RUHR-UNIVERSITÄT BOCHUM

RUB

Fakultät für Philologie
Sprachwissenschaftliches Institut

THE ANSELM PROJECT

Tools for Automatic Analysis of a Parallel Corpus in Early New High German

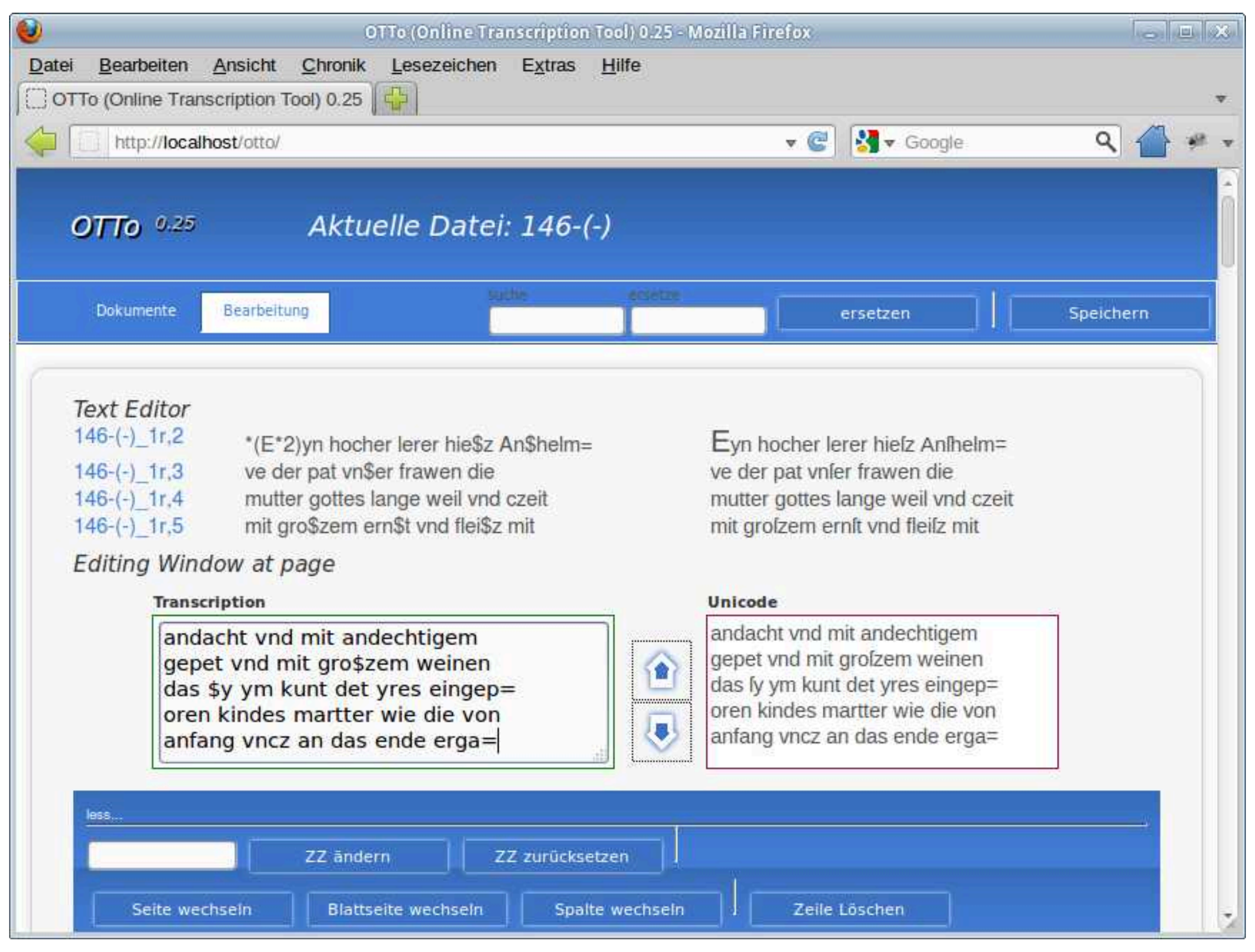
Marcel Bollmann, Stefanie Dipper, Julia Krasselt, Florian Petran
{bollmann,dipper,krasselt,petran}@linguistics.rub.de

The Project

In the Anselm project, we develop tools for the processing and automatic analysis of the medieval German treatise "St. Anselmi Fragen an Maria" (St. Anselm's questions to Mary), in collaboration with historical linguists from Ruhr-Universität Bochum.

