Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison

Stefanie Dipper

Linguistics Department
Ruhr-University Bochum

5 January, 2012
Workshop on Annotation of Corpora
for Research in the Humanities
Heidelberg

Goals

- A morphological and a part-of-speech (POS) tagger for texts from Middle High German (MHG, 1050–1350)
  - Ex: *sach* 'saw': `3.Sg.Past.Ind` (Morph)
    `VVFIN` (POS)

Goals

- A morphological and a part-of-speech (POS) tagger for texts from Middle High German (MHG, 1050–1350)
  - Ex: *sach* 'saw': `3.Sg.Past.Ind` (Morph)
    `VVFIN` (POS)
- Use of available state-of-the-art taggers
  - without any adaption
  - some preprocessing of the annotated training data

Goals

- A morphological and a part-of-speech (POS) tagger for texts from Middle High German (MHG, 1050–1350)
  - Ex: *sach* 'saw': `3.Sg.Past.Ind` (Morph)
    `VVFIN` (POS)
- Use of available state-of-the-art taggers
  - without any adaption
  - some preprocessing of the annotated training data
- Challenge: highly variant data
  - no spelling conventions, e.g. *sitzen, sizzen* 'sit'
  - different dialects, e.g. *bruoder, pruder* 'brother'

  → Data sparseness

Goals

- A morphological and a part-of-speech (POS) tagger for texts from Middle High German (MHG, 1050–1350)
  - Ex: *sach* 'saw': `3.Sg.Past.Ind` (Morph)
    `VVFIN` (POS)
- Use of available state-of-the-art taggers
  - without any adaption
  - some preprocessing of the annotated training data
- Challenge: highly variant data
  - no spelling conventions, e.g. *sitzen, sizzen* 'sit'
  - different dialects, e.g. *bruoder, pruder* 'brother'

  → Data sparseness
- Questions:
  - (how much) does normalization help?
    – original ("diplomatic") vs. normalized wordforms
  - does POS preprocessing help morphological tagging?

Outline

Outline

Project context

- Projects “Mittelhochdeutsche Grammatik” and “Referenzkorpus Mittelhochdeutsch” (Universities of Bochum and Bonn)

Project context

- Projects “Mittelhochdeutsche Grammatik” and “Referenzkorpus Mittelhochdeutsch” (Universities of Bochum and Bonn)
- Goals
  - a balanced, annotated reference corpus of MHG
  - diplomatic transcriptions
  - final size: 300 texts, 2 million wordforms
  - available via the internet (ANNIS)

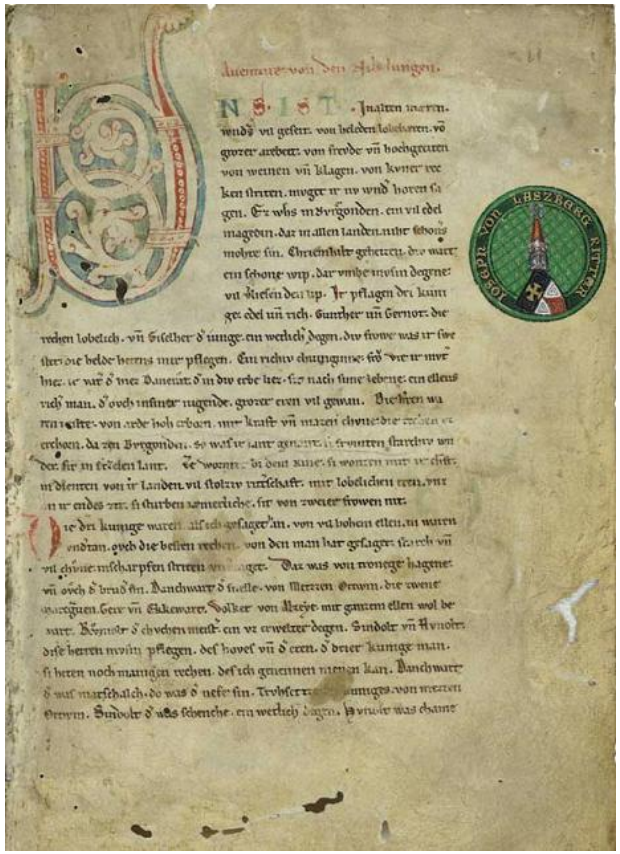
Project context

- Projects “Mittelhochdeutsche Grammatik” and “Referenzkorpus Mittelhochdeutsch” (Universities of Bochum and Bonn)
- Goals
  - a balanced, annotated reference corpus of MHG
  - diplomatic transcriptions
  - final size: 300 texts, 2 million wordforms
  - available via the internet (ANNIS)
- Annotations
  - parts of speech (POS)
  - morphological tags
  - lemma
  - normalized wordform

