RUHR-UNIVERSITÄT BOCHUM

Fakultät für Philologie
Sprachwissenschaftliches Institut

CorA: A web-based annotation tool for historical and other non-standard language data

Marcel Bollmann, Florian Petran, Stefanie Dipper, Julia Krasselt

Motivation & Purpose

CorA (short for "Corpus Annotator") is a web-based annotation tool. It is designed for the annotation of non-standard language data, e.g., historical texts, computer-mediated communication (CMC), or texts from language learners.

Tokenization is often problematic with these types of texts. For historical text, there can be OCR or transcription errors that are only detected during annotation. Also, automatic (pre-)annotation is often difficult due to the sparseness of training data.

CorA tries to alleviate these issues by providing options to edit the source document during annotation (including tokenization changes). Annotation software can be integrated, with options for retraining on newly annotated passages, and reannotating passages in already imported texts.

We plan to make CorA available as open source, including an English user interface. CorA is based on JavaScript, PHP 5, and MySQL.

Features

- Web-based
- XML-based import & export
- CorA XML can be imported into the ANNIS corpus search tool (Zeldes et al., 2009)

Annotation Types

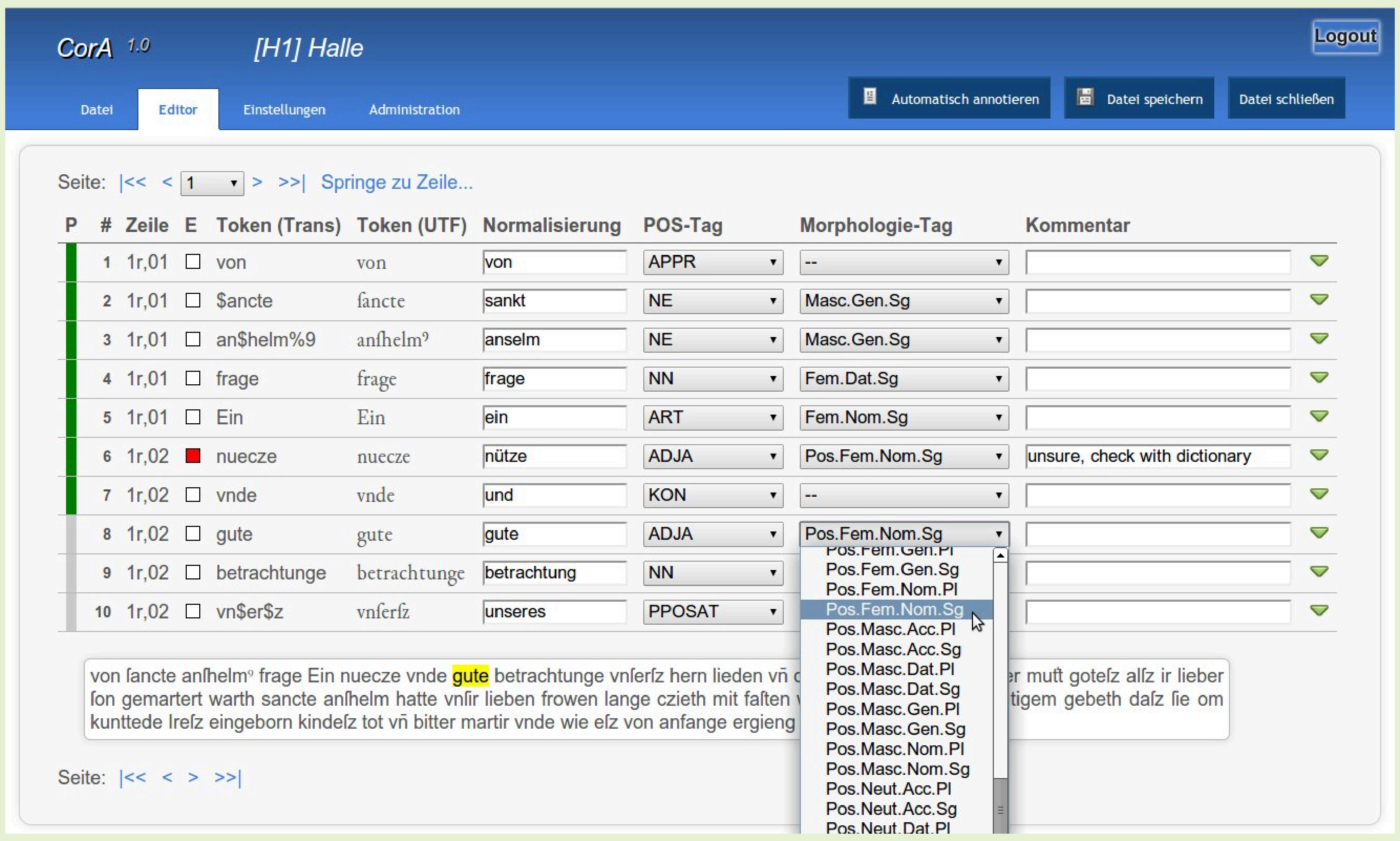
- Part-of-speech and morphological tags
- Lemmatization
- Lemma part-of-speech
- Normalization
- Modernization and modernization category

