The *Anselm Corpus*: Methods and Perspectives of a Parallel Aligned Corpus

Stefanie Dipper[1], *Simone Schultz-Balluff*[2]

(1) Department of Linguistics, Ruhr University Bochum
(2) German Department, Ruhr University Bochum

dipper@linguistics.rub.de, simone.schultz-balluff@rub.de

ABSTRACT
This paper presents ongoing work in the *Anselm project* at Ruhr-University Bochum, which deals with a parallel corpus of historical language data. We first present our corpus, which consists of about 50 versions of the medieval text *Interrogatio Sancti Anselmi de Passione Domini* ('Questions by Saint Anselm about the Lord's Passion'), written in different dialects from Early New High German, Middle Low German, and Middle Dutch. The versions were transcribed in a diplomatic way, and are currently being normalized and annotated with lemma and part of speech. In addition, the versions are being aligned at different levels of granularity (paragraph, sentence, phrase, word). We describe two use cases that profit from the annotations: one use case from historical lexical semantics, the other from historical syntax. We finally sketch further application scenarios from the historico-cultural domain of Digital Humanities.

KEYWORDS: Parallel corpus, Early New High German, lexical semantics, extraposition.

Stefanie Dipper and Simone Schultz-Balluff 2013. The *Anselm Corpus*: Methods and perspectives of a parallel aligned corpus. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 27–42.

1 Introduction[1]

This paper deals with the *Anselm Corpus*, a corpus consisting of more than 50 texts of the medieval tract *Interrogatio Sancti Anselmi de Passione Domini* ('Questions by Saint Anselm about the Lord's Passion'). The corpus is being created and annotated in the context of two cooperating projects from linguistics and German medieval studies at Ruhr-University Bochum (Schultz-Balluff and Dipper, 2013).[2]

The different texts are not just one-to-one copies from some source(s) but show considerable variation and, at least in parts, seem to be independent creations.[3] As a consequence, we treat all texts equally, in contrast to most other historical text editions. One of the project goals is a digital edition which gives equal access to all texts of the corpus. Users of the edition will be able to search for important concepts, such as the Last Supper, and compare the different terms used for this concept in the different texts. The edition will also support linguistically-motivated queries, e.g. for investigating the positions of verb arguments or the relative order of auxiliary–verb sequences.

This paper includes a description of the corpus and its annotations. The main focus, however, is on illustrating how this data can be exploited for different kinds of research questions. Based on a small passage, we illustrate research questions from two different areas. One research question concerns historical lexical semantics (the vocabulary), and investigates the different terms used for the concept 'Last Supper', and their temporal and regional distributions, and strategies of conceptualization. The second research question concerns historical syntax, and deals with the distribution of complements and adjuncts. In addition to these research questions from the linguistic domain, we sketch further application scenarios from the historico-cultural domain of Digital Humanities.

The paper is structured as follows. In Sec. 2, we address issues related to editions of historical texts. In Sec. 3, we describe the corpus and its annotations. Sec. 4 and 5 present the two linguistic research questions (use cases) in detail, followed by a sketch of further application scenarios in Sec. 6. Sec. 7 presents an outlook.

2 Text editions

German Medieval Studies usually have a focus on general philological rather than purely-linguistic issues. Since the 19th century, the main goal was on reconstructing the (lost) original source of a historical text. Hence, there is a long tradition in Germany for edition philologists to concentrate on the oldest witnesses of a text. Due to a paradigm shift in the late 20th century, which was initiated by the schools of *New Philology* and *New Historicism*, German Medieval Studies started to look more and more at texts other than the courtly minnesong and epic poetry. Moreover, they were now increasingly interested in the complete tradition of a text, which can span several centuries, and investigated the transmission history and genesis of a text (Bumke, 1996; Quast, 2001). The former premise—"the older the more valuable"—was deprioritized (for an overview of the history of German edition philology, see Bein (1995)).

[1] The research reported here was financed by Deutsche Forschungsgemeinschaft (DFG), Grants DI 1558/4-1 and SCHU 2524/1-1. We would like to thank the anonymous reviewers and the participants of two workshops for helpful comments.

[2] Project URLS: `http://www.ruhr-uni-bochum.de/schultz-balluff/sanktanselmus.html` and `http://www.linguistics.ruhr-uni-bochum.de/anselm`.

[3] Nevertheless, we call the corpus a *parallel corpus* because we treat the texts as parallel texts, by aligning corresponding passages across the texts, see Sec. 3.

Despite the paradigm shift, the goal of digitizing (i.e. transcribing) complete traditions has not yet been achieved because it is a highly time-consuming task.

There is a fairly complete online synopsis of the *Nibelungenlied* (‘Song of the Nibelungs’), provided by the University of Vienna[4], which comprises most of the complete and some of the fragmentary manuscripts. However, this project does not aim at diplomatic transcriptions but presents normalized versions, to facilitate readability. For instance, the transcriptions standardize capitalization and the character pairs <i>/<j> and <u>/<v>, and omit many diacritics and some superposed characters. As a consequence, the transcriptions are not suitable for linguistic research.

Another long-term enterprise at the University of Bern[5] aims at digitizing the complete tradition of the epic *Parzival* by Wolfram von Eschenbach. Sample online synopses present fragments of the epic. One of the project goals is to apply methods from *New Phylogeny* to determine stemmatic interrelations.