Project context

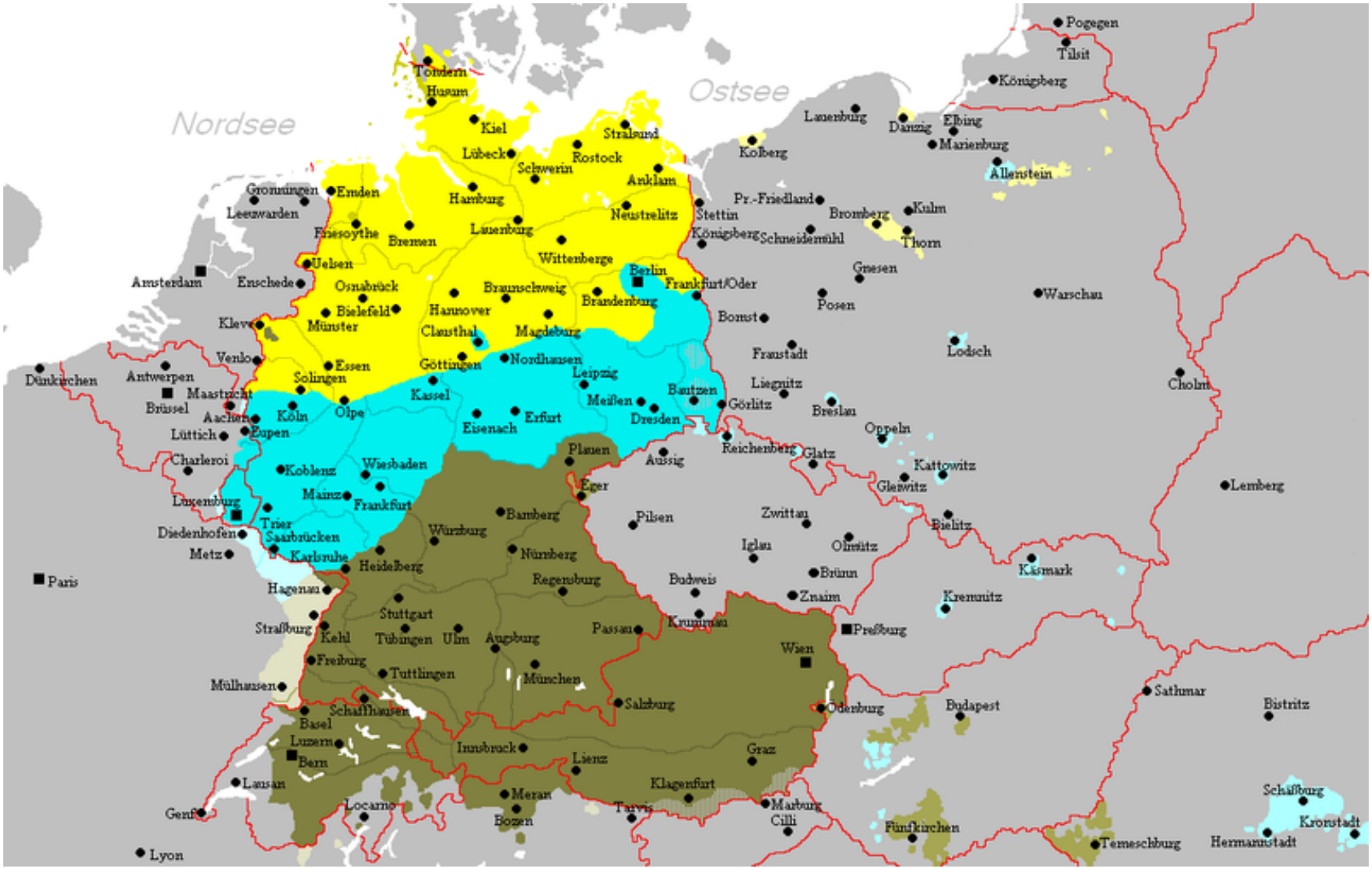
- Projects “Mittelhochdeutsche Grammatik” and “Referenzkorpus Mittelhochdeutsch” (Universities of Bochum and Bonn)
- Goals
  - a balanced, annotated reference corpus of MHG
  - diplomatic transcriptions
  - final size: 300 texts, 2 million wordforms
  - available via the internet (ANNIS)
- Annotations
  - parts of speech (POS)
  - morphological tags
  - lemma
  - normalized wordform
- Currently: semi-automatic annotation
  - tools by Thomas Klein, Bonn (2001)
  - require a lot of human intervention