Tagger Integration

- Integration of external annotation tools (e.g., POS taggers)
- Call taggers directly from the web interface
- Retraining taggers on existing annotations within the same project

Editing the Document

- Edit, add, or delete tokens in the source document
- Custom embedded scripts can perform validity checks, UTF-8 conversion, tokenization, etc.



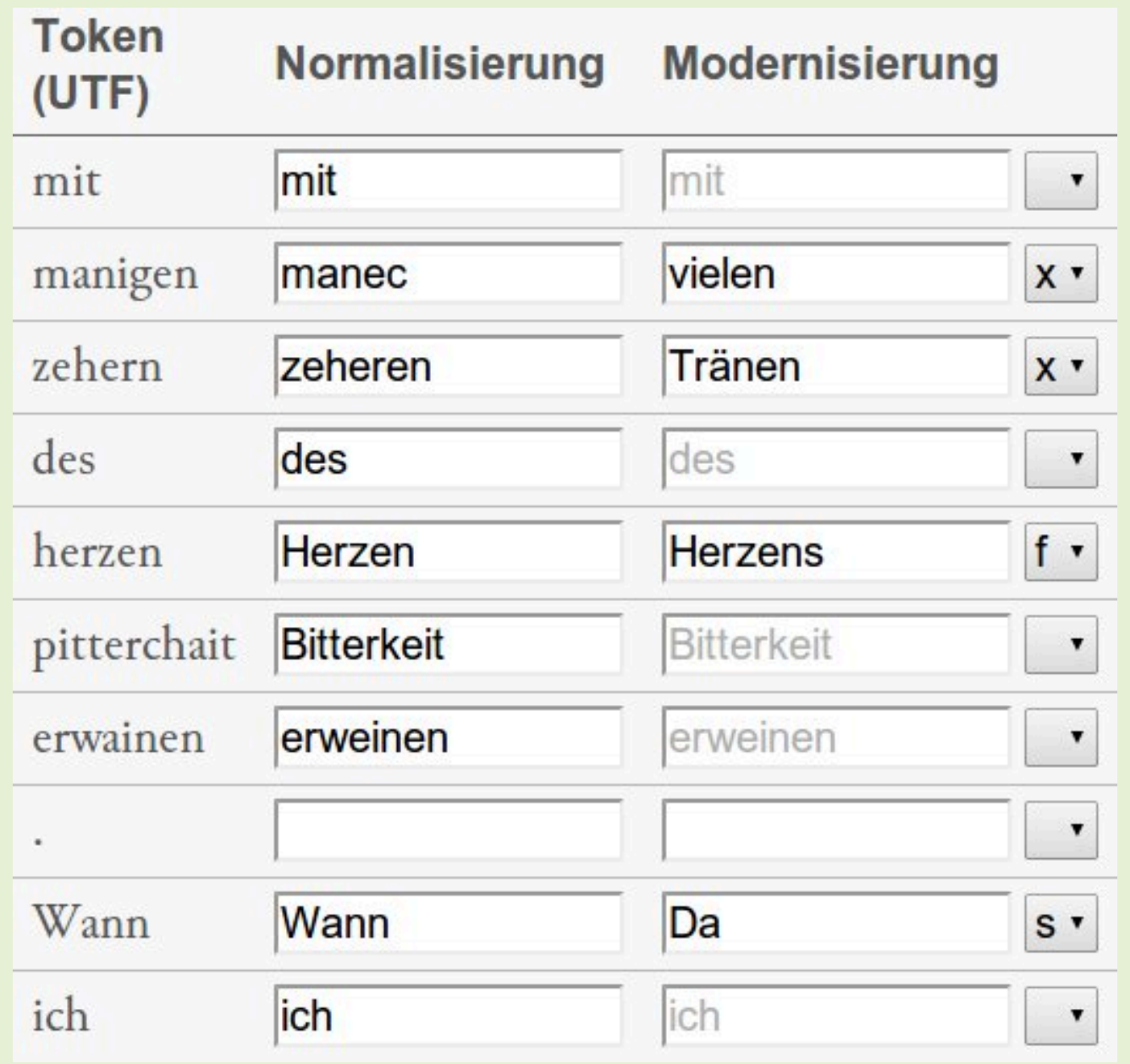
Seite: |<< < 1 > >>| Springe zu Zeile...

