# Coreference relations in EN, DE and RU, and mapping between them

Yulia Grishina

Applied Computational Linguistics
FSP Cognitive Sciences
University of Potsdam / Germany

# Coreference resolution

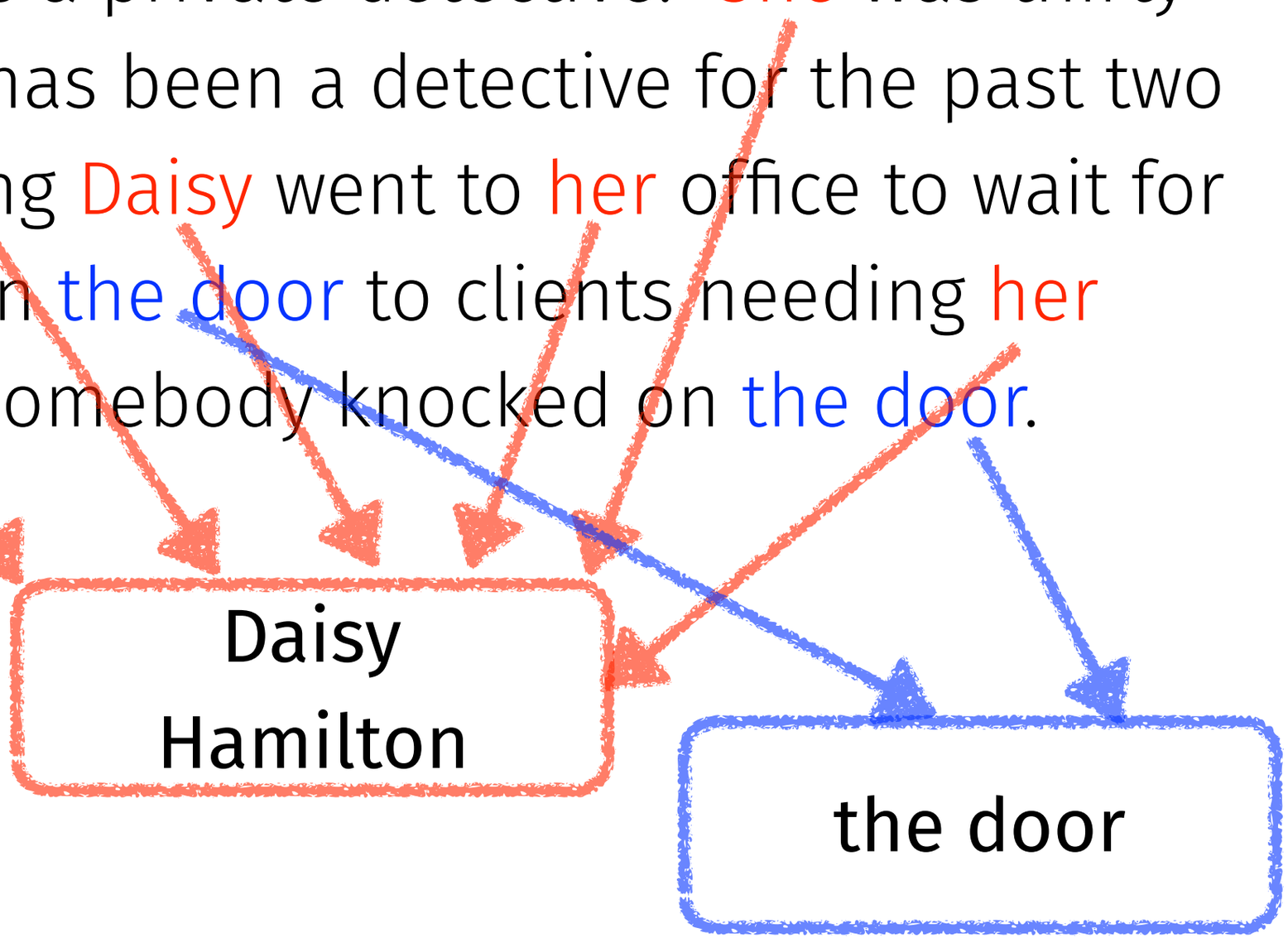
Daisy Hamilton was a private detective.  She was thirty years old and she has been a detective for the past two years. Every morning Daisy went to her office to wait for phone calls or open the door to clients needing her services. One day somebody knocked on the door.