The data



- 51 texts with 211,000 tokens from the MHG Reference Corpus
- From two dialect regions: Upper (UG) and Central German (CG)

Upper and Central (and Lower) German



Source: Wikimedia

Spelling variation

E.g. (normalized) *wolte* ‘wanted’

- *wolt*
- *wolta*
- *wolte*
- *woltt*
- *wolti*
- *wolthe*
- *walde*
- *uuolde*
- *volde*
- *. . .*

Spelling variation

E.g. (normalized) *wolte* ‘wanted’

- *wolt*
- *wolta*
- *wolte*
- *woltt*
- *wolti*
- *wolthe*
- *walde*
- *uuolde*
- *volde*
- *. . .*

Normalization: mapping to a virtual, idealized historical wordform

The data: some statistics

Texts	Tokens	Types *diplomatic*	Types *normalized*	[NHG]
51 total	211,000	40,500	20,500	
		.19	.10	[.14]
27 CG	91,000	22,000	13,000	
		.24	.14	[.18]
20 UG	67,000	15,000	8,500	
		.22	.13	
4 mixed	53,000			

The data: some statistics

Texts	Tokens	Types *diplomatic*	Types *normalized*	[NHG]
51 total	211,000	40,500 .19	20,500 .10	[.14]
27 CG	91,000	22,000 .24	13,000 .14	[.18]
20 UG	67,000	15,000 .22	8,500 .13	
4 mixed	53,000			

- On average:
  roughly 2 spelling variants (diplomatic) per wordform (normalized)

The data: some statistics

Texts	Tokens	Types *diplomatic*	Types *normalized*	[NHG]
51 total	211,000	40,500 .19	20,500 .10	[.14]
27 CG	91,000	22,000 .24	13,000 .14	[.18]
20 UG	67,000	15,000 .22	8,500 .13	
4 mixed	53,000			

- On average:
  roughly 2 spelling variants (diplomatic) per wordform (normalized)
- Type-token ratio: higher ratio $\rightarrow$ more diverse data
  – CG more diverse than UG
  [– cf. modern German (NHG): TTR = .14/.18]

Predictions I

1. Normalized vs. diplomatic:
   tagging normalized data should be easier

Predictions I

1. Normalized vs. diplomatic:
   tagging normalized data should be easier

2. CG vs. UG vs. NHG: rather unclear
   a) pro CG: more training data available
   b) pro UG: less diverse (lower type/token ratio)
   c) pro MHG (normalized, equal size): less diverse than NHG

Tagsets

POS

- based on the STTS tagset (standard German tagset)
- `NN, NE`
  “normal noun, proper noun”
  `VVFIN, VVINF, VVIMP, VVPP`
  “finite full verb, infinitive, imperative, past participle”

Tagsets

POS

- based on the STTS tagset (standard German tagset)
- `NN, NE`
  “normal noun, proper noun”
  `VVFIN, VVINF, VVIMP, VVPP`
  “finite full verb, infinitive, imperative, past participle”

Morphology

- “large” STTS tagset
- `Comp.Fem.Acc.Sg`
  “(adjective:) comparative form, feminine, accusative, singular”
  `3.Sg.Past.*`
  “(verb:) 3rd singular past tense, unspecified for mood”

Underspecification

- POS and morph: more underspecified tags in MHG than in NHG (no native speakers)
- Gender of nouns: not yet as fixed as nowadays
  - Example: *slange* 'snake': masc/fem

daz	*si*	*slangen*	*bizzen*
–	*.Acc.Pl	MascFem.Nom.Pl	3.Pl.Past.*
that	them	snakes	bit

