perspective**, the Germanic languages provide a test-bed to develop strategies for novel algorithms for alignment and annotation projection. In particular, the abundance of parallel (Bible) texts for all major language stages of most Germanic languages, the excellent NLP support for modern Germanic languages, and the availability of a considerable body of annotated historical texts allow us to study the impact of the factor of *diachronic relatedness* when building resources for low-resource languages.

2 Corpus Data

Along with annotated corpora provided by third parties (Tab. 1), two important data sets have been constructed in the course of our research. These include a massive, verse-aligned Bibles corpus

[1] http://www.digital-humanities-hessen.de

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 22–31,
Gothenburg, Sweden, April 26 2014.

covering all Germanic languages, and the Old German Reference Corpus. In additional, a thematical alignment of quasi-parallel text within and across biblical texts was extrapolated from the literature.

2.1 Germanic parallel Bible corpus

Bible data 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. Hence, we began compiling a corpus of Bible texts, excerpts and fragments for all Germanic languages marked up with IDs for verses (if possible), 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 XML format. At the moment, 271 texts with about 38.4M tokens have already been processed (Tab. 2). Copyright prevents redistributing most of this data under a free or an academic license, but we plan to share the extraction and conversion scripts we used. . Except for automatically parsed Bibles in modern English, German and Swedish, the texts in this collection are not annotated. Where annotations are available from other corpora (Tab. 1), however, these were aligned with our Bibles.

2.2 Old German Reference Corpus

The Old German Reference Corpus (*Referenzkorpus Altdeutsch*) (Mittmann, 2013) is a joint project in cooperation between HU Berlin, U Frankfurt and U Jena, conducted in the wider context of the Deutsch Diachron Digital (DDD) initiative. The DDD initiative aims to provide deeply-annotated reference corpora of different historical stages of German. The Old German Reference Corpus comprises all preserved texts from the oldest stages of continental Western Germanic (OHG and OS) dating from ca. 750 to 1050 CE, 650,000 tokens in total. Among the largest coherent subcorpora are Tatian (OHG), Otfrid of Weissenburg (OHG) and the Heliand (OS). From these, only Tatian can be verse-aligned with the gospels (and is included in Tab. 1 and 2), while the Heliand and Otfrid are free renderings of the gospels. For these, the literature provides a section-level alignment only.

The DDD builds on the earlier efforts of the TITUS project (Thesaurus of Indo-European Text and Language Materials, *Thesaurus Indogermanischer Text- und Sprachmaterialien*) that provided digitized editions of texts in old Germanic languages as well as other Indo-European and selected non-Indo-European languages (Gippert, 2011).[2]

The annotations are mostly derived from the literature and existing glossaries that provide grammatical information for all known OHG and OS words, together with their exact source. These have been digitized, automatically applied to the text, manually refined using the annotation software ELAN,[3] augmented with metadata, and finally published via the ANNIS database (Linde and Mittmann, 2013).

The annotated corpus is published under a CC-BY-SA license over `http://www.laudatio.org`, where ELAN and relANNIS files are provided. So far, the OHG Tatian is available, further data sets are currently in preparation.

2.3 Thematical alignment within and across biblical texts

Translations of religious texts are well-suited for language comparison as well as NLP experiments exploiting parallel data as they are not only faithfully translated, but also, they come with a verse-level alignment which can serve as a basis for statistical word-level alignment, using, e.g., GIZA++ (Och and Ney, 2003). Where such a verse-level is not explicitly given, it can be automatically identified for actual translations. However, for independent compositions such as gospel harmonies, alignment is harder to identify and can only be established at the level of sections. In addition, similar links also exist *between* different parts of the Bible, e.g., parallel passages in different gospels.