# Coreference resolution

Daisy Hamilton was a private detective. She was thirty years old and she has been a detective for the past two years. Every morning Daisy went to her office to wait for phone calls or open the door to clients needing her services. One day somebody knocked on the door.

# Coreference resolution

Daisy Hamilton was a private detective. She was thirty years old and she has been a detective for the past two years. Every morning Daisy went to her office to wait for phone calls or open the door to clients needing her services. One day somebody knocked on the door.



| language | POS | syntax | coreference |
|---|---|---|---|
| Basque | √ | √ | - |
| Bulgarian | √ | √ | √ |
| Catalan | √ | √ | √ |
| Croatian | √ | √ | - |
| Czech | √ | √ | - |
| Danish | √ | √ | - |
| Dutch | √ | √ | √ |
| English | √ | √ | √ |
| Estonian | √ | √ | - |
| Finnish | √ | √ | - |
| French | √ | √ | √ |
| Galician | √ | √ | - |
| German | √ | √ | √ |
| Greek | √ | √ | - |
| Hungarian | √ | √ | √ |
| Icelandic | √ | √ | - |
| Irish | √ | √ | - |
| Italian | √ | √ | √ |
| Latvian | √ | √ | - |
| Lithuanian | √ | - | - |
| Maltese | √ | - | - |
| Norwegian | √ | √ | √ |
| Polish | √ | √ | √ |
| Portuguese | √ | √ | √ |

(Zhekova, 2013)

# Motivation

# Idea

High-quality coreference resolvers  exist for a small number of languages. How about other languages (in particular, low-resource)?

# Idea

High-quality coreference resolvers  exist for a small number of languages. How about other languages (in particular, low-resource)?

-> require multilingual resources

# Idea

High-quality coreference resolvers  exist for a small number of languages. How about other languages (in particular, low-resource)?

-> require multilingual resources

-> resource transfer (annotation projection)

# Outline

(I) direct coreference in English, German, Russian

(II) annotation projection

(III) indirect coreference (bridging & near-identity)

# I.

# direct coreference in EN, DE, RU

# The parallel corpus

- 38 parallel texts
- 3 languages: English, German, Russian
- 3 text genres: newswire[1], narratives[2], medicine instruction leaflets[3] (only EN-DE)

[1] multilingual newswire agency Project Syndicate (www.project-syndicate.org)

[2] short narratives for second language acquisition Daisy stories (http://www.lonweb.org)

[3] EMEA subcorpus of the OPUS collection of parallel corpora (Tiedemann, 2009)