Goals:

- Diplomatic transcription of all residual manuscripts and printings in German
- Annotation with bibliographic and linguistic information, e.g. POS tagging
- Alignment on paragraph, sentence, and word level
- Linguistic analysis of dialectal variation
- Publication of a digital edition



Aktuelle Datei: 146-(-)

Text Editor

146-(-)_1r,2	*(E*2)yn hocher lerer hie$z An$helm=	Eyn hocher lerer hieſz Anſhelm=
146-(-)_1r,3	ve der pat vn$er frawen die	ve der pat vnſer frawen die
146-(-)_1r,4	mutter gottes lange weil vnd czeit	mutter gottes lange weil vnd czeit
146-(-)_1r,5	mit gro$zem ern$t vnd flei$z mit	mit groſzem ernſt vnd fleiſz mit

Editing Window at page

Transcription

```
andacht vnd mit andechtigem
gepet vnd mit gro$zem weinen
das $y ym kunt det yres eingep=
oren kindes martter wie die von
anfang vncz an das ende erga=
```

Unicode

```
andacht vnd mit andechtigem
gepet vnd mit groſzem weinen
das ſy ym kunt det yres eingep=
oren kindes martter wie die von
anfang vncz an das ende erga=
```

Screenshot of OTTo (Online Transcription Tool)