Other similar projects have just started. For instance, a project at the University of Cambridge will publish a complete edition of the *Kaiserchronik* (‘Chronicle of the Emperors’).

Our project follows the new paradigm in that it aims at editing all available German and Dutch witnesses of the Anselm text. It goes beyond the above-mentioned projects, though, in that we further enrich the texts by linguistic annotations, at the morphological, syntactic and semantic level, and plan to align the witnesses at different levels: at the level of sections, sentences, phrases and characters. This will enable us to perform comparative linguistic investigations, as is illustrated in this paper.

3 The corpus

Interrogatio Sancti Anselmi de Passione Domini is a tract of the passion, in the form of a dialogue. St. Anselm fasts and prays and implores Virgin Mary to reveal the events of the passion. She finally appears to him and grants his wish. He then starts asking questions, which she answers, about the Passion of Christ, beginning with the Last Supper and ending with the entombment.

The handwritten and printed documents date from the 14th–16th century and are written in a great variety of dialects in Early New High German (ENHG), i.e. Upper, Central, and Low German, and in Middle Dutch. There are 50 texts in total, with an average length of 6,000 tokens. The texts have not yet been investigated in research to any mentionable extent.

3.1 Versions

The tract has been preserved in different versions, which put emphasis on different aspects of the narration, e.g. focussing on Jesus’ sufferings or on Mary in her role as the Mother of Sorrows.[6] Based on the different foci and other general properties, the 50 texts can be grouped in 3 different versions: (i) verse versions (“V”); (ii) short prose versions (“PS”); (iii) long prose versions (“PL”).[7] They differ with regard to content and distribution:

[4] `http://germanistik.univie.ac.at/index.php?id=14531`

[5] `http://www.parzival.unibe.ch`

[6] See Schiewer (2005) for discussions and criteria of the concepts *version, copy, editing*, and *edition*.

[7] Note that the terms “short” and “long” do not refer to the texts’ length but to the texts’ content. Long versions are those that add certain specific details to the basic content.

- Verse versions (V) focus on Christ's sufferings whereas prose versions (P) focus on the sorrows of Mary.

- Since verses are written in rhyme, V-versions are rather homogeneous. In contrast, prose as a less formal text form promotes extending the basic content in various ways.

- The opening of the V-versions is very detailed and emphasizes Anselm's scariness and emotions at the moment of Mary's appearance. It includes a justification by Anselm for invoking Mary.

  PS-versions only contain the basic content; details such as Mount Sion or the Golden Gate are mentioned only in PL-versions. PL-versions also often address practical issues worth knowing, such as: What exactly are the "Ismaelitic pennies"? Why can't Mary be alone in the streets after nightfall? How big were the nails used to crucify Jesus?

- V-versions have been preserved from the north and center of Germany, PS-versions from the north and east parts of the German-speaking countries (including Austria and Switzerland). PL-versions stem from the central and southern parts. Only eastern regions produced texts of different versions.

Anselm's first question Both use cases presented in the next sections focus on Anselm's first actual question, which has been preserved in 44 German and 3 Dutch versions. In this question, Anselm asks Mary to describe the beginning of Jesus' martyrdom. Mary starts by describing the Last Supper and the betrayal of Judas. Depending on the respective version (V, PS, PL), the answer can provide further details, such as elaborate explanations of the "Ismaelitic pennies" (which Judas receives for his betrayal), or it contains supplementary elements, such as the footwashing by Jesus.

Table 1 shows Anselm's first question and the beginning of Mary's answer in a verse version, and short and long prose versions.[8]

3.2 Annotations

Normalization Currently, the texts are being annotated semi-automatically with normalized word forms, by mapping the dialect-specific historical word forms to corresponding word forms from modern German (Bollmann et al., 2011). The semi-automatic mapping first produces a simplified word form, where historic characters are replaced by their modern equivalents (e.g. *ſ* is replaced by *s*) and certain abbreviations are spelt out (e.g. ꝰ becomes *us*). In addition, since capitalization and punctuation marks are used in an inconsistent way in the Anselm texts,

[8] The versions are:

- Oldenburg ("O1"): verse version written in Low German, 2nd half of the 14th century; Landesbibliothek Oldenburg, Cim I 74.
- München ("M9"): short prose version in East Upper German, 15th century; Staatsbibliothek München, Cgm 4701.
- Wien ("W1"): long prose version in East Upper German, early 15th century; Österreichische Nationalbibliothek Wien, Cod. 2969.

In the context of the Anselm Corpus, we defined new text sigla O1, M9, W1, etc., that we use throughout this paper. A complete list of the Anselm sigles can be found at `http://www.ruhr-uni-bochum.de/schultz-balluff/sanktanselmus.html`.