For these, an index providing a coarse-grained thematical alignment at the level of sections was extrapolated from the literature. This index can be exploited to increase the coverage of the alignment: where no exact translation is available (historical language data is often fragmentary), a thematically matching section is retrieved. Furthermore, consulting the verse under consideration together with renderings of quasiparallel parts of the same text allows historical linguists to grasp the degree of grammatical variability for the phenomena they are interested in. Language comparison can thus be particularly well accomodated if mul-

[2] `http://titus.uni-frankfurt.de/texte/texte2.htm#ahd` and `#asachs`

[3] `http://www.lat-mpi.eu/tools/elan`

language	period		syntax	tok.	corpus
English	Modern	19th	CS	21K	(Kroch et al., 2010)
	British	18th	CS	32K	(Kroch et al., 2010)
	Early	17th	CS	22K	(Kroch et al., 2004)
	Modern	16th	CS	21K	(Kroch et al., 2004)
	Middle	14th	CS	66K	(Kroch and Taylor, 2000)
	Old	10th	CS	78K	(Taylor et al., 2003b)
			DS	7K	(Haug and Jøhndal, 2008)
Icelandic	Middle	16th	CS	40K	(Rögnvaldsson et al., 2012)
High German	Early Mod.	16th	CS	27K	(Light, 2013)
	Old	9th	CH	41K	Sect. 2.2
Gothic		4th	DS	56K	(Haug and Jøhndal, 2008)

Table 1: Verse-aligned older Germanic Bible texts from various corpora with manual annotations for morphosyntax and syntax (CH chunks, CS constituents, DS dependencies)

	after 1900	1800-1900	1600-1800	1400-1600	1100-1400	before 1100
				Insular West Germanic		
English	2	2	2	6	3 (+2)	1
Pidgin/Creol	2					
Scots	(6)			(1)		
Frisian	2 (+8)	(12)		**Continental West Germanic**		
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)

tiple versions of the same passage in the same language can be provided.

To exploit redundancy and to enlarge the number of parallel and quasi-parallel passages for a given phenomenon searched in the corpus, cross-references within the Bible and between the Bible and derived texts have been identified. For example, coarse-grained thematical alignment between different gospels is provided by the Eusebian Canon Tables and their subordinate Ammonian sections and are extendable to the Latin Tatian. For OS Heliand, a free adaptation of gospels, we have only a section-level thematical alignment with Tatian provided by Sievers (1872).

Information on these cross-references has been digitized and employed to create an interlinked index of thematically similar sections in the gospels and the OS and OHG gospel harmonies. Our Bible

	West Germanic					other		reconstr.	
lexicon	OE	OHG	OS	OLF	OFr	ON	Got	PGmc	PIE
entries (XML, in K)	25	24	9	2	13	12	5	9	7
triples (RDF, in M)	1.2	1.6	.6	.2	.6	.7	.4	.2	.2
lemon:Words & links (in K)									
OE	25					1			
OHG	2	26	7	2	3	1			
OS	1	4	9	1	2	1			
ON	1				1	14			
Got	1	1			1	1	6		
PGmc	5	3	3	1	2	4	2	8	
PIE	2	1	1	1	1	1	1		8
German	16	23	8	4	10	12	7	6	3
English		10	4	2	5		9		2
symmetric closure of etym. links (triples *per lang.* in K)	+11	+14	+11	+5	+9	+8	+5	+21	+9
links to (L)LOD data sets (triples *per data set* in K)									
OLiA	24	22	8	2	12	11	5	8	7
lexvo	132	186	82	21	68	82	49	14	15
Glottolog	15	11	8	3	7	11	6	9	13

Table 3: Statistics on the etymological dictionaries, including Old Low Franconian (OLF), Old Frisian (OFr), Old Norse (ON), Gothic (Got), Proto-Germanic (PGmc) and Proto-Indo-European (PIE)

data is thus accompanied with an index that links disparate texts from different time periods and in distinctive styles and variant languages on the basis of thematical similarity as identified in the literature. For gospels and gospel harmonies, we identified 4560 inter-text groups made up of the related chunks between all the originals and languages involved that represents the basis for a more fine-grained level of alignment (Price, 2012).