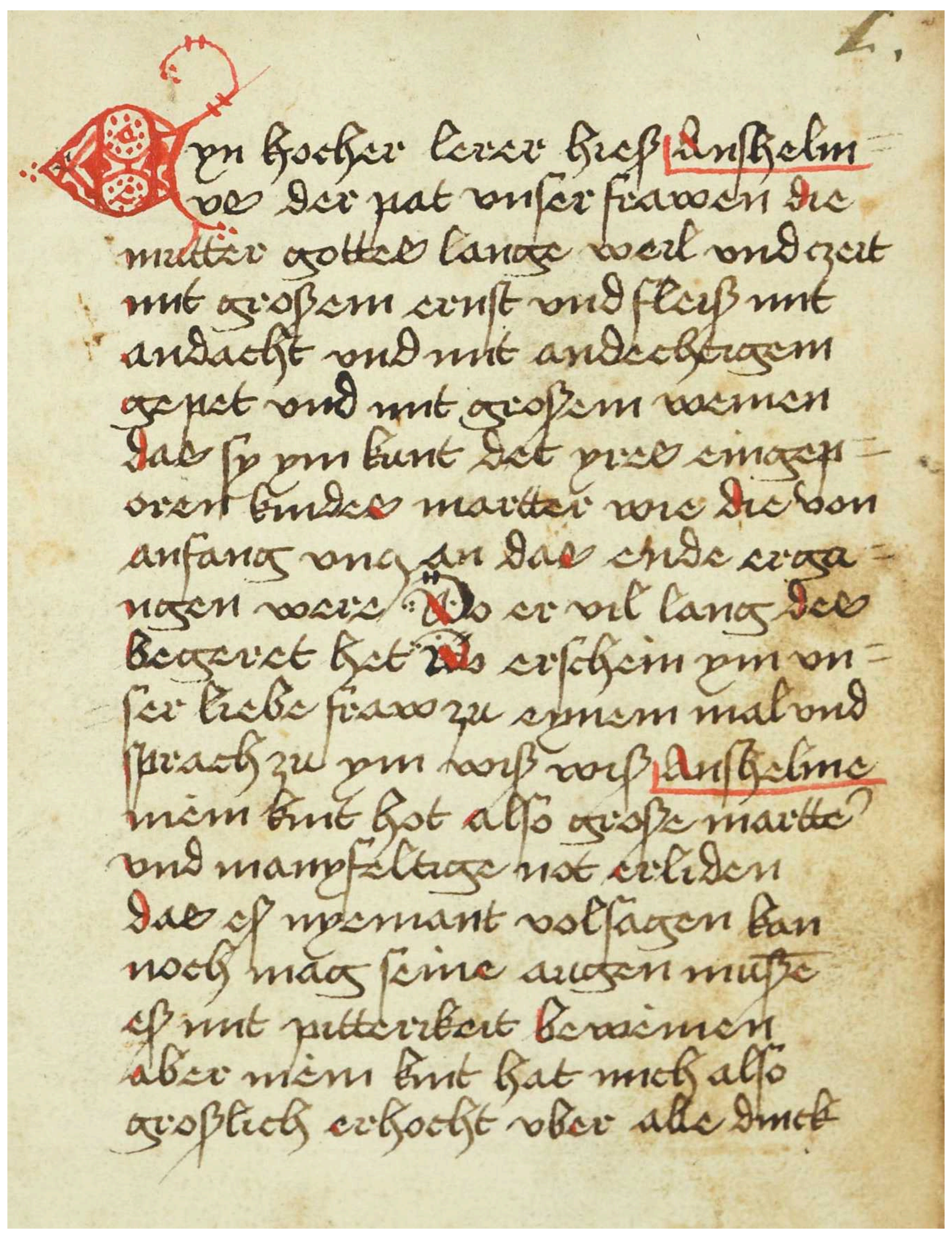
Transcription

- Diplomatic transcription with twofold collation
- Medieval texts use spelling conventions no longer in use today, e.g. long s ('ſ') or macrons ('ē')
- Graphematic peculiarities are potentially of interest for users of the corpus (e.g. for dialectal analysis) and should therefore be retained in transcriptions

The Corpus

In the texts, Anselm of Canterbury asks questions to the Virgin Mary concerning the Passion of Jesus Christ. She answers him in the form of longer monologues.

- 44 German manuscripts and prints
- Written between 14th to 16th centuries, in Early New High German (ENHG)
- Verse and prose versions of different lengths
- Comparable content, logical structure, and even (semi-)parallelity in sentence structure and wording
- No fixed spelling conventions
- Dialectal variations in graphematics, phonology, morphology, and syntax



1.

Eyn hocher lerer hieß Anſhelm
ve der pat vnſer frawen die
mutter gottes lange weil vnd czeit
mit großem ernſt vnd fleiß mit
andacht vnd mit andechtigem
gepet vnd mit großem weinen
das ſy ym kunt det yres eingep=
oren kindes martter wie die von
anfang vncz an das ende erga=
ngen were. Do er vil lang des
begeret het Do erſchein ym vn=
ſer liebe fraw zu eynem mal vnd
ſprach zu ym weiß weiß Anſhelme
mein kint hot alſo groſze martter
vnd manyfeltige not erliden
das eſ nyemant volſagen kan
noch mag ſeine augen muſze
eß mit pitterkeit beweinen
aber mein kint hat mich alſo
groſlich erhocht vber alle dinck

Sample from an Anselm manuscript (Upper German dialect), Weimar, Herzogin Anna Amalia Bibliothek, Cod. Oct. 4, fol.1r