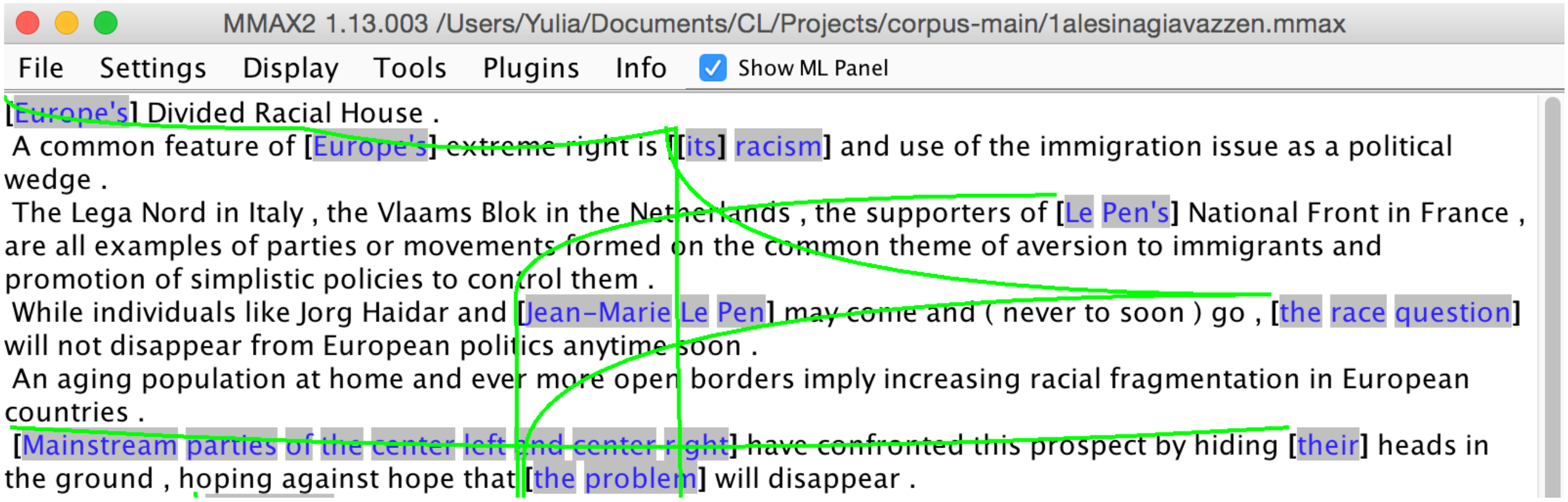
# Annotation

- common coreference annotation guidelines
- uniform annotations in 3 languages
- related annotation schemes: OntoNotes (Hovy et al., 2006) PoCoS (Krasavina & Chiarcos, 2007), ParCor (Guillou et al., 2014)
- identity relation
- annotation tool: MMAX-2 (Müller & Strube, 2006), subsequently converted into CoNLL-2012 format

# Annotation guidelines

- NP coreference: full NPs, proper names, pronouns
- no generic NPs annotated
- no singletons annotated
- set of attributes defined

# Annotation workflow



# Annotation workflow

# The parallel corpus

| | Newswire | | | Narratives | | | Medicine | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN | DE | RU | EN | DE | RU | EN | DE | EN | DE | RU |
| **Tokens** | 5903 | 6268 | 5763 | 2619 | 2642 | 2343 | 3386 | 3002 | *11908* | *11912* | *8106* |
| **Sentences** | 239 | 252 | 239 | 190 | 186 | 192 | 160 | 160 | *589* | *598* | *431* |
| **REs** | 558 | 589 | 606 | 470 | 497 | 479 | 322 | 309 | *1350* | *1395* | *1085* |
| **Chains** | 124 | 140 | 140 | 45 | 45 | 48 | 90 | 88 | *259* | *273* | *188* |
| **Tokens/REs (%)** | | | | | | | | | *8.82* | *8.5* | *7.47* |

(Grishina and Stede, 2015)

# The parallel corpus

| | Newswire | | | Narratives | | | Medicine | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN | DE | RU | EN | DE | RU | EN | DE | EN | DE | RU |
| **Tokens** | 5903 | 6268 | 5763 | 2619 | 2642 | 2343 | 3386 | 3002 | *11908* | *11912* | *8106* |
| **Sentences** | 239 | 252 | 239 | 190 | 186 | 192 | 160 | 160 | *589* | *598* | *431* |
| **REs** | 558 | 589 | 606 | 470 | 497 | 479 | 322 | 309 | *1350* | *1395* | *1085* |
| **Chains** | 124 | 140 | 140 | 45 | 45 | 48 | 90 | 88 | *259* | *273* | *188* |
| **Tokens/REs (%)** | | | | | | | | | *8.82* | *8.5* | *7.47* |

# The parallel corpus

| | Newswire | | | Narratives | | | Medicine | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN | DE | RU | EN | DE | RU | EN | DE | EN | DE | RU |
| **Tokens** | 5903 | 6268 | 5763 | 2619 | 2642 | 2343 | 3386 | 3002 | *11908* | *11912* | *8106* |
| **Sentences** | 239 | 252 | 239 | 190 | 186 | 192 | 160 | 160 | *589* | *598* | *431* |
| **REs** | 558 | 589 | 606 | 470 | 497 | 479 | 322 | 309 | *1350* | *1395* | *1085* |
| **Chains** | 124 | 140 | 140 | 45 | 45 | 48 | 90 | 88 | *259* | *273* | *188* |
| **Tokens/REs (%)** | | | | | | | | | *8.82* | *8.5* | *7.47* |

# The parallel corpus

| | Newswire | | | Narratives | | | Medicine | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN | DE | RU | EN | DE | RU | EN | DE | EN | DE | RU |
| **Tokens** | 5903 | 6268 | 5763 | 2619 | 2642 | 2343 | 3386 | 3002 | *11908* | *11912* | *8106* |
| **Sentences** | 239 | 252 | 239 | 190 | 186 | 192 | 160 | 160 | *589* | *598* | *431* |
| **REs** | 558 | 589 | 606 | 470 | 497 | 479 | 322 | 309 | *1350* | *1395* | *1085* |
| **Chains** | 124 | 140 | 140 | 45 | 45 | 48 | 90 | 88 | *259* | *273* | *188* |
| **Tokens/REs (%)** | | | | | | | | | *8.82* | *8.5* | *7.47* |