3 Linked Lexicon Data

A large lexical database of etymologically linked dictionaries of old Germanic languages (OS, OHG, OE, Gothic, Old Norse, Old Frisian, Old Low Franconian, Proto-Germanic; also Proto-Indo-European) has been developed in the context of the LOEWE cluster 'Digital Humanities' at the U Frankfurt. Building on the etymological and translational dictionaries of Old Germanic languages by Gerhard Köbler,[4] the project 'Historical Linguistic Database' developed user-friendly means of comparing etymologically related forms between historical dialects and their daughter languages (Price, 2012). The original PDF data were converted into an XML representation, cross-references have been resolved and the results are

[4] http://www.koeblergerhard.de/ahdwbhin.html

imported into an XML database. A web interface has been developed, that transforms user queries into XQuery and visualizes the results in a convenient way using XSLT.

To provide a machine-readable representation of the etymological dictionaries, an **RDF version** has been compiled. Applying the Linked Data paradigm (Bizer et al., 2009) to etymological lexicons is particularly promising as they are characterized by a heavy linkage across different languages, so that etymological lexicons for different languages are very likely to complement each other. RDF provides the means to represent the cross-language linking using a uniform formalism, and subsequently, to facilitate information aggregation over multiple etymological lexicons as well as language-specific lexical resources.

We converted the Köbler lexicons to RDF in conformance to the Lemon model (McCrae et al., 2011), an LMF-based vocabulary to represent machine-readable lexicons by using Semantic Web standards. This conversion followed the three main objectives:

(i) linkability: XML-based query languages such as XQuery and XPath, used to create the user interface to the lexicons, limit our lexicon to a tree-structure representation. However, as our lexicons complement each other, it would be desirable to provide explicit cross-references between these entries, and to allow them to be queried jointly. Within the RDF data model, the relations within and beyond a single lexicon can be represented and queried with equal ease, surmounting constraint imposed by XML.

(ii) interoperability: Instead of resource-specific abbreviations for languages and grammatical categories, we represent linguistic information and meta data by reference to community-maintained vocabularies publicly available as part of the (Linguistic) Linked Open Data cloud, namely lexvo (de Melo, to appear, ISO 639-3 language codes), Glottolog (Nordhoff and Hammarström, 2011, language families) and OLiA (Chiarcos, 2008, linguistic categories). Reusing vocabularies shared among many parties over the Web of Data has the advantage that resources dealing with related phenomena in the same language can be easily identified and their information integrated without additional conversion steps.

(iii) inference: The original lexicons were distributed in individual PDF files, and the XML representation was created as a faithful representation of their content, augmented with markup for relevant linguistic features. These files, however, provided complementary information, so that, say, a lexicon entry in the OS dictionary provided a reference to an etymological corresponding OHG entry, but this reference was not found in the OHG dictionary. Such gaps can be easily detected (and filled) through symmetric closure in the RDF data model.

The results of this conversion are summarized in Tab. 3. In the original XML (first row), every entry corresponds to a lemma of the language under consideration, with different etymologies (and/or senses) being associated with it. In RDF (second row), each of these homographs (together with its definition number) is defined as a `lemon:Word` with a homography relation with the homograph set (represented by a `lemon:Word` *without* definition number). The number of `lemon:Words` is thus slightly higher than the number of entries in the original dictionaries. Differently from the XML, however, information from different data sets can be easily aggregated, and triples originating from one document can be complemented with triples from another, shown here for the symmetric closure of etymological relations (third row) that can be easily generated using a simple SPARQL pattern like `CONSTRUCT { ?o ?p ?s } WHERE {?s ?p ?o}`. The last row shows links to other data sets from the (Linguistic) Linked Open Data cloud. Most original entries were complemented with grammatical information using different (and not fully consistent) abbreviations. For the most frequent abbreviations used, a link to the corresponding OLiA concept was generated. These definitions are thus *interoperable* beyond these lexicons and can be compared, e.g., with those of lexical-semantic resources for Modern German and English as compiled in (Eckle-Kohler et al., to appear). Similarly, language abbreviations were mapped to ISO 639-3 codes (in lexvo), or, where these were not available, to Glottolog. Even though the number of data in historical languages is constantly increasing and there is a demand for fine-grained language codes for them, neither of the aforementioned resources provide such codes. So we had to use a link to the corresponding language family instead.