P	#	Zeile	E	Token (Trans)	Token (UTF)	Normalisierung	POS-Tag	Morphologie-Tag	Kommentar
🟩	1	1r,01	☐	von	von	von	APPR	--	
🟩	2	1r,01	☐	$ancte	ſancte	sankt	NE	Masc.Gen.Sg	
🟩	3	1r,01	☐	an$helm%9	anſhelmꝰ	anselm	NE	Masc.Gen.Sg	
🟩	4	1r,01	☐	frage	frage	frage	NN	Fem.Dat.Sg	
🟩	5	1r,01	☐	Ein	Ein	ein	ART	Fem.Nom.Sg	
🟩	6	1r,02	🟥	nuecze	nuecze	nütze	ADJA	Pos.Fem.Nom.Sg	unsure, check with dictionary
🟩	7	1r,02	☐	vnde	vnde	und	KON	--	
	8	1r,02	☐	gute	gute	gute	ADJA	Pos.Fem.Nom.Sg	
	9	1r,02	☐	betrachtunge	betrachtunge	betrachtung	NN		
	10	1r,02	☐	vn$er$z	vnſerſz	unseres	PPOSAT		

von ſancte anſhelmꝰ frage Ein nuecze vnde gute betrachtunge vnſerſz hern lieden vñ c er mutt goteſz alſz ir lieber
ſon gemartert warth sancte anſhelm hatte vnſir lieben frowen lange czieth mit faſten v tigem gebeth daſz ſie om
kunttede Ireſz eingeborn kindeſz tot vñ bitter martir vnde wie eſz von anfange ergieng