'that snakes bit them'

Tagsets: some statistics (normalized data)

POS

	# Tags	Ø Tags/wofo	Median (max)
CG *norm*	44	1.10 ± 0.37	1 (7)
UG *norm*	41	1.10 ± 0.35	1 (6)
NHG (210K)	53	1.05 ± 0.23	1 (6)
(90K)	51	1.04 ± 0.21	1 (6)

Tagsets: some statistics (normalized data)

POS

	# Tags	Ø Tags/wofo	Median (max)
CG *norm*	44	1.10 ± 0.37	1 (7)
UG *norm*	41	1.10 ± 0.35	1 (6)
NHG (210K)	53	1.05 ± 0.23	1 (6)
(90K)	51	1.04 ± 0.21	1 (6)

Morphology

	# Tags	Ø Tags/wofo	Median (max)
CG *norm*	245	1.40 ± 1.16	1 (23)
UG *norm*	219	1.46 ± 1.28	1 (33)
NHG (210K)	230	1.37 ± 0.97	1 (26)
(90K)	205	1.32 ± 0.86	1 (18)

Predictions II

1. Normalized vs. diplomatic:
   tagging normalized data should be easier

2. CG vs. UG vs. NHG: rather unclear
   a) pro CG: more training data available
   b) pro UG: less diverse (lower type/token ratio)
   c) pro MHG (normalized, equal size): less diverse than NHG

3. Morphology vs. POS:
   tagging POS should be easier (lower ambiguity rate)

Predictions II

1. Normalized vs. diplomatic:
   tagging normalized data should be easier

2. CG vs. UG vs. NHG: rather unclear
   a) pro CG: more training data available
   b) pro UG: less diverse (lower type/token ratio)
   c) pro MHG (normalized, equal size): less diverse than NHG

3. Morphology vs. POS:
   tagging POS should be easier (lower ambiguity rate)

4. CG vs. UG vs. NHG: again, rather unclear
   a) CG and UG rather similar
   b) POS: pro UG, morph: pro CG (lower maxima)
   c) pro NHG: lower ambiguity rates

Outline

Other approaches

- Usually: map the historical wordforms to modern wordforms and apply a modern (trained) tagger (e.g. Rayson et al. 2007, Pilz et al. 2006)

Other approaches

- Usually: map the historical wordforms to modern wordforms and apply a modern (trained) tagger (e.g. Rayson et al. 2007, Pilz et al. 2006)
- Here: train and apply a tagger to historical wordforms (diplomatic and normalized)

The tagger: TreeTagger (Schmid 1994, 1995)

The tagger: TreeTagger (Schmid 1994, 1995)

- Training on annotated data

The tagger: TreeTagger (Schmid 1994, 1995)

- Training on annotated data
- TreeTagger assigns tags on the basis of
  1. a fullform lexicon
     e.g. `saw:  VERB see NOUN saw`
  2. prefix and suffix information
     e.g. `-ous` → 96% `ADJ`, 4% `NOUN`
  3. context information from *n* preceding tags (bigrams, trigrams, . . .)
     e.g. `DET ADJ ?` → 70% `NOUN`, 10% `ADJ`, . . .

The tagger: TreeTagger (Schmid 1994, 1995)

- Training on annotated data
- TreeTagger assigns tags on the basis of
  1. a fullform lexicon
     e.g. `saw: VERB see NOUN saw`
  2. prefix and suffix information
     e.g. `-ous` → 96% `ADJ`, 4% `NOUN`
  3. context information from *n* preceding tags (bigrams, trigrams, . . .)
     e.g. `DET ADJ ?` → 70% `NOUN`, 10% `ADJ`, . . .
- Performance:
  - 97.53% accuracy for modern German (POS tagging)
  - best published tagging accuracy for German (as of 2009)

The tagger: TreeTagger (Schmid 1994, 1995)

- Training on annotated data
- TreeTagger assigns tags on the basis of
  1. a fullform lexicon
     e.g. `saw: VERB see NOUN saw`
  2. prefix and suffix information
     e.g. `-ous` → 96% `ADJ`, 4% `NOUN`
  3. context information from *n* preceding tags (bigrams, trigrams, . . .)
     e.g. `DET ADJ ?` → 70% `NOUN`, 10% `ADJ`, . . .
- Performance:
  - 97.53% accuracy for modern German (POS tagging)
  - best published tagging accuracy for German (as of 2009)
- Easy to get and to use