# Example: news

- NEW YORK – The terrorist sanctuary in [the South Waziristan region] of [Pakistan's] tribal frontier with [Afghanistan] is coming apart . It took a while for [the Pakistani Army] to move against the [region's] rising violence and chaos , but [[its] campaign in [South Waziristan]] is making progress .

**-> tokens/RE = 10.58**

**-> REs/chain = 4.5**

# Example: medicine

- [Abilify] is a medicine containing [the active substance aripiprazole] . [It] is available as 5 mg, 10 mg, 15 mg and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets. [Abilify] is used to treat adults with the following mental illnesses: [schizophrenia, a mental illness with a number of symptoms], …

**-> tokens/RE = 10.5**

**-> REs/chain = 3.5**

# Example: narrative

- [Daisy] had got up early that spring morning because [she] was working on a case in the nearby town. [She] arrived at [[her] office] with a paper bag in [her] hand containing fresh cream buns at a quarter to eight and was dying for a cup of coffee.

**-> tokens/RE = 5.57**

**-> REs/chain = 10.4**

->

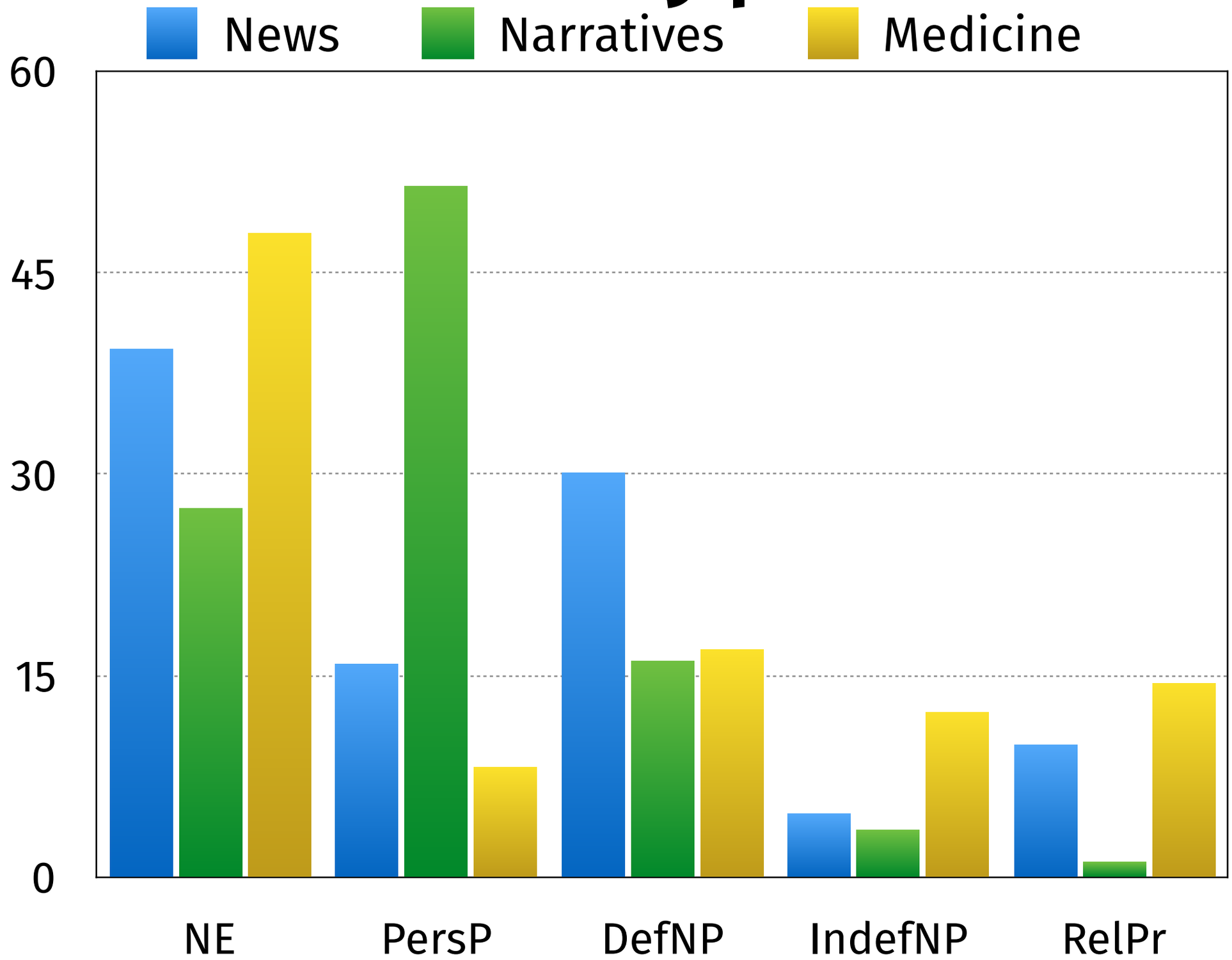
- **Average number of tokens per RE per lang:** EN (8.82), DE (8.5), RU (7.47)