Seite: |<< < > >>|

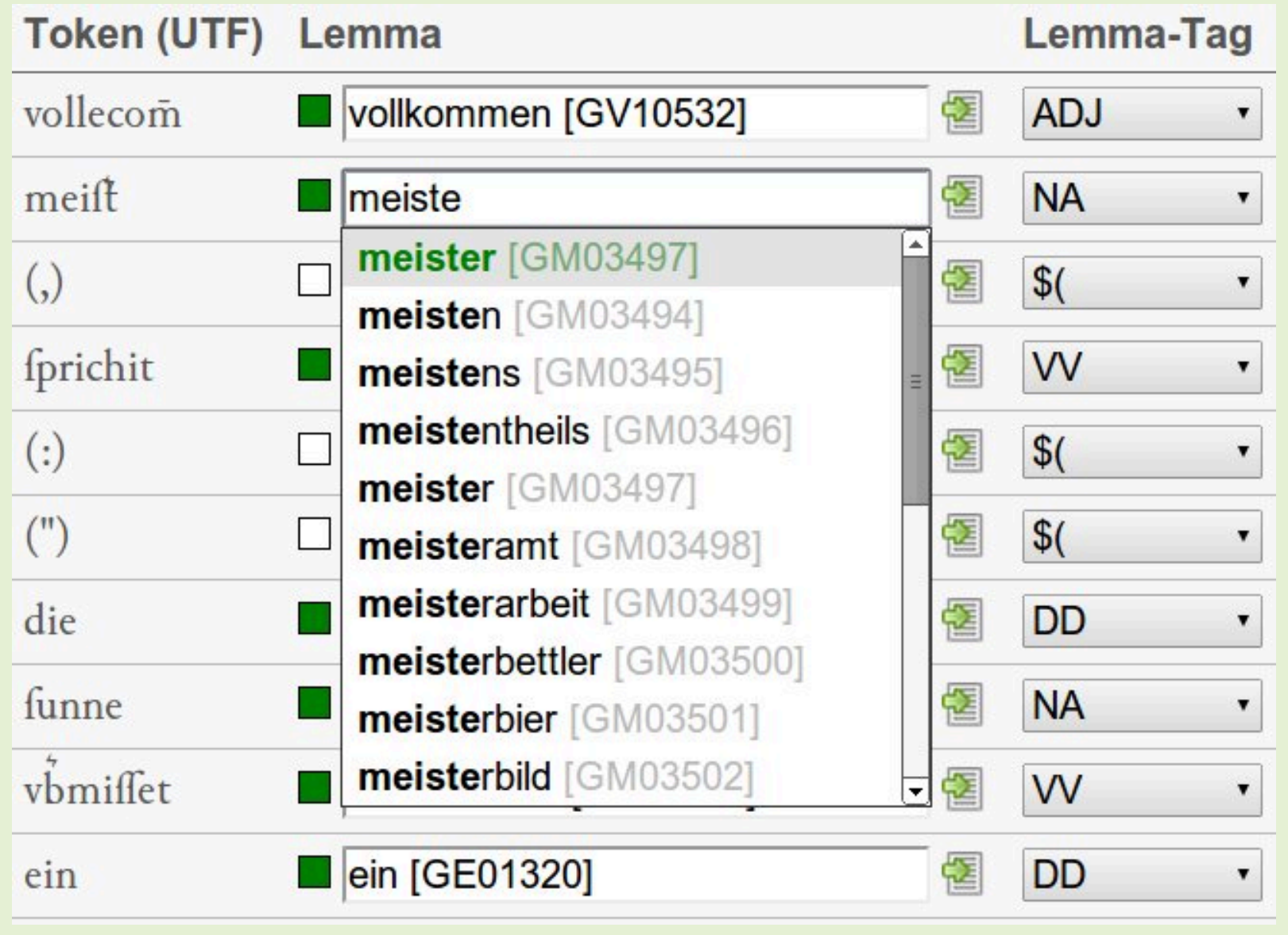
Main editor view of CorA, showing a selection of available columns



Token (UTF)	Normalisierung	Modernisierung	
mit	mit	mit	
manigen	manec	vielen	x
zehern	zeheren	Tränen	x
des	des	des	
herzen	Herzen	Herzens	f
pitterchait	Bitterkeit	Bitterkeit	
erwainen	erweinen	erweinen	
.			
Wann	Wann	Da	s
ich	ich	ich	

◂ Editor view with normalization, modernization, and a category field describing the type of modernization (x = extinct word, f = inflection, s = semantics)

Editor view showing lemma and lemma POS columns. ▸
A list of lemma forms is used to retrieve suggestions, with previously chosen lemmas for the same word in green. IDs shown in grey link to entries in the Deutsches Wörterbuch.
(http://www.woerterbuchnetz.de/DWB/)



Token (UTF)		Lemma	Lemma-Tag
vollecom̄	🟩	vollkommen [GV10532]	ADJ
meiſt	🟩	meiste	NA
(,)	☐		$(
ſprichit	🟩		VV
(:)	☐		$(
(")	☐		$(
die	🟩		DD
ſunne	🟩		NA
vbmiſſet	🟩		VV
ein	🟩	ein [GE01320]	DD