The Tools

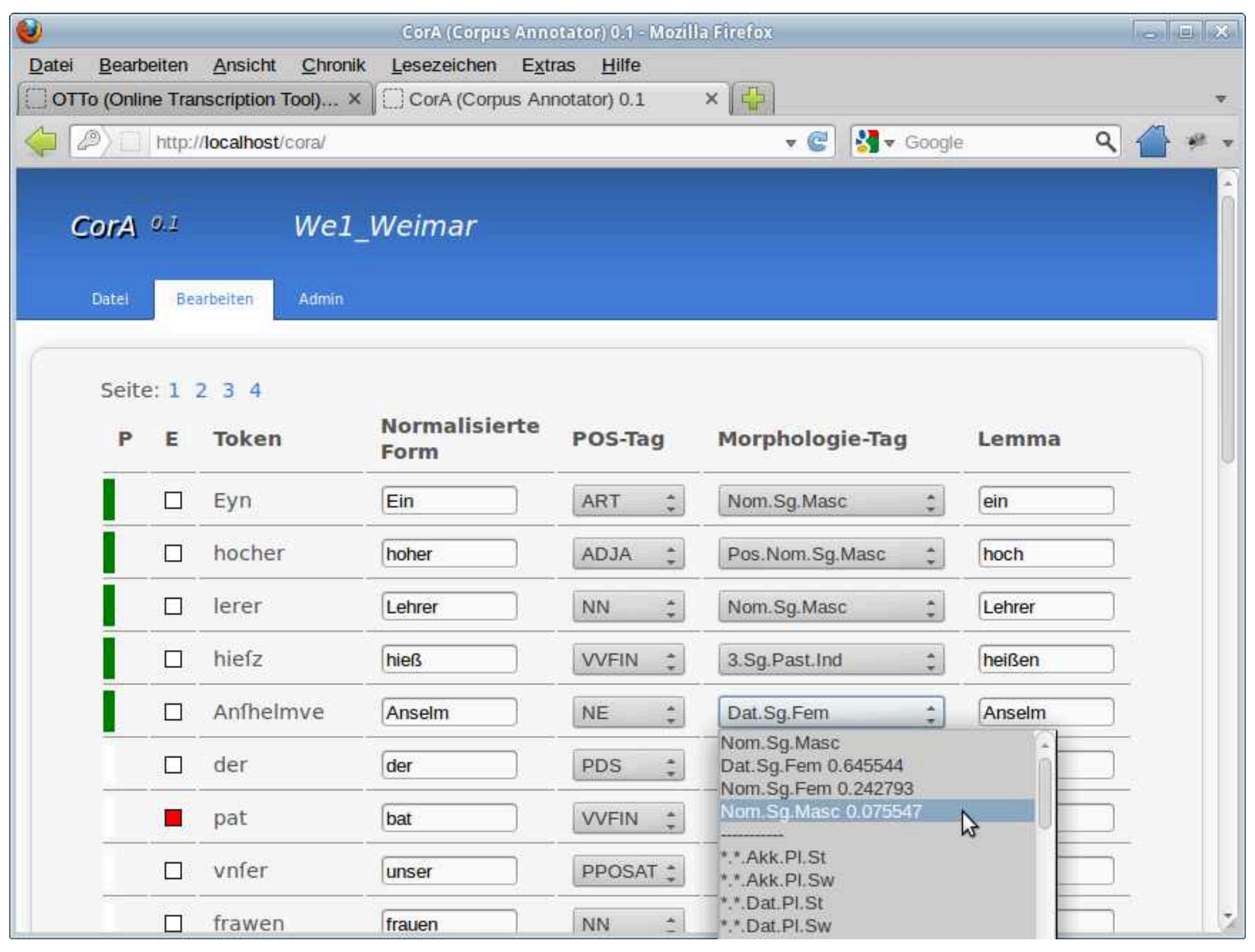
- Browser-based
  - No installation required by the user
  - Centrally maintained

OTTo (Online Transcription Tool)

- Aids the diplomatic transcription of documents
- Special characters can be represented with user-definable escape sequences (e.g. '$' for 'ſ')
- Visual feedback: escape sequences are converted to Unicode characters and displayed instantly

CorA (Corpus Annotator)

- Supports semi-automatic normalizations and annotations of various types



CorA (Corpus Annotator) 0.1 - Mozilla Firefox
Datei
Bearbeiten
Ansicht
Chronik
Lesezeichen
Extras
Hilfe
OTTo (Online Transcription Tool)...
CorA (Corpus Annotator) 0.1
http://localhost/cora/
Google
CorA 0.1
Datei
Bearbeiten
Admin

We1_Weimar

Seite: 1 2 3 4

P	E	Token	Normalisierte Form	POS-Tag	Morphologie-Tag	Lemma
🟩	□	Eyn	Ein	ART	Nom.Sg.Masc	ein
🟩	□	hocher	hoher	ADJA	Pos.Nom.Sg.Masc	hoch
🟩	□	lerer	Lehrer	NN	Nom.Sg.Masc	Lehrer
🟩	□	hieſz	hieß	VVFIN	3.Sg.Past.Ind	heißen
🟩	□	Anſhelmve	Anselm	NE	Dat.Sg.Fem	Anselm
	□	der	der	PDS		
	🟥	pat	bat	VVFIN		
	□	vnſer	unser	PPOSAT		
	□	frawen	frauen	NN		

```
Nom.Sg.Masc
Dat.Sg.Fem 0.645544
Nom.Sg.Fem 0.242793
Nom.Sg.Masc 0.075547
------------
*.*.Akk.Pl.St
*.*.Akk.Pl.Sw
*.*.Dat.Pl.St
*.*.Dat.Pl.Sw
```

Screenshot of CorA (Corpus Annotator)

Alignment

As the texts are semi-parallel, alignment is often possible even at the word level.

Purpose:

- Projection of normalizations and annotations from one text over the whole collection
- Comparative linguistic and literary studies on the texts

Problems for traditional alignment methods:

- Missing or inserted sections
- Spelling differences hinder recognition of corresponding wordforms
- No sentence boundary markings

Solution:

- Extract a cognate dictionary
- Construct alignment sequences from closely occuring translation pairs
- Gradually expand the sequences to find the longest possible subsequence of alignment pairs
- Assign a confidence score and keep only the highest scoring sequence in case of conflict

Although it is a work in progress, this approach already performs better than traditional alignment methods based on heuristic text segmentation.