language	period	scheme	corpus reference
English	Modern	PTB	(Taylor et al., 2003a; Kroch et al., 2010)
	Early Mod.	PPCEME	(Kroch et al., 2004)
	Middle	PPME2	(Kroch and Taylor, 2000)
	Old	YCOE	(Taylor et al., 2003b)
		PROIEL	(Taylor et al., 2003b)
High German	Modern	STTS	(Schiller et al., 1999)
	Early Mod.	PCENHG	(Light, 2013)
	Old	Sect. 2.2	
		T-CODEX	(Petrova et al., 2009)
Dutch	Modern	Alpino	(Bouma et al., 2001)
Old Norse		Menota	(Haugen et al., 2008)
Danish	Modern	EAGLES	(Leech and Wilson, 1996)
Swedish	Modern	Mamba	(Nivre et al., 2006)
Icelandic		IcePaHC	(Rögnvaldsson et al., 2012)
Gothic		PROIEL	(Haug and Jøhndal, 2008)

(a) Morphosyntactic annotations

language	period	scheme	corpus reference
English	Modern	PTB	(Taylor et al., 2003a; Kroch et al., 2010)
		Stanford deps	(De Marneffe and Manning, 2008)
		Penn2Malt deps	(Johansson and Nugues, 2007)
	Early Mod.	PPCEME	(Kroch et al., 2004)
	Middle	PPME2	(Kroch and Taylor, 2000)
	Old	YCOE	(Taylor et al., 2003b)
		PROIEL	(Taylor et al., 2003b)
High German	Modern	TIGER	(Brants et al., 2004)
		Tüba-D/Z	(Telljohann et al., 2003)
		NEGRA	(Skut et al., 1997)
	Early Mod.	PCENHG	(Light, 2013)
Dutch	Modern	Alpino	(Bouma et al., 2001)
Swedish	Modern	Mamba	(Nivre et al., 2006)
Icelandic		IcePaHC	(Rögnvaldsson et al., 2012)
Gothic		PROIEL	(Haug and Jøhndal, 2008)

(b) Syntactic annotations

Table 4: List of annotation schemes represented as OWL2/DL ontologies and relevant Germanic corpora

4 NLP methods applied

We sketch selected NLP applications developed on the data described before, the automated phrase-level alignment of quasi-parallel text, and two experiments on annotation projection on parallel text. All of these experiments are still in a relatively early stage.

4.1 Automated phrase-level alignment of quasi-parallel text

The needs of historical lingustics demand a more fine-grained alignment than the currently available thematical alignment of Heliand with Tatian and the gospels. We thus investigate parallel phrase detection between Heliand (OS) and Tatian (OHG), resp., Heliand and the West Saxon gospels (OE).

To identify cognate phrases, we explore 6 types of similarity metrics $\delta(w_{OS}, w_{OHG})$ for every OS word w_{OS} and its potential OHG cognate w_{OHG}.

1. geometry δ_g = difference between the relative positions of w_{OS} and w_{OHG}.

2. identity $\delta_i(w_{OS}, w_{OHG}) = 1$ iff $w_{OHG} = w_{OS}$ (0 otherwise)

3. lexicon $\delta_{lex}(w_{OS}, w_{OHG}) = 1$ iff $w_{OHG} \in W$ (0 otherwise) where W is a set of possible OHG translations for w_{OS} suggested by a lexicon, i.e., either

direct etymological link in (the symmetric closure of) the etymological dictionaries, or

indirect shared German gloss in the etymological dictionaries

4. orthography similarity measure based on character replacement likelihood:

relative Levenshtein similarity $\delta_{lev}(w_{OS}, w_{OHG}) = 1 - \frac{ld}{|w_{OS}|+|w_{OHG}|}$ where ld is the standard Levenstein distance and $|w_{OS}|$ and $|w_{OHG}|$ are the number of characters in each word.

statistical character replacement probability as approximated by a character-based statistical machine translation system (Neubig et al., 2012)

5. normalization $\delta_{norm}(w_{OS}, w_{OHG}) = \delta_i(w'_{OS}, w_{OHG})$, with w'_{OS} being the OHG 'normalization' of the original w_{OS}. Here, normalization uses a weighted Levenshtein distance and a fixed list of OHG target words (Bollmann et al., 2011).