Oldenburg (O1; V)	**München (M9; PS)**	**Wien (W1; PL)**
2 Maria erſt wil ik di vraghen ik bidde dattu mi willeſt ſaghen 3 Wu quam it erſt to den pranghen dat din ſone wart ge vanghen	1 Do fragt anzhelm[us] vnd ſprach 2 O aller liebſtew fraw 3 wie huͤb ſich an dez erſten deins liebn̄ chindes marter	1 Sand Anſhelm was von herczn vrö und ſprach 2 ſag mir liebe fraw 3 wie was der anfankch der marter dynes libn chindes
5 Ancelme hore dat ik di ſaghe Dat ſchude an dem guden donerſdaghe Dat he mit ſinen jungheren ſaat Lepliken dat he mit on aat He gaf on ſin vleiſch vnd ok ſin bloͮt Dat he vedder vor vns goͮt. [. . .] 6 Maria do ſe de rede dreuen Wur was judas do ge bleuen Judas de leip alto hant Dar he de Vorſten der jodden vant	4 Do ſprach ma[r]ia 5 Do mein chind an dem antloz tag daz leczt ezzen het mit ſeinē iung[er]n vnd von dem tiſch gie 6 Do gie iudas zu den iuden piſchofen	4 vnſer fraw ſprach 5 da mein libs chind het geeſſenn mit ſeinen Jungern vor ſeiner marter daz leſt mal und da ſy von tiſch auf ſtunden 6 da gieng Judas ſcarioth zw den furſten der Juden

Translation of the Wien-PL text: ‘1 St. Anselm was very glad and said: 2 Tell me, dear woman, 3 how was the beginning of the martyrdom of your dear child. 4 Our woman said: 5 As my dear child had eaten the Last Supper with his disciples before his martyrdom, and as they left the table, 6 Judas Iscariot went to the high priests of the Jews.’

Table 1: Anselm’s first question and the beginning of Mary’s answer. Excerpts from a verse version (left column), a short prose version (central column), and a long prose version (right column). Corresponding clauses in the parallel texts are annotated by the same numbers; the PL text serves as the reference text.

all letters are lower-cased and punctuation marks are removed. Next, a cascade of mapping steps are applied to create normalized (= modern) word forms (see below). Table 2 shows a fragment from an Anselm text from East Central German and its simplified and normalized equivalents.

ENHG-orig	ENHG-simpl	NORM		POS	Morph	Lemma
Do	*do*	*da*	then	ADV	–	da
ſprach	*sprach*	*sprach*	said	VVFIN	1.Sg.Past.Ind	sprechen
ſente	*sente*	*sankt*	Saint	NE	Nom.Sg.Masc	Sankt
anſhelmꝰ	*anshelmus*	*anselm*	Anselm	NE	Nom.Sg.Masc	Anselm

Table 2: A fragment from an Anselm text from Early New High German (ENHG). Column ENHG-orig shows the original word form, ENHG-simpl its simplified version and NORM the manually-normalized (modern) equivalent. Further annotations include part of speech (POS), morphology, and lemma.

The cascade of mapping steps (see Bollmann (2012) for details) starts with a word-list mapper that makes use of a lexicon of historical–modern word pairs. Second, word forms not covered by the lexicon are input to a set of character rewrite rules. The rewrite rules are derived from a set of manually-normalized word forms. Ex. (1) displays a rewrite rule that was derived from the first word pair *do–da* from Table 2.

(1) `o → a / d _ #`
('o' is replaced by 'a' between 'd' and the right word boundary ('#'))

Finally, weighted Levenshtein distance is used to map remaining historical word forms to their "closest" modern counterparts. The edit operations map sequences of up to three characters of historical to modern word forms. Weights for the edit operations are derived from the manually-normalized word forms.

Using small training corpora of 500 manually-normalized word forms, the cascade results in accuracies between 60.71% and 69.34%, depending on the individual text (see Bollmann (2012)).

POS and morphological tagging, lemmatization Normalizing the word forms results in a text that is already close to modern German. It is therefore possible to exploit POS and morphological taggers that have been trained on modern German corpora. Since the normalization procedure only produces lower-case letters and no punctuation marks, Bollmann (2012) retrained the RFTagger (Schmid and Laws, 2008) on modified versions of the TIGER corpus (Brants et al., 2004) and the TüBa-D/Z (Hinrichs et al., 2004) where all letters have been lower-cased and punctuation marks have been removed. Trained on the original data, the RFTagger achieves an accuracy of 96.85%, as compared to 95.74% when trained on the modified data. Removing punctuation especially affects the tagging of relative pronouns, which are often confused with the definite article (Bollmann, 2012).

Applying the RFTagger to the Anselm corpus, the tagger currently achieve accuracies of around 87% on manually-normalized texts and around 76% on automatically-normalized texts (Bollmann, 2012). One reason for the decrease in accuracy, as compared to tagging modern data, is the fact that the training data — which contains newspaper texts — differ considerably from the Anselm texts. For instance, the Anselm corpus contains direct speech (Anselm addressing Mary and vice versa), which frequently involves imperatives, as in *nu hor anſhelmus* 'now listen, Anselm'. In addition, many personal pronouns occur never or only rarely in the training data.

We further plan to annotate morphology and lemma information, see Table 2, as well as syntactic information.

Key words Important concepts and key words, such as *Abendmahl* 'Last Supper', *Karfreitag* 'Good Friday', will be marked in all texts, to facilitate research on these concepts.

Alignment Moreover, corresponding passages are being aligned semi-automatically across the parallel texts (Petran, 2012a). The fact that the texts do not contain punctuation marks to indicate (modern) sentence boundaries poses a special challenge (Petran, 2012b). However, automatic alignment can profit from the normalized word forms.

Semi-automatic alignments are being created at a high level between related questions asked by Anselm, and at a more fine-grained level between corresponding clauses (see the numbering in the fragments in Table 1), phrases (see Table 3) or words. The alignments represent the core annotation of our corpus. They support comparative investigations of the various texts and versions that are part of our corpus, as is illustrated in the next sections.