- **Average number of tokens per RE**: newswire (10.58) > medicine (10.5) > narrative (5.57)

- **Average number of REs per chain**: narrative (10.4) > newswire (4.5) > medicine (3.5)

- tokens/RE, REs/chain are stable across languages

# Annotation: types of NPs



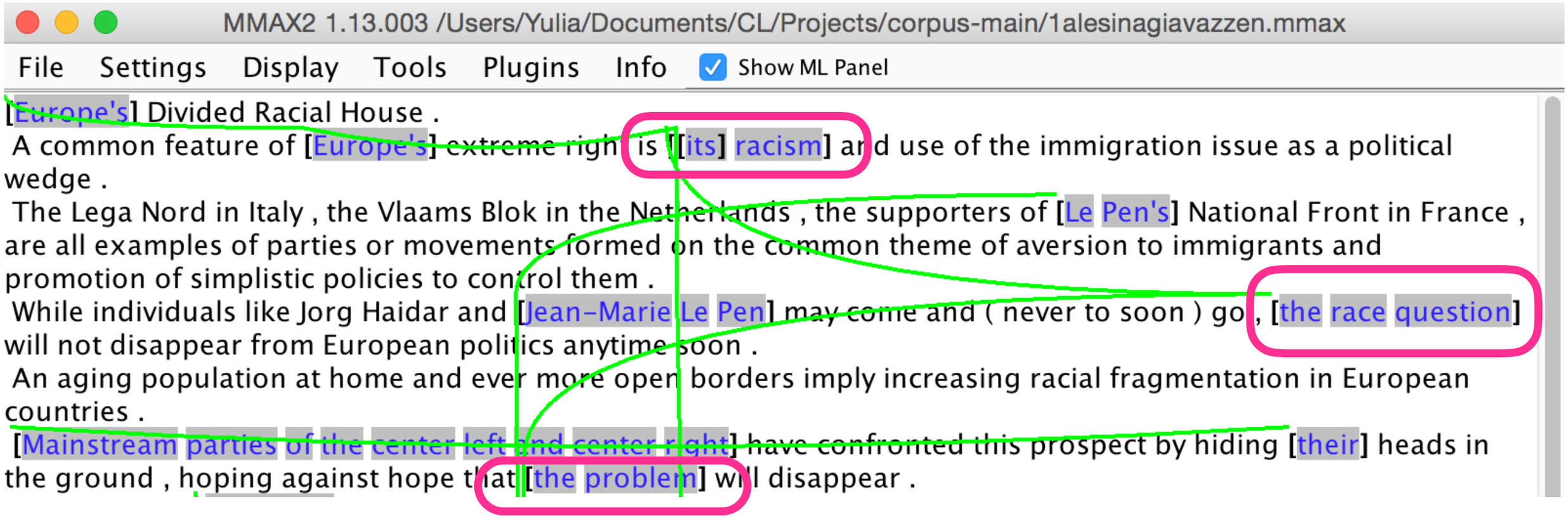
# Inter-annotator agreement

- EN-DE: each text annotated by 2 lightly trained annotators (students of linguistics)

| | English | German |
|---|---|---|
| Binary overlap $\kappa$ | 0.87 | 0.86 |
| Proportional overlap $\kappa$ | 0.81 | 0.81 |
| MUC F-score | 77.28 | 73.91 |