6. cooccurrences $\delta_p(w_{OS}, w_{OHG}) = P(w_{OS}|w_{OHG})P(w_{OHG}|w_{OS})$, calculated on thematically aligned sections from both texts.

For any two thematically aligned OS and OHG word vectors, we thus span up a similarity matrix between both word vectors on the basis of these metrics. On the matrices, different operations can be applied to calculate similarity derived metrics, including point-wise multiplication or addition, thresholds and a smoothing operator, that aligns words due to the similarity of its neighbors. The resulting matrix is then decoded by a greedy algorithm that aligns the words with the highest score, and then iterates for the remaining words.

At the moment, we provide a graphical interface over a webpage that allows a philologist to dynam-

ically define an alignment function and that provides a graphical visualization of the result. During a partial qualitative evaluation a historical linguist was asked to compare the results of alignment based on various metrics applied to a small text passage. He took into consideration the overall match of the topic of the aligned passages as well as the number of parallel passages that the metrics failed to align. Eventually, it was indicated that the best results can be achieved by combining multiple metrics. A combination of either direct lexicon-based or normalization-based alignment and geometrical alignment appears to be particularly promising. Yet, systematic experiments to automatically explore this feature space are still being prepared and depend on the availability of a gold alignment for selected verses.

4.2 Projecting dependency relations

As shown in Tab. 1, we only possess shallow syntactic annotations of OHG (and OS) text. We are thus particularly interested in establishing richer syntactic annotations. A challenging aspect in this respect is the limited availability of parallel training data for historical language stages. However, due to diachronic relatedness, we may expect that syntactic patterns of Old Germanic languages are preserved in their modern descendants. Such an approach requires a consistent hyperlemmatization, e.g., against a modern language

We tested this idea on Bible texts from four corpora with closely related annotation schemes for syntax (Tab. 1, corpora with CS-syntax): Icelandic (IS), Early Modern High German (DE), Middle English (ME) and Old English (OE). These schemes originate in the Penn Treebank scheme (Taylor et al., 2003a), and we thus parsed a modern English Bible with a parser trained on the Penn Treebank. As older Germanic languages are characterized by a higher degree of word order flexibility than Modern English, we converted historical and modern annotations to dependency relations using standard tools for this task (Johansson and Nugues, 2007). Word-alignment was obtained with GIZA++ and 1:1 alignment was enforced using the translation table. Then, we projected dependency relations and the English words as hyperlemmas for the historical texts. The historical texts had comparable POS annotation that was only slightly normalized across the corpora as it preserved more morphological information than Modern English POS tags.

On these projections, a fragment-aware parser was trained using the English (hyper)lemmas and the original POS tags (Spreyer and Kuhn, 2009). We limited the amount of parallel data available to a training set of 437 sentences per language and a test set of 174 per language. Our hypothesis was that in this setting, (projected) training data from related languages can be used *in place of* training data for the language under consideration, *if* the amount of data is sufficient *and* the languages are sufficiently closely related. Furthermore, we assumed that with an increasing number of languages considered (and thus training set size), the quality of the projected annotations would continuously improve *as long as* the languages are sufficiently closely related.

For evaluation, we employed the unlabeled attachment score (UAS) (Collins et al., 1999) on the test data and compared with the (dependency version of) the original annotation in these corpora. Tab. 5 compares the performance of a parser trained on target language data with parsers trained on (hyperlemmatized) related languages. The scores in the second column are the baseline UAS where the parser was applied to the same language as it was trained on. The third column shows the difference with the parser applied to a language but trained on projections into another language. The fourth and the fifth column provides the results of the parser trained on one or two additional related languages respectively.