1. Wien (W1)	2. Halle (H1)	3. München (M4)	4. Karlsruhe (Ka1)
da mein libs chind	do mein lieber son iheſuſz	do meī kint	Do min kint
het geeſſenn	daſz nachtmal	mit ſeinē iungern̄	**hatte gezen**
mit ſeinen Jnngern	mit ſienen iüngern	**het ge eſſen**	daz ivngeſte maz
vor ſeiner marter	*am heiligen grün dornſtage*	*vor ſeiner marter*	mit sinen ivng'n
daz leſt mal	**geſſen hatte**	das iüngſt eſſen	*vor sinᷣ martᷣ*

Table 3: Phrase alignments between four PL texts. Aligned constituents which correspond to one another are highlighted in the same fashion.

Note that the annotations that the comparative investigations are based on are not yet completed at the time of writing. Hence, for the purposes of the pilot studies presented here we created the necessary annotations manually.

4 Use case I: the term and concept *Last Supper*

This section investigates the terms used for the 'Last Supper', by looking at temporal, regional, and version-related distributions and variance (cf. Besch (1967)), and by analyzing strategies of conceptualization (cf., e.g., Busse et al. (1994)). In contrast to the studies by Besch (1967), which focus on High German dialects, the Anselm Corpus also contains texts from Low German and Dutch.

The term *Abendmahl* 'Last Supper', denoting the last meal of Jesus and his disciples, is established in church language only in the early 16th century, heavily influenced by Martin Luther. Prior to that, different, but unambiguous terms had been in use to denote Jesus' farewell dinner. In the following, the different verbalizations of this concept in selected versions of the Anselm text are analyzed, focussing on their temporal and spatial distribution in relation to the three versions V, PS, and PL. The distributions of the variants are displayed in Table 4. The table displays the regions according to their actual locations, starting with Alemanic (1a) and Bavarian (1b) at the bottom (= in the south), and ending with North Low German (5) on the top (= the north).

- The table shows that in almost all **verse versions** (V, in the north: 3a–5), the Last Supper is simply described as the fact that Jesus *mit en at* 'ate with them', i.e. with his disciples.

- In the P-versions, rather fixed (but different) phrases are used. German **short prose versions** (PS, east: 1b–4b)[9] mainly use two terms: combinations of *lezt* 'last' plus *essen* 'meal' in regions 1b (Bavarian) and 2b (East Franconian), and *abent* 'evening' plus *essen* 'meal' in region 1b.

[9] PS-version D4 is from region 2a (west) and seems to be singular in several aspects.

5. North Low German

14		–
15	V	*meth em ath* (f1), *mit en at* (Kh1), *myt en at* (Arnd1494)
16	V	*mit en ath* (Arnd1521)

4a. Middle Dutch

14	PS	*auontmael* (Am1)
15	PS	*auont mael* (Le1)
16	PS	*auont maeltijt* (Berntsz1523)

3a. Ripuarian

14		–
15	V	*mit yn as* (KoeldÄ1492), *mit yn as* (KoeldJ1499)
16	V	*myt yn as* (Neuss1500), *mit yn as* (Neuss1509), *mit yn as* (Neuss1514), *myt yn as* (Neuss1514/17)

2a. Rhenish Franconian

14		–
15	PS	*nach mal* (D4)
	PL	*iungſte maſze/abend eſzen* (B2), *Jungeſte was* (St1)
16		–

1a. Alemanic

14		–
15	PL	*Iung maſz* (Be1), *ivngeſte maz* (Ka1), *Iungſt maſſ* (Stu1), *iungſt mal* (N4), *iüngſt maſz* (sa1), *iungſt <...>* (Sa1), *iüngst male* (Schau1496/97)
16	PL	*nacht mal* (SG1)

4b. West/Eastphalian

14	V	*mit em at* (D1), *mit em at* (D2), *mit on aat* (O1)
	PS	*auent ſpiſe* (Wo1)
15		–
16		–

3b. East Central German

14		–
15	V	*abint eſſin* (D3)
	PS	*obent brot* (B1)
	PL	*nachtmal* (H1)
16		–

2b. East Franconian

14		–
15	PS	*letzt eſſen* (Ba1), *leczt eſſen* (Ba2), *leczt eſſen* (N2), *leczſt eſſē* (N3)
	PL	*iüngſt eſſen* (M4), *oſterlamp* (We1)
16		–

1b. Bavarian

14	PL	*iungiſt mal* (M1)
15	PS	*leczt eſſen* (B3), *leſt eſſen* (Me1), *letz eſſen* (M5), *het geſſen* (M6), *abent eſſen* (M7), *abent eſſen* (M8), *leczt ezzen* (M9), *leczt ezzen* (M10)
	PL	*jungſte mal* (M2), *iwngiſt was* (M3), *abent eſſn̄* (Sb1), *leſt mal* (W1)
16	PL	*iungſten mal* (Hk1)

Table 4: Terms and phrases denoting the "Last Supper", used in different regions, time spans, and versions.

- Similarly, the short versions from Dutch (4a) all use combinations of *abent* ‘evening’ plus *mal(tijt)* ‘meal(time)’. Interestingly, the compound *abentmal* is already spelled in one word in the 14th century version.
- In the **long prose versions** (PL, south and center: 1–2), the combination of *iungst* ‘youngest’ plus *mal/maz* ‘meal’ is predominant, occurring in regions 1a, 1b, and 2a.
- Most other occurrences are singular, e.g. *auent spise* ‘evening dish’ (4b, PS) or *osterlamp* ‘paschal lamb’ (2b, PL).