Projects using CorA

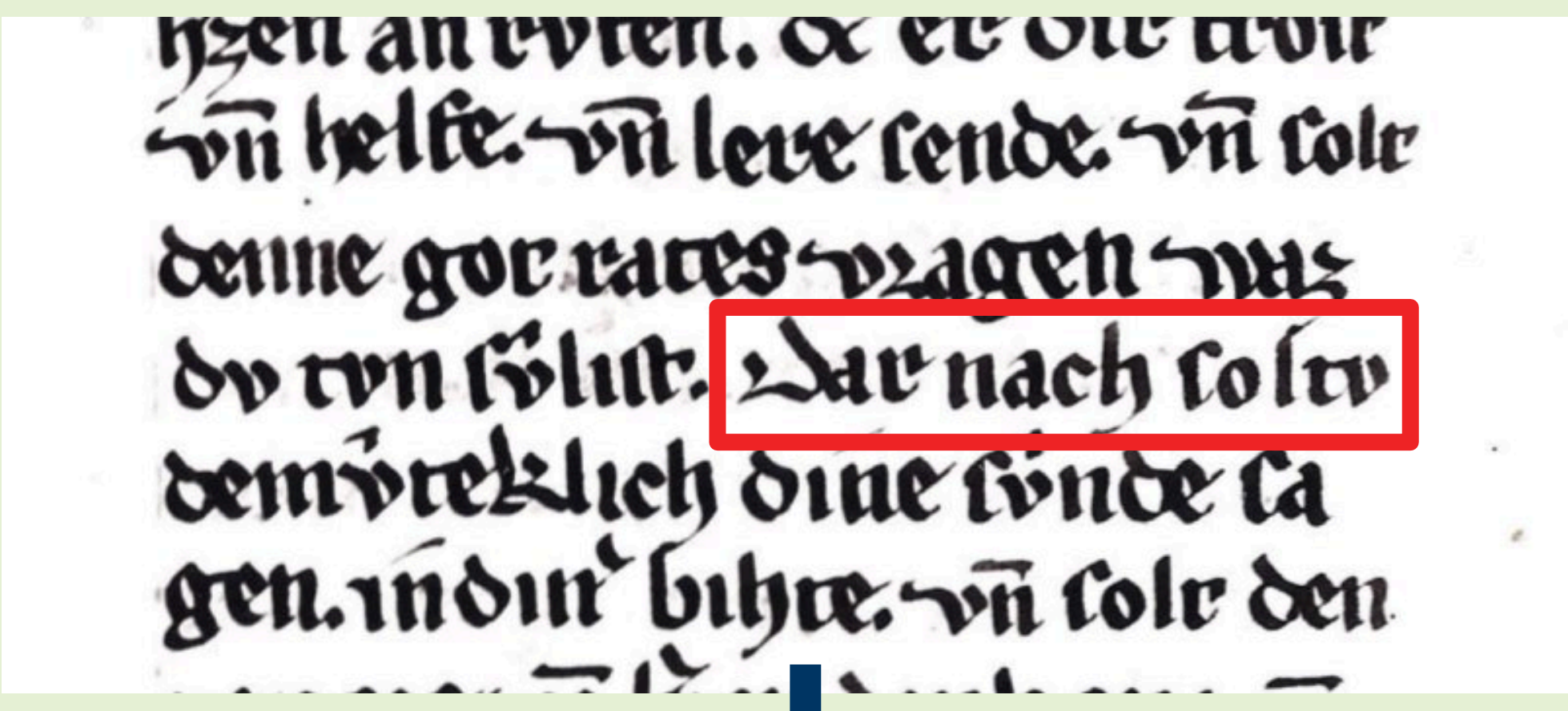
- A reference corpus of Early New High German (Projekt Referenzkorpus Frühneuhochdeutsch, n. d.)
  - Annotated with POS, morphology, and lemma
  - Integrated lemma suggestions and linking to an external dictionary
- A parallel corpus of Early New High German (Anselm project; see below)
  - Annotated with POS, normalization, and modernization (including category)
- German chat corpus
  - Annotated with POS
- German learner corpus
  - Focuses on categorization of spelling variation

Future Work

- Internationalization of the user interface (soon)
- Export to TEI format
- Closer integration with the ANNIS corpus tool for late error correction
- Performing inter-annotator agreement calculations

Use Case: The Anselm Project (Dipper & Schultz-Balluff, 2013)

1.



```
vn\- helfe. vn\- lere $ende. vn\- $olt
denne got rates vragen waz
dv tvn $v\ili$t. Dar#nach $olt||v
demv\iteklich dine $v\inde $a(=)
gen. in din' bihte. vn\- $olt den(=)
```

From the historical manuscript, a diplomatic transcription is created manually by specialists for German medieval studies.