The results showed that, among the West Germanic languages (but not IS), a parser trained on two or more related languages can reach the same performance or even outperforms a parser trained on the target language. Furthermore, a parser trained on (projected) annotations from two or more related languages is likely to outperform a parser trained on a single related language. Accordingly, in absence of parallel texts for the target language, the parser can be successfully trained on annotation projections from two or more related languages. It should be noted, however, that the overall performance of the parser was relatively poor. This may be, however, an artifact of the great grammatical divergency between Modern English (and, to a limited degree, ME: reduced morphology, strict word order) and older Germanic languages (rich morphology, flexible word order).

Subsequent experiments will thus address the

inclusion of richer morphological features, projections from other languages and evaluation against another set of dependency (DS) annotations for Gothic and Old English (Tab. 1), for which related annotation schemes for Latin, Greek and Czech are available – all of these languages are characterized by rich morphology and flexible syntax.

Tgt	on Tgt lang.	on related languages: Best monoling. model	ΔUAS	Best biling. model	ΔUAS	Triling. model	ΔUAS
DE	.41	IS	$+.02^{\text{n.s.}}$	+ME	$+.05^{*}$	+OE	$+.04^{**}$
IS	.32	ME	$-.06^{***}$	+DE	$-.03^{\text{n.s.}}$	+OE	$-.04^{*}$
ME	.60	IS	$-.04^{***}$	+OE	$-.01^{\text{n.s.}}$	+DE	$-.02^{\text{n.s.}}$
OE	.30	ME	$.00^{\text{n.s.}}$	+IS	$.00^{\text{n.s.}}$	+DE	$.00^{\text{n.s.}}$

Table 5: Performance of parsing models (UAS difference vs. 2nd col. with χ^2: * $p < .05$, ** $p < .01$, *** $p < .005$)

4.3 Harmonization of grammatical features

Another line of studies addresses the projection of grammatical features as represented in POS tags and dependency labels. Unfortunately, modern and historical language stages are annotated according to a great variety of annotation schemes which can not be trivially mapped to a generalization without substantial loss of information (as, e.g., in the approach by Petrov et al., 2012). For processing of multilingual corpora the problem of heteroginity of linguistic annotations is very acute. Above, we described an experiment that used PTB style annotations only. This limitation was imposed by the annotation schema of the target corpora that had PTB style syntactic annotations.

We thus follow Chiarcos (2008) and represent the most relevant Germanic annotation schemes as OWL2/DL ontologies, and link these to an overarching Reference Model. Unlike a tagset, whose string-based annotations require disjoint categories at a fixed level of granularity, this ontology-based approach allows to decompose the semantics of annotations and consider all aspects independently. For example, a tagger may correctly identify plural agreement but incorrectly assume that it pertains a noun, as in the Penn Treebank tag `NNS`. In the original tagset, a corresponding tag for, say, adjectives, does not exist, but using the ontology, a plural adjective could nevertheless be represented in the form of different RDF triples. With lexicon data being available in RDF and linked to the OLiA Reference Model, as well (Sect. 3), the incorrect word class can be spotted, and corrected, but the agreement information could remain unaffected.

These annotations have also been successfully employed in ensemble combination architectures, where information from different sources (say, NLP tools) was integrated on the basis of the Reference Model and disambiguated using ontological axioms (Chiarcos, 2010; Pareja-Lora and Aguado de Cea, 2010). In an annotation projection scenario, these sources could be projections from different languages annotated according to different schemes, e.g., German, English, Swedish or Latin. These experiments are currently being conducted, but Annotation Models for several schemes are already available (Tab. 4).

5 Digital Humanities

Our ultimate goal is to facilitate studies of historical and empirical linguists and philologists.