POS and morphology experiments: 3 parameters

POS and morphology experiments: 3 parameters

1. CG vs. UG data

POS and morphology experiments: 3 parameters

1. CG vs. UG data
2. Diplomatic vs. normalized wordforms

POS and morphology experiments: 3 parameters

1. CG vs. UG data
2. Diplomatic vs. normalized wordforms
3. Training on all data vs. dialect-specific training ("generic" vs. "specific" taggers)

Results (summary)

- Top results
  - POS: 92.91% (UG *norm*, specific)
  - Morph: 80.84% (CG *norm*, generic)
  - cf. NHG (210K): 95.67% (POS), 76.95% (morph)

Results (summary)

- Top results
  - POS: 92.91% (UG *norm*, specific)
  - Morph: 80.84% (CG *norm*, generic)
  - cf. NHG (210K): 95.67% (POS), 76.95% (morph)

→ Predictions 2/4: results mirror ambiguity rates

→ Prediction 3 is confirmed: gap of $>$ 10 percentage points between POS and morph

Results (summary)

- Top results
  - POS: 92.91% (UG *norm*, specific)
  - Morph: 80.84% (CG *norm*, generic)
  - cf. NHG (210K): 95.67% (POS), 76.95% (morph)

  → Predictions 2/4: results mirror ambiguity rates

  → Prediction 3 is confirmed: gap of $>$ 10 percentage points between POS and morph

- Normalization: considerable improvements
  - 4–7 percentage points ($p < .001$)

  → Prediction 1 is confirmed

Results (summary)

- Top results
  - POS: 92.91% (UG *norm*, specific)
  - Morph: 80.84% (CG *norm*, generic)
  - cf. NHG (210K): 95.67% (POS), 76.95% (morph)

  → Predictions 2/4: results mirror ambiguity rates

  → Prediction 3 is confirmed: gap of $>$ 10 percentage points between POS and morph

- Normalization: considerable improvements
  - 4–7 percentage points ($p < .001$)

  → Prediction 1 is confirmed

- Generic vs. specific taggers: more (but heterogeneous) training data helps (significant in most scenarios)

Morphology experiment: an additional parameter

Integrate POS information:

Morphology experiment: an additional parameter

Integrate POS information:

1. Original version: no use of POS; bigrams

```
werde     3.Sg.Pres.Subj
disemo    Neut.Dat.Sg                    'would this'
```

Morphology experiment: an additional parameter

Integrate POS information:

1. Original version: no use of POS; bigrams

```
werde     3.Sg.Pres.Subj
disemo    Neut.Dat.Sg                    'would this'
```

2. Successive pairs of <wofo, POS><wofo, morph>; trigrams

```
werde     VAFIN
werde     3.Sg.Pres.Subj
disemo    PD
disemo    Neut.Dat.Sg
```

Morphology experiment: an additional parameter

Integrate POS information:

1. Original version: no use of POS; bigrams

```
werde     3.Sg.Pres.Subj
disemo    Neut.Dat.Sg                    'would this'
```

2. Successive pairs of <wofo, POS><wofo, morph>; trigrams

```
werde     VAFIN
werde     3.Sg.Pres.Subj
disemo    PD
disemo    Neut.Dat.Sg
```

3. Merged pairs of <wofo.POS, morph>; bigrams

```
werde.VAFIN    3.Sg.Pres.Subj
disemo.PD      Neut.Dat.Sg
```

Morphology experiment: an additional parameter

Integrate POS information:

1. Original version: no use of POS; bigrams

```
werde     3.Sg.Pres.Subj
disemo    Neut.Dat.Sg                    'would this'
```

2. Successive pairs of <wofo, POS><wofo, morph>; trigrams

```
werde     VAFIN
werde     3.Sg.Pres.Subj
disemo    PD
disemo    Neut.Dat.Sg
```

3. Merged pairs of <wofo.POS, morph>; bigrams

```
werde.VAFIN    3.Sg.Pres.Subj
disemo.PD      Neut.Dat.Sg
```

4. (Upper bound: merged pairs with gold POS)