2. The transcribed document is converted to CorA XML format, preserving different layers of tokenization.

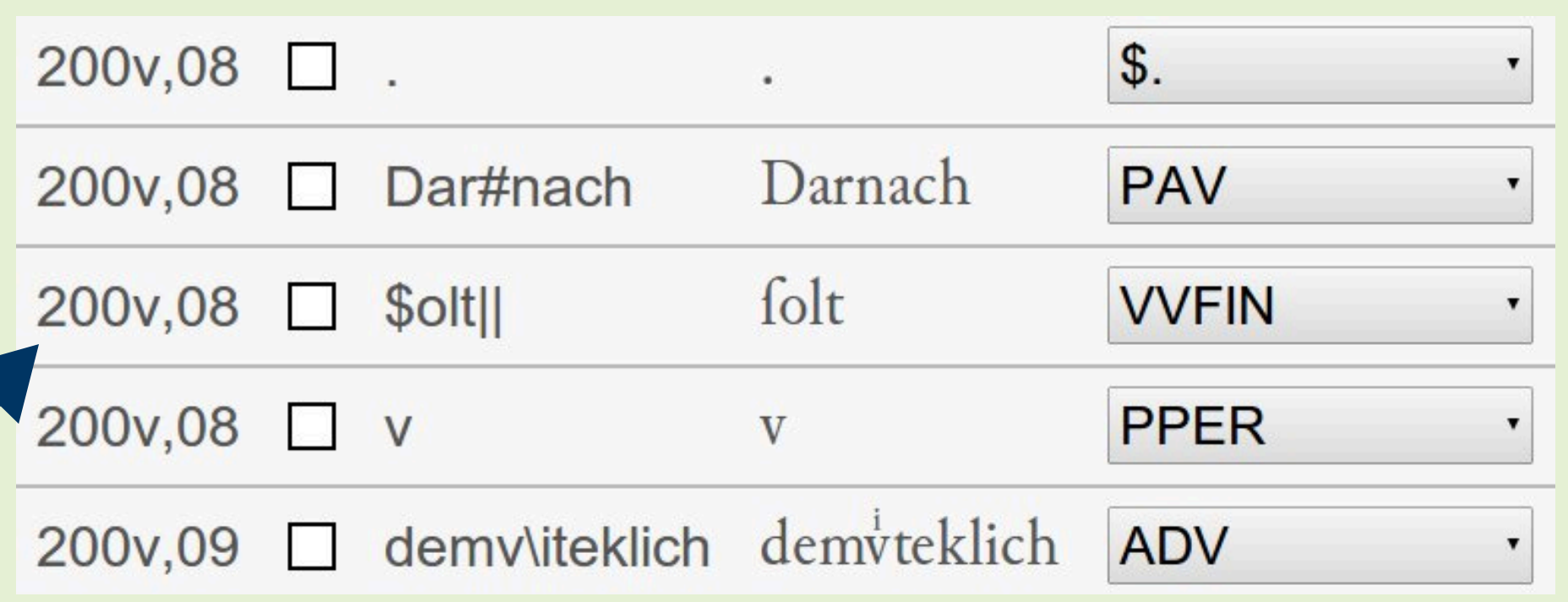
```
<token trans="Dar#nach">
   <dipl trans="Dar#"     utf="Dar"/>
   <dipl trans="nach"     utf="nach"/>
   <mod  trans="Dar#nach" utf="Darnach"/>
</token>

<token trans="$olt||v">
   <dipl trans="$olt||v" utf="ſoltv"/>
   <mod  trans="$olt"    utf="ſolt"/>
   <mod  trans="v"       utf="v"/>
</token>
```

- *dipl*: tokenization in the historical manuscript
- *mod*: tokenization following modern conventions
- *trans*: original transcription format
- *utf*: UTF-8 representation