One research question under consideration is whether the Heliand influenced Luther (Price, 2012), who, apparently, possessed one copy. Based on a thorough comparison of thematically aligned passages, evidence for or against this hypothesis may be gathered, and this investigation can be simplified by limiting the search to parallel phrases automatically identified (Sect. 4.1).

Another research question pertains to divergencies between, e.g., OHG texts and their Latin source. As most OHG material is translated in a literal fashion, and the word order was relatively flexible, the OHG syntax may have been adjusted to mirror the Latin original. Research of OHG syntax thus concentrates on passages where OHG syntax differs from the Latin source (Hinterhölzl and Petrova, 2009).

Different types of divergencies have been identified by qualitative research. Early translations unlike modern ones tend to be very literal, often not being only word by word translation but also preserving the syntax of the original. Nevertheless, due to strong grammatical differences between two languages, various divergencies on (morpho)syntactic and lexical levels were unavoidable. Such, the transition from the Latin synthetic to OHG analytic wordforms in case of the deponent verbs is systematically observed. Also the changes of the word position as well as missing a word in translation or adding a word that is not present in the Latin original can be frequently found. Such divergencies can be often explained

by stylistic or pragmatic reasons as well as by personal preferences of the translator.

This line of research is currently supported through automated word-level alignment between the OHG and Latin versions of Tatian. We built a parallel corpus using GIZA++ and used the TreeAligner (Lundborg et al., 2007) for search and evaluation. On this basis, a philological comparison of OHG Tatian and its Latin source is being conducted. More helpful, however, would be a comparison of different syntactic patterns in OHG and Latin which motivates our experiments in annotation projection (Sect. 4.2).

Finally, our experiments in the ontology-based harmonization of different annotation schemes (Sect. 4.3) will facilitate subsequent typological and linguistic comparison across corpora with manual annotations for syntax and/or morphology according to different schemes.

6 Summary

We sketched major research directions on the development of resources, NLP tools and algorithms to facilitate the study Old Germanic languages currently pursued at the Goethe-University Frankfurt in the context of two related research initiatives, the LOEWE cluster 'Digital Humanities' and the project 'Old German Reference corpus'.

Our efforts resulted in the creation of the following resources:

- a massive **parallel corpus** of TEI-conformant Bibles including all contemporary Germanic languages as well as early stages of Germanic languages (Sect. 2.1).
- an exhaustive, **morphosyntactically annotated corpus of OHG and OS** with morphosyntactic annotations. Annotations were automatically derived from glossaries and manually refined (Sect. 2.2).
- an index providing a **thematical alignment** of the four gospels with each other as well as with OHG and OS gospel harmonies (Sect. 2.3). This high quality alignment provides a solid basis for further more fine-grained automatic alignment (Sect. 4.1).
- XML versions of **lexical resources**, including etymological dictionaries of Old Germanic languages (Sect. 3)
- an RDF-based **linked etymological database** of Old Germanic languages compiled from the latter (Sect. 3)
- a Linked Data representation of **annotation schemes** for corpora, NLP tools and grammatical features in the linked lexicon data (Sect. 3, 4.3)

The resources created provide an excellent testbed for various NLP algorithms, particularly for experiments on alignment and annotation projection techniques: We developed different metrics for **quasi-parallel alignment** applied to the corpus of gospel harmonies (Sect. 4.1). For subsequent analysis, evaluation and refinement by historical linguists, we provide a graphical visualization and user interface in a form of a webpage. This is an on-going project and further research will aim at refining metrics and their combination.

Our massive parallel corpus is a perfect prerequisite for **annotation projection** (Sect. 4.2). Our experiments on annotation projections and cross-lingual parser adaptation showed that it is possible to use (hyperlemmatized) training data from multiple closely related languages *in place of* training data for the language under consideration, and on small sets of parallel training data available, this did not lead to a significant loss of performance. The only exception in the experiment (IS) is also most remote from the other languages considered.

A severe limitation of this experiment was that it required operating on (variants of) the same annotation scheme. Another line of our research is focused on researching of ways to surmount such restrictions. We thus adopt a modular approach with annotation schemes linked to the OLiA Reference Model to harmonize annotations and grammatical features from lexicons (Sect. 4.3).