The data shows that term selection depends on the region *and* the type of version in combination. Beyond the dominance of certain terms that we mentioned above, no continuity of terms spanning larger regions or time periods can be observed.

The variance that we observe across versions but also within the prose versions seem to suggest that at that time, no general term had yet been established. Terms used already in the 14th century continue to be used in the 15th and 16th centuries; besides them, new forms and combinations were coined.

In one PL version (B2 from region 2a), the term used for the Last Supper is explicitly addressed, see Ex. (2).

(2) Da myn kint hatte geſzen mit ſynen iungern daz iungſte maſze daz da heiſzet daz abend eſzen
‘As my child had eaten with his disciples the youngest meal which is called the evening meal’

Iungſte maſze ‘youngest meal’ is probably a general term, whereas *abend eſzen* ‘evening meal’ seems to be a more special term, highlighted by the author. However, *abend eſzen* is not a fixed term as can be seen from the variance observed in regions 1b and 3b in the 15th century.

The unsteadiness of the terms is also reflected by the fact that most instantiations are spelled in two words, and only few “real” compounds can be observed: *auontmal* (Am1, 4a) and *nachtmal* (H1, 3b), from the 14th and late 15th centuries, respectively. Moreover, these compounds reoccur later, but spelled in two words (in Le1, 4a, and SG1, 1a).

It is remarkable that the term that has finally been established in standard German is the term from Middle Dutch.

Strategies of specification As we have seen, the prose versions, short and long, seem to struggle for verbalizing the concept *Last Supper* but do not arrive at a common, “standardized” term. The verse versions follow another strategy: they use the unspecific phrase *dat he mit on aat* ‘that he ate with them’ but add specific temporal information when this happened: *an dem guden donerſdaghe* ‘on the good Thursday’ (O1).[10]

A similar specification strategy is also followed by some of the prose versions. 11 PS versions[11] add *an dem antlaz tag* ‘on the indulgence day’. *Antlaz tag* in general means ‘day of

[10] D3 (written in East Central German) represents a special case: It uses rhymes but otherwise shows characteristics of the prose versions. Especially its vocabulary deviates from the other verse versions. This suggests that D3 should be considered separately from the verse versions, and in connection with the prose versions.

[11] Texts B3, Ba1, Ba2, M5, M7, M8, M9, M10, Me1, N2, N3.

release/indulgence', and it can be used to refer to Holy Thursday in particular. Two PL versions (H1, SG1), which use the term *nachtmal* 'night meal', add the specifications *am heiligen gründornſtage* (H1), *am hailgen gruͤnen donstag* (SG1) 'on the Holy Thursday'.

This data shows that the fact that there is not yet a mandatory agreed-upon term is compensated by specification strategies. We propose that the specific strategy used in a subset of the versions can be used as a defining criterion for the version *verse*.

5 Use case II: constituents in situ and extraposed

In the second use case, we select the first sentence of Mary's first answer and compare its different syntactic realizations in all PL versions of our corpus (20 versions in total). In particular, we investigate the positions of verb arguments and adjuncts. Note that this pilot study has been done by aligning the data manually and assigning syntactic functions and positions by hand. In the long run, we plan to do quantitative investigations that cover the entire corpus, by relying on the semi-automatically assigned POS tags and syntactic annotations, and exploiting the alignments at the phrasal level.

In (modern) German, the *right verbal bracket* indicates the boundary between nominal and prepositional arguments and adjuncts that occur *in situ* (preceding the right verbal bracket) or *extraposed* (following the verbal bracket).

In subordinate clauses, the right verbal bracket is filled by verbal components (verbs and auxiliaries), see Ex. (3a). In main clauses, the finite verb or auxiliary takes the second position after some other constituent, filling the *left verbal bracket* (this construction is called *verb-second*). Further verbal components, such as infinite verb forms, verb particles, can occupy the right verbal bracket, see Ex. (3b).[12] The left and right verbal brackets are underlined in the examples. Constituents occurring in situ are marked by "INS", extraposed constituents by "EX". The examples illustrate that in modern German, arguments such as the subject and object occur in situ, whereas adjuncts can be extraposed (optionally).

(3)	a.	*als*	$[_{INS}$	*Jesus]*	$[_{INS}$	*das Abendmahl]*	*gegessen hatte*	$[_{EX}$	*mit*	*seinen Jüngern]*
		as		Jesus		the Last_Supper	eaten had		with	his disciples

'as Jesus had eaten the Last Supper with his disciples'

b.	*Jesus*	*hatte*	$[_{INS}$	*das Abendmahl]*	*gegessen*	$[_{EX}$	*mit seinen Jüngern]*
	Jesus	had		the Last_Supper	eaten		with his disciples

'Jesus had eaten the Last Supper with his disciples'

The verb-second pattern can already been observed in Old High German, in addition to verb-first, verb-third, and verb-final patterns. In Middle High German, the verb-second pattern has been established as the common structure of main clauses. Verb-final patterns in subordinate clauses are predominant from the earliest stages on. However, as can be seen from Ex. (3a), extraposed constituents can occur after the final verb.

It is well known that arguments and adjuncts occurred in extraposed positions much more frequently in older language stages than nowadays.