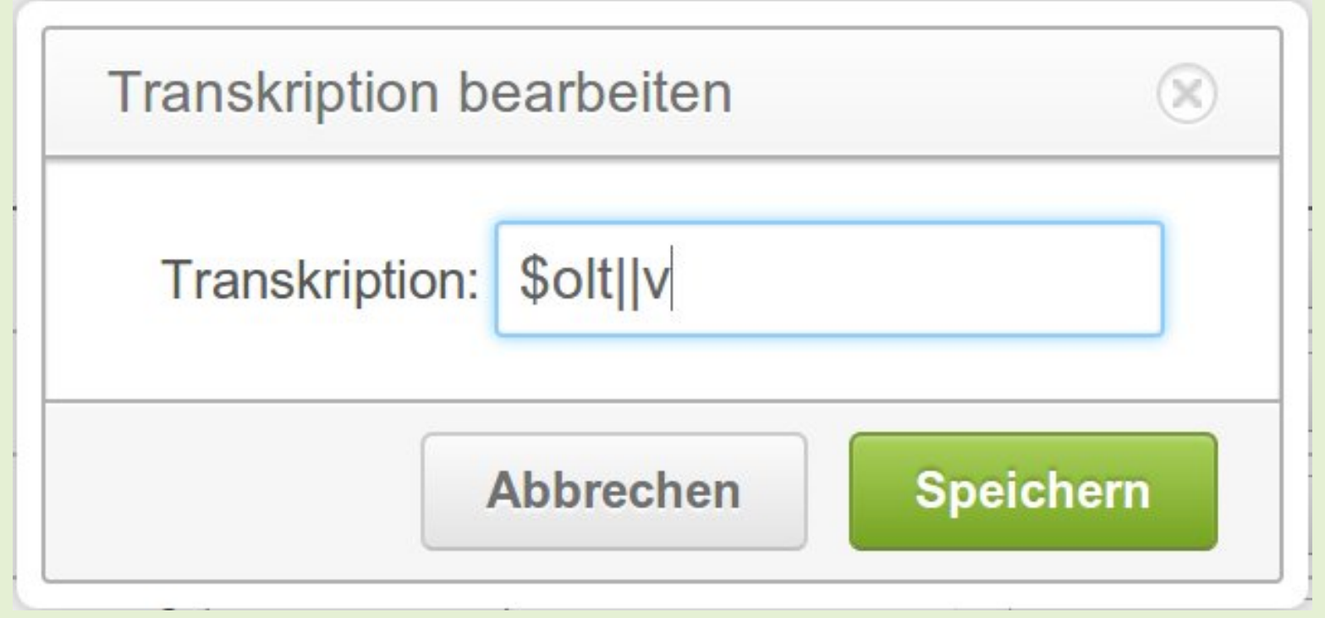
3. The XML file is imported into CorA.

Only the modernized tokenization is shown for annotation purposes:



200v,08	☐	.	.	$.
200v,08	☐	Dar#nach	Darnach	PAV
200v,08	☐	$olt||	ſolt	VVFIN
200v,08	☐	v	v	PPER
200v,09	☐	demv\iteklich	demvͥteklich	ADV

Errors in the transcription can be corrected directly within CorA.



Transkription: $olt||v

- Stefanie Dipper and Simone Schultz-Balluff (2013). *The Anselm Corpus: Methods and perspectives of a parallel aligned corpus.* In: Proceedings of the Workshop on Computational Historical Linguistics at NoDaLiDa 2013, NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings, pp. 27-42. Oslo, Norway. See also: http://www.linguistics.rub.de/anselm/
- Projekt Referenzkorpus Frühneuhochdeutsch (n. d.). http://www.ruhr-uni-bochum.de/wegera/ref/
- Amir Zeldes, Julia Ritz, Anke Lüdeling und Christian Chiarcos (2009). *ANNIS: a search tool for multi-layer annotated corpora.* In: Proceedings of Corpus Linguistics. Liverpool, UK.

The research reported here was financed by Deutsche Forschungsgemeinschaft (DFG), Grant No. DI 1558/5-1.