Normalization

We map ENHG wordforms to New High German (NHG) wordforms, with the idea of ...

- applying already existing annotation tools for German (e.g. TreeTagger for POS tagging) to our data;
- facilitating the alignment process by standardizing the spelling; and
- measuring the distance between old and modern wordforms for purposes of linguistic analysis.

Manual

- Manual normalization done by student assistants for evaluation purposes
- Two levels of normalization are distinguished:
  1) Modern wordform that is closest to the historical form ('vnſer frawen' → 'unser Frauen')
  2) Wordform adjusted with regard to inflection and semantics to form grammatical NHG phrases ('vnſer frawen' → 'unsere Frau')

Automatic

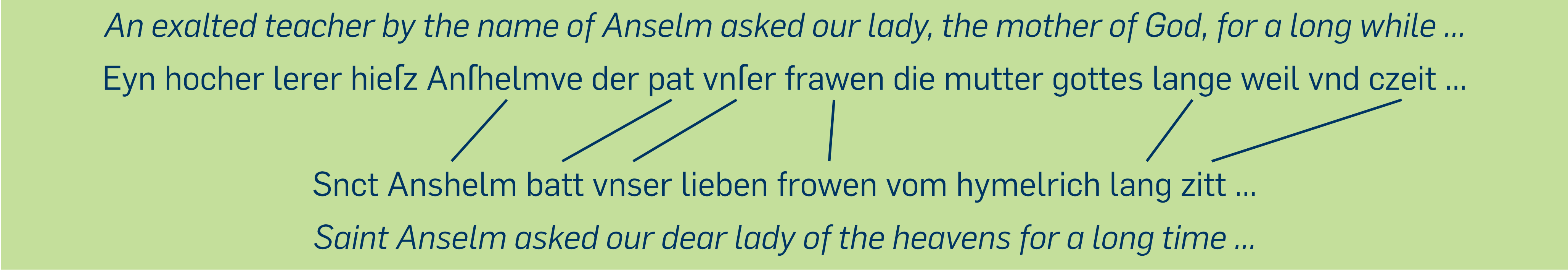
- Idea: Learn characteristic spelling variations
- Character rewrite rules, sensitive to context

v → u / # _ n

(historic 'v' becomes modern 'u' between a word boundary and 'n', e.g. in the mapping 'vnd' → 'und')

- Rules learned from an aligned Luther bible corpus
  - Based on the 1545 version and a modernized equivalent from 1892, both freely available
  - Alignment on paragraph, sentence, and word level
- Evaluation showed performance superior to a pure dictionary-based word substitution method
- Example mappings generated by our method:

etleich	→	*etliche*	
vnse	→	*unser*	
zitt	→	**zittern*	(instead of the correct 'zeit')



An exalted teacher by the name of Anselm asked our lady, the mother of God, for a long while ...

Eyn hocher lerer hieſz Anſhelmve der pat vnſer frawen die mutter gottes lange weil vnd czeit ...

Snct Anshelm batt vnser lieben frowen vom hymelrich lang zitt ...

Saint Anselm asked our dear lady of the heavens for a long time ...

Sample passage from two different manuscripts, illustrating word alignment

- *OTTo: Online Transcription Tool.* http://www.linguistics.rub.de/otto/.
- Florian Petran (2012). *Aligning the un-alignable — a pilot study using a noisy corpus of nonstandardized, semi-parallel texts.* In: Alexander Gelbukh (ed.). Computational Linguistics and Intelligent Text Processing, Vol. 2. Springer: Berlin/Heidelberg.
- Marcel Bollmann, Florian Petran, and Stefanie Dipper (2011). *Applying Rule-Based Normalization to Different Types of Historical Texts — An Evaluation.* In: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznań, Poland.
- Marcel Bollmann, Florian Petran, and Stefanie Dipper (2011). *Rule-Based Normalization of Historical Texts.* In: Proceedings of the RANLP Workshop on Language Technologies for Digital Humanities and Cultural Heritage. Hissar, Bulgaria.

This project is supported by Deutsche Forschungsgemeinschaft, Grant DI 1558/4-1.

DFG