Based on data from Gothic, Old English, and different stages from German, Behaghel (1932) shows that short constituents, consisting of one word, predominantly occur in situ, whereas

[12] For a description of the German sentence structure, e.g. see Höhle (1986).

long, “heavy” constituents, e.g. constituents involving coordination, tend to be extraposed. Ebert (1986) examines two texts from the 14th century and finds that around 20% of subordinate clauses contain extraposed constituents, predominantly PPs, but also NP complements.[13]

In the 17th century, the sentence-final position of the verb in subordinate clauses has been established in standard language (Behaghel, 1932, p. 133). That is, since that time, extraposition is limited to clausal arguments and PP adjuncts.

Ex. (4), taken from Behaghel (1932, p. 132), shows an example from Martin Luther with an extraposed object. This construction would be highly marked in modern standard German.

(4) *wenn du erkenntest [$_{EX}$ die Gabe Gottes und wer der ist, der zu dir sagt, gib mir trinken]*
‘if you knew the gift of God and who it is that asks you for a drink’ (John 4:10)

The presentation above shows that in Early New High German, extraposition is still applied to a range of arguments. Hence, it is interesting to investigate the amount of extraposition and the type of arguments that are extraposed in the different Anselm texts. To do this, we analyse the first sentence of Mary’s first answer in detail, see Table 5. The table displays the W1-text in the first column, organized by constituents, a translation in the second column, each constituent’s function in the third column, and its position in the forth column. The sentence consists of two subordinate clauses, followed by the main clause.

As can be seen from the table, the three subjects occur in situ. The remaining constituents of the first subordinate clause are extraposed, in contrast to the constituents of the second subordinate clause. The positions of the main clause constituents cannot be determined in this example because the right verbal bracket is not filled (but see below).

The distribution of the constituents, as realized in this text, is in fact the “default” distribution, which shows up in 11 of the 20 PL texts.[14] In three texts,[15] all constituents occur in situ. Interestingly, these texts share another unique feature: the verbal components of the first clause (line 3) show the modern order *verb participle > finite auxiliary*, e.g. *geſſen hett* ‘eaten had’, in contrast to all other texts.

In 14 texts, the NP-object of the first clause (line 6) occurs after the PP-adjuncts (lines 4 and 5). In five texts,[16] the NP-object occurs in front of the PP-adjuncts, and in one text,[17] it occurs between both PPs.

In three texts,[18] the right verbal bracket of the main clause is filled by a verb particle. In these cases, the locative PP, which denotes the goal of the movement (line 15), is extraposed, see Ex. (5).

[13] The terms *in situ* and *extraposed* suggest that one of the positions is the “original”, unmarked one, while the other is a secondary position, derived from the first, e.g., by a relation called “extraposition”. For modern German, the unmarked positions of NP and PP constituents are clearly in front of the rigth verbal bracket, and positions behind the right verbal bracket are exceptional. In former stages of German, however, the situation is not as clear. Hence, the reader is asked to interpret the terms *in situ* and *extraposed* as referring to pre- and postverbal positions, without implications about the actual analysis.

[14] From region 1a: N4, Stu1, Schau1496/97; region 1b: M1, M2, M3, W1, Hk1; region 2a: B2, St1; region 2b: We1.

[15] From region 1a: SG1; region 1b: Sb1; region 3b: H1.

[16] Region 1a: Be1, Ka1, sa1, SG1; region 3b: H1.

[17] Region 1a: Sa1.

[18] All from region 1a: Be1, Ka1, Sa1

	Wien ("W1", PL)		**Function**	**Position**	**Clause**
1	*da*	as	subord		
2	*mein libs chind*	my dear child	NP-subj	INS	Subord 1 (1–6)
3	*het geeffenn*	has eaten	verb	right VB	
4	*mit ſeinen Jungern*	with his disciples	PP-adjunct	EX	
5	*vor ſeiner marter*	before his martyrdom	PP-adjunct	EX	
6	*daz leſt mal*	the Last Supper	NP-obj	EX	
7	*und*	and	coord		
8	*da*	as	subord		Subord 2 (8–11)
9	*ſy*	they	NP-subj	INS	
10	*von tiſch*	off table	PP-adjunct	INS	
11	*auf ſtunden*	up stood	verb	right VB	
12	*da*	then	adverb		
13	*gieng*	went	verb	left VB	Main (12–15)
14	*Judas ſcarioth*	Judas Iscariot	NP-subj	?	
15	*zw den furſten der Juden*	to the princes of_the Jews	PP-goal	?	

Table 5: The beginning of Mary's first answer: 'As my dear child had eaten the Last Supper with his disciples before his martyrdom, and as they left the table, Judas Iscariot went to the high priests of the Jews'. INS: in situ, EX: extraposed, VB: verbal bracket.

(5) *Do giench [$_{INS}$ ivdas ſcarioth] vz. [$_{EX}$ zv den fv̄rſten d͛ ivden].*
then went Judas Iscariot out to the princes of_the Jews
'Then Judas went out to the high priests of the Jews'

Further differences between the texts include: absence of the second subordinating conjunction (line 8); absence of the subject in the second subordinate clause (line 9).

To sum up the findings of this small comparison, we have seen that extraposition of the object NP seems to be the unmarked case, in contrast to modern German. Ignoring the case of subject NPs (which seem to be extraposed only rarely) and unclear positions, the numbers of constituents in situ vs. extraposed are almost equal in the default order: two PPs and one object NP occur extraposed, two PPs are in situ. Our small study seems to indicate that the Anselm texts exhibit considerably more extraposed constituents than the texts examined by Ebert (1986), possibly due to the fact that the Anselm texts contain direct speech. Fully-annotated texts will allow us to investigate such questions to a greater extent and in more detail.