Finally, applications of these algorithms and resources in research questions in philology, historical linguistics and comparative linguistics were sketched in Sect. 5.

While most resources described in this paper have been developed for several years at the Goethe-University Frankfurt, the increased focus on NLP and Linked Data represent novel developments pursued by the newly established Applied Computational Linguistics Lab at the Goethe University Frankfurt. Different aspects of research sketched in this paper thus describe on-going activities at different degrees of completion.

Acknowledgements

The research of Christian Chiarcos, Maria Sukhareva, Tim Price, Gaye Detmold, and Jens Chobotsky described in this paper was supported by the research cluster 'Digital Humanities' at the Goethe-University Frankfurt, funded through the LOEWE programme of the federal state of Hessia. The research of Roland Mittmann was conducted in the project 'Old German Reference corpus', funded by the Deutsche Forschungsgemeinschaft (DFG).

References

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data – The story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22.

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.

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

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2(4):597–620.

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

Christian Chiarcos. 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.

Michael Collins, Lance Ramshaw, Jan Hajič, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999)*, pages 505–512, Maryland, June.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the COLING-2008 Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Gerard de Melo. to appear. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web Journal*, pages 1–7.

Judith Eckle-Kohler, John McCrae, and Christian Chiarcos. to appear. lemonUby – A large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal: Multilingual Linked Open Data*.

Jost Gippert. 2011. The TITUS Project. 25 years of corpus building in ancient languages. In *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften*, pages 169–192, Berlin, December.

Dag TT Haug and Marius Jøhndal. 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.

Odd Einar Haugen, Tone Merete Bruvik, Matthew Driscoll, Karl G Johansson, Rune Kyrkjebø, and Tarrin Wills. 2008. The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources.

Roland Hinterhölzl and Svetlana Petrova. 2009. *Information Structure and Language Change: New Approaches to Word Order Variation in Germanic*. Mouton de Gruyter.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of the 16th Nordic Conference on Computational Linguistics (NoDaLiDa-2007)*, pages 105–112, Tartu, Estonia, May.

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

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2010. The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE). Department of Linguistics, University of Pennsylvania. CD-ROM.

Geoffrey Leech and Andrew Wilson. 1996. EAGLES guidelines: Recommendations for the morphosyntactic annotation of corpora.

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

Sonja Linde and Roland Mittmann. 2013. Old German Reference Corpus. Digitizing the knowledge of the 19th century. In Paul Bennett, Martin Durrell, Silke

Scheible, and Richard J. Whitt, 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.

Joakim Lundborg, Torsten Marek, Maël Mettler, and Martin Volk. 2007. Using the Stockholm TreeAligner. In *Proceedings of the 6th Workshop on Treebanks and Linguistic Theories (TLT-2007)*, pages 73–78.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*, pages 245–259. Springer.

Roland Mittmann. 2013. Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(2):39–52.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 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.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 1392–1395.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science 2011 (LISC-2011)*.

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

Antonio Pareja-Lora and Guadalupe Aguado de Cea. 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.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 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.

Svetlana Petrova, Michael Solf, Julia Ritz, Christian Chiarcos, and Amir Zeldes. 2009. Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. *TAL*, 50(2):47–71.

Timothy Blaine Price. 2012. Multi-faceted alignment: Toward automatic detection of textual similarity in gospel-derived texts. In *Proceedings of Historical Corpora 2012*, Frankfurt, Germany.

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

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universitäten Stuttgart und Tübingen.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 88–95.

Kathrin Spreyer and Jonas Kuhn. 2009. Data-driven dependency parsing of new languages using incomplete and noisy training data. In *Proc. of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 12–20, Boulder, CO, June.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003a. The Penn Treebank: An overview. In Anne Abeill, editor, *Treebanks*, pages 5–22. Springer, Dordrecht.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003b. The York-Toronto-Helsinki parsed corpus of Old English prose.

Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2003. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Germany.