- EN-RU: only 1 annotator available

# Annotation discrepancies




**-> need to incorporate near-identity and bridging (part III of the talk)**

# -> Outcomes I

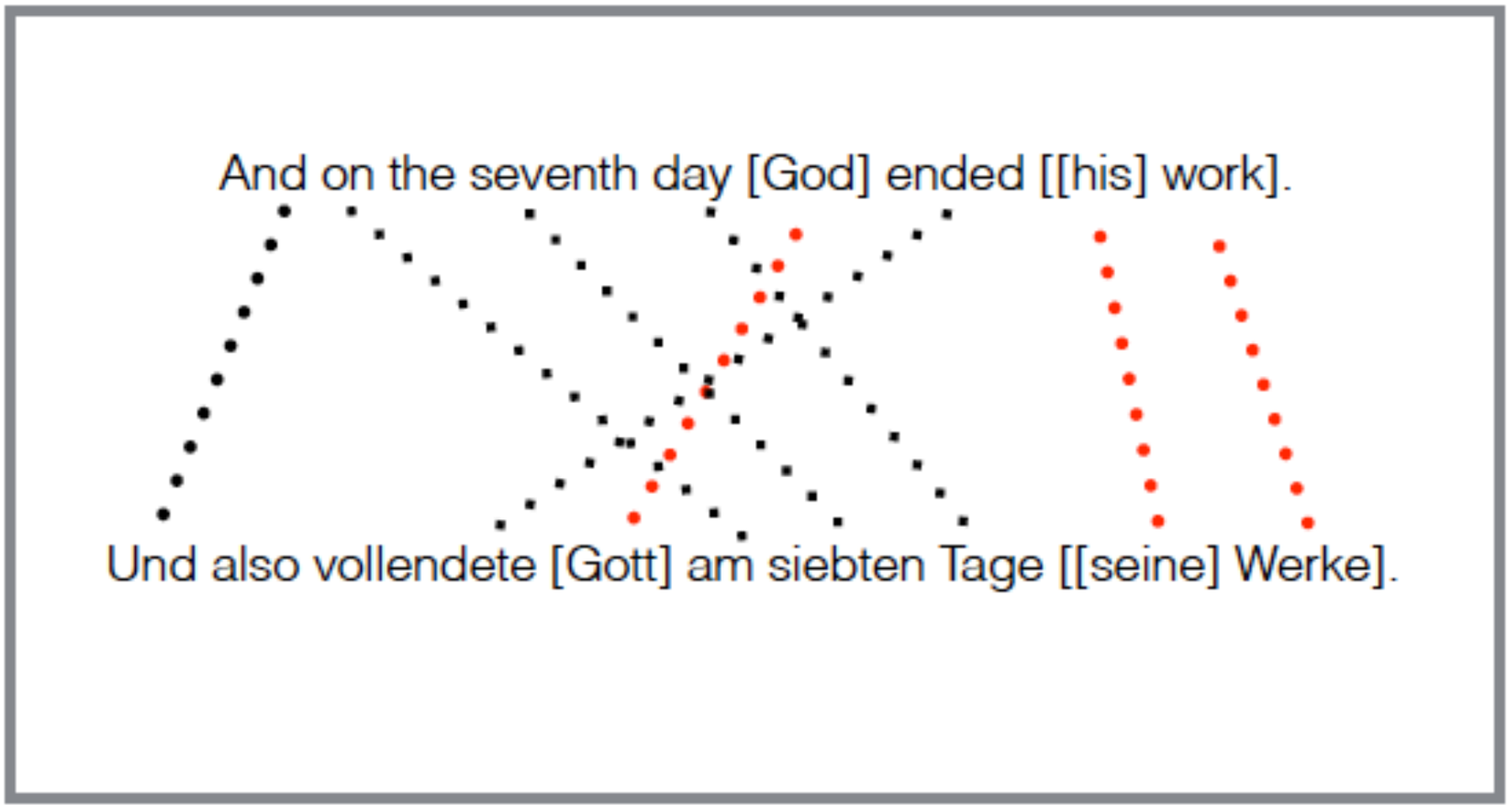
- created common annotation guidelines
- built a parallel coreference corpus of 3 genres
- compared annotations in 3 languages

# II.

# annotation projection