6 Further application scenarios

The use cases presented in Sec. 4 and 5 dealt with linguistic issues. In this section, we want to sketch the use of the Anselm Corpus and its annotations for research in other fields of the humanities, such as reception history or narratology. In particular, we show that Mary and Jesus are addressed and referenced in different ways in the individual texts. Taken together with other characteristics of the texts, we can deduce from the different ways of reference that the texts were composed for different groups of recipients.

Forms of address The forms of address for **Mary** vary considerably between the different texts. In the prose versions, Anselm addresses Mary by forms indicating devotion: *leue vrowe* ‘dear woman’ (Wo1), or *aller liebſtew fraw* ‘most dearest woman’ (M9). When talking about Jesus, Anselm emphasizes Mary’s role of the mother: *deins liben kindes* ‘your dear child’ (N3). This culminates in the description of Mary as the mother of all humans: *liebe mutt[er]* ‘dear mother’ (form used by Anselm, H1) or as the Mother of God: *Mutt[er] gotteſz* (form used by the narrator, H1). Finally, the narrators of the prose texts involve the recipient, e.g. by phrases such as *vnſer fraw* ‘our woman’ (e.g. W1) or *vnſe liebe frauwe* ‘our dear women’ (e.g. B2).

In the verse versions, the relation to Mary remains more reserved. She is addressed exclusively by her name. Still, her mother role is present in that Jesus is referenced by *din ſon* ‘your son’.

When talking about **Jesus**, Mary refers to him by *min kint* ‘my child’ (Be1), *mein libs kint* ‘my dear child’ (W1) or even *min alre lieffte kint* ‘my most dearest child’ (Am1). In the verse versions, Jesus is predominantly referred to by the personal pronoun.

In short, the prose versions emphasize the relation mother–son, described from the point of view of the sorrowing mother, and establish a mother relation between Mary and the recipient of the text. The personal relationship is intensified by elaborate passages of lamentation. The idea of compassion and the role of Mary as Mater Dolorosa plays a stronger role in the prose than in the verse versions, which remain more distant in general.

Incipit and explicit The forms of the incipits and explicits (the opening and closing words of the texts) also show that the texts were composed and intended for different groups of users. For instance, the explicit of the text Wo1 begins with the words *Swe dit het lest de vordenet seszdusentseshundertvndsesvndsestindich iar afflates* ‘whoever has read this deserves 6,666 years of indulgence’. This is followed by a compact depiction of Jesus’ sufferings. The text ends with the words *des danke ic di here* ‘I thank you for that, lord’, emphasizing the prayerlike style of the explicit. One can suppose, therefore, that the text served purposes of intensive internalization and spiritualization, and could have been used in a monastic context. In fact, the text is part of a larger manuscript consisting of multiple individual texts; the composition of this manuscript also supports the presumption that it was used in a spiritual, religious context, presumably in a nunnery in the Eastphalian region (Schultz-Balluff, 2013).

As we have just seen, the recipient of the text Wo1 is actively involved. The explicit of the text Kh1 shows a different picture. The text ends with a plea to God to send peace to the human beings, and states that everyone who would not enjoy the tract should remain a fool forever: *Dyt is sunte ansylmus vrage/ weme se nicht en behage/ De blyue en schalk al syne daghe* ‘This is St. Anselm’s question. Who does not like it remains a fool for the rest of his life’.

According to some of the texts, reading *St Anselms Question’s about the Lord’s Passion* can also be helpful for the everyday life. For instance, it facilitates childbirth (M8): *den frawn̄ die do ſchwanger ſein vnd ſwarlig kinder gepern den iſt dyſz büchlen alſz nuͤz als ob ſy andere ding theten dy den frawn̄ hylff geben* ‘For pregnant women who have difficulties in bearing children this book is beneficial just like other things that might help women in such a situation’. Similarly, reading the tract can protect the house against harm caused by water or other disaster (M8): *in welgē hauſz das buch mit andacht wurt geleſen vnd in welgē haüſz es iſt dem ſelbn̄ hauſz kan kain waſſer ader kain vngehewr geſchadn̄* ‘Houses in which this book is read with devotion, and houses in which it is present, cannot be harmed by water or disaster’.

These examples illustrate that the tract in its different forms served different kinds of recipients:

in monasteries and convents, other clerical circles, and also in the secular part of society.

To sum up, the forms of address, style, and differences in content allow us to draw conclusions with regard to the context of use of individual texts or versions, and to the image of Mary and the envisaged recipients.

7 Conclusion

In this paper, we presented a parallel corpus of texts from Early New High German and Middle Low German. We argued that alignments at different levels (question–answer pairs, sentences, phrases, words) can support comparative investigations in different areas. This was illustrated by different use cases from historical lexical semantics (comparing terms used for the Last Supper), historical syntax (comparing the distribution of constituents), and from a historico-cultural perspective (comparing the ways Mary and Jesus are addressed and referenced, and the specific forms of the incipits and explicits). The last scenarios showed that linguistic annotations can also benefit non-linguistic research in Digital Humanities.