Sample tagging rules

Successive pairs:

```
tag[-1] = ART
tag[-2] = Masc.Nom.Sg
                        Masc.Nom.Sg      0.625717
                        Pos.Masc.Nom.Sg  0.207722
```

Sample tagging rules

Successive pairs:

```
tag[-1] = ART
tag[-2] = Masc.Nom.Sg
                        Masc.Nom.Sg      0.625717
                        Pos.Masc.Nom.Sg  0.207722
```

Merged pairs:

```
n.PIS
          *.Nom.Sg 0.863558
          *.Acc.Sg 0.106894
          *.*.*    0.029548
e.PIS
          *.Dat.Sg 0.831818
          *.Nom.Sg 0.090909
          *.Acc.Sg 0.077273
```

Exemplary results

Results with CG *norm*, generic:

(i)	No use	79.70
(ii)	Successive pairs	80.84
(iii)	Merged pairs	79.81
(iv)	Gold POS	82.19

- Some improvements with successive pairs
- No improvement with merged pairs

Summary and Outlook

Summary and Outlook

- POS tagging
  - satisfiable results ($>$ 91%)
  - clearly better results with modern data (modern data: no cross-validation)

Summary and Outlook

- POS tagging
  - satisfiable results ($>$ 91%)
  - clearly better results with modern data
    (modern data: no cross-validation)
- Morphological tagging
  - needs more sophisticated tagging methods ($>$ 79%)
  - e.g. RFTagger (Schmid & Laws 2008): analyzes complex morphological tags
  - better results with historical data
    (maybe due to greater extent of underspecified annotations)

Summary and Outlook

- POS tagging
  - satisfiable results ($>$ 91%)
  - clearly better results with modern data
    (modern data: no cross-validation)
- Morphological tagging
  - needs more sophisticated tagging methods ($>$ 79%)
  - e.g. RFTagger (Schmid & Laws 2008): analyzes complex morphological tags
  - better results with historical data
    (maybe due to greater extent of underspecified annotations)
- Normalization: increases accuracy by 4–7 percentage points (with POS and morphological tagging)

Summary and Outlook

- POS tagging
  - satisfiable results ($>$ 91%)
  - clearly better results with modern data
    (modern data: no cross-validation)
- Morphological tagging
  - needs more sophisticated tagging methods ($>$ 79%)
  - e.g. RFTagger (Schmid & Laws 2008): analyzes complex morphological tags
  - better results with historical data
    (maybe due to greater extent of underspecified annotations)
- Normalization: increases accuracy by 4–7 percentage points (with POS and morphological tagging)
- TODO: Error analysis

Thank you!

(Semi-)automatic annotation

How many correct tags are among the top *n* most probable tags?

Part of Speech		
Rank	# Word forms	
1	8160	92.0%
2	370	4.2%
3	27	0.3%
None	303	3.4%

Morphology (merged)		
Rank	# Word forms	
1	7467	79.6%
2	600	6.4%
3	99	1.1%
None	1122	12.0%

$\rightarrow$ Top-3 ranks: 96.5% (POS), 87.1% (morph)

POS experiments: results

Dialect	Tagger	Word Forms		
		diplomatic	*normalized*	[NHG]
CG	*generic*	86.92	91.66	[95.67]
	specific	86.62	91.43	[94.39]
UG	*generic*	88.88	92.83	
	specific	89.16	**92.91**	

Morphological experiments: results

Scenario	Dialect	Tagger	Word Forms *diplomatic*	Word Forms *normalized*	[NHG]
(i) No use	CG	*gen*	73.91	*79.70*	[76.95]
		spec	72.64	78.43	[75.71]
	UG	*gen*	73.85	78.28	
		spec	73.23	78.15	
(ii) Succ. pairs	CG	*gen*	74.23	**80.84**	
		spec	72.37	79.47	
	UG	*gen*	74.17	79.11	
		spec	73.27	78.63	
(iii) Merged pairs	CG	*gen*	74.39	*79.81*	
		spec	72.86	78.48	
	UG	*gen*	74.07	77.63	
		spec	73.14	77.02	
(iv) Gold POS	CG	*gen*	77.14	*82.19*	
		spec	75.54	80.80	
	UG	*gen*	76.79	80.83	
		spec	75.79	80.26	