The (manually-created) annotations that were made use of in the pilot studies are:

- Use case 1 ("Last Supper"): alignment at paragraph level (Anselm's questions) and key words
- Use case 2 (extraposition): alignment at chunk/phrase level
- Further application scenario 1 (forms of address): key words
- Further application scenario 2 (incipit and explicit): alignment at paragraph level

We plan to create a digital edition of the entire corpus (Stolz et al., 2007). Users will be able to select texts from the collection and search for specific word forms, parts of speech etc. The query results will be presented in the form of a synopsis, which places aligned passages next to each other.

We think that the alignments can also support semi-automatic creation of a critical apparatus used in a print edition. The variance observed between the three versions (verse, short prose, long prose) suggests that all three versions would be edited. The variance could also lead to considerations whether we actually deal with one or three texts.

The Anselm corpus is especially well suited for a pilot project exploring comparative linguistic research based on an aligned parallel corpus. The Anselm texts are rather short but written in a great variety of dialects. In future work, we would like to apply our method to other texts with many surviving witnesses, such das *Die 24 Alten* from Otto von Passau ('The 24 Elders'; 142 witnesses, among them 11 fragments) or *Unser vrouwen klage* ('Mary's lamentations'; 26 witnesses, 4 fragments). Next, we would like to look at other types of texts, e.g. pharmacopoeias, which have a long and broad tradition throughout the entire Middle Ages, e.g. *Bartholomäus* (with 32 witnesses, 6 fragments).

The long-term goal would be to analyze long texts using our methods, such as *Parzival, the Song of the Nibelungs* or *the Chronicle of the Emperors*.

References

Behaghel, O. (1932). *Deutsche Syntax. Eine geschichtliche Darstellung. Band IV: Wortstellung. Periodenbau*. Winter, Heidelberg.

Bein, T., editor (1995). *Altgermanistische Editionswissenschaft*. Peter Lang, Frankfurt/Main, New York.

Besch, W. (1967). *Sprachlandschaften und Sprachausgleich im 15. Jahrhundert. Studien zur Erforschung der spätmittelhochdeutschen Schreibdialekte und zur Entstehung der neuhochdeutschen Schriftsprache*. Francke, München.

Bollmann, M. (2012). Automatic normalization for linguistic annotation of historical language data. Master's thesis, Ruhr-Universität Bochum.

Bollmann, M., Petran, F., and Dipper, S. (2011). Applying rule-based normalization to different types of historical texts — an evaluation. In Vetulani, Z., editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 339–344, Poznan, Poland.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Bumke, J. (1996). Der unfeste Text. Überlegungen zur Überlieferungegeschichte und Textkritik der höfischen Epik im 13. Jahrhundert. In Müller, J.-D., editor, *Aufführung und Schrift in Mittelalter und Früher Neuzeit*. Metzler, Stuttgart, Weimar.

Busse, D., Hermanns, F., and Teubert, W., editors (1994). *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik*. Westdeutscher Verlag, Opladen.

Ebert, R. P. (1986). *Historische Syntax des Deutschen II: 1300-1750*. Peter Lang, Frankfurt.

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and Trushkina, J. (2004). Recent developments in linguistic annotations of the TüBa-D/Z Treebank. In *Proceedings of TLT 2004*, Tübingen, Germany.

Höhle, T. (1986). Der Begriff 'Mittelfeld'. Anmerkungen zur Theorie der topologischen Felder. In Schöne, A., editor, *Kontroversen, neue und alte. Akten des 7. Internationalen Germanistenkongresses Göttingen 1985*, pages 329–340. Niemeyer, Tübingen.

Petran, F. (2012a). Aligning the un-alignable — a pilot study using a noisy corpus of nonstandardized, semi-parallel texts. In Gelbkuh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 2. Springer, Berlin, Heidelberg.

Petran, F. (2012b). Studies for segmentation of historical texts: Sentences or chunks? In *Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), 2012*, Lisbon, Portugal.

Quast, B. (2001). Der feste Text. Beobachtungen zur Beweglichkeit des Textes aus Sicht der Produzenten. In Peters, U., editor, *Text und Kultur. Mittelalterliche Literatur 1150–1450*. Metzler, Stuttgart, Weimar.

Schiewer, H. J. (2005). Fassung, Bearbeitung, Version und Edition. In Schubert, M. J., editor, *Deutsche Texte des Mittelalters zwischen Handschriftennähe und Rekonstruktion. Berliner Fachtagung 1.-3. April 2004*. de Gruyter, Tübingen.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, Manchester, UK.

Schultz-Balluff, S. (2013). Auf dem Wandbord einer Nonne — Ein Passionstraktat in täglichem Gebrauch. In *Rosenkränze und Seelengärten. Bildung und Frömmigkeit in niedersächsischen Frauenklöstern*, Ausstellungskataloge der Herzog August Bibliothek 95, pages 147–155. Herzog August Bibliothek Wolfenbüttel, Wiesbaden.

Schultz-Balluff, S. and Dipper, S. (2013). 'St. Anselmi Fragen an Maria' — Schritte zu einer (digitalen) Erschließung, Auswertung und Edition der gesamten deutschsprachigen Überlieferung (14.–16. Jh.). In Bohnenkamp-Renken, A., editor, *Medienwandel/Medienwechsel in der Editionswissenschaft*, Beihefte zu editio, pages 173–191. Berlin, Boston: de Gruyter.

Stolz, M., Lucas, M., and Loop, J., editors (2007). *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien. Eine Standortbestimmung*. germanistik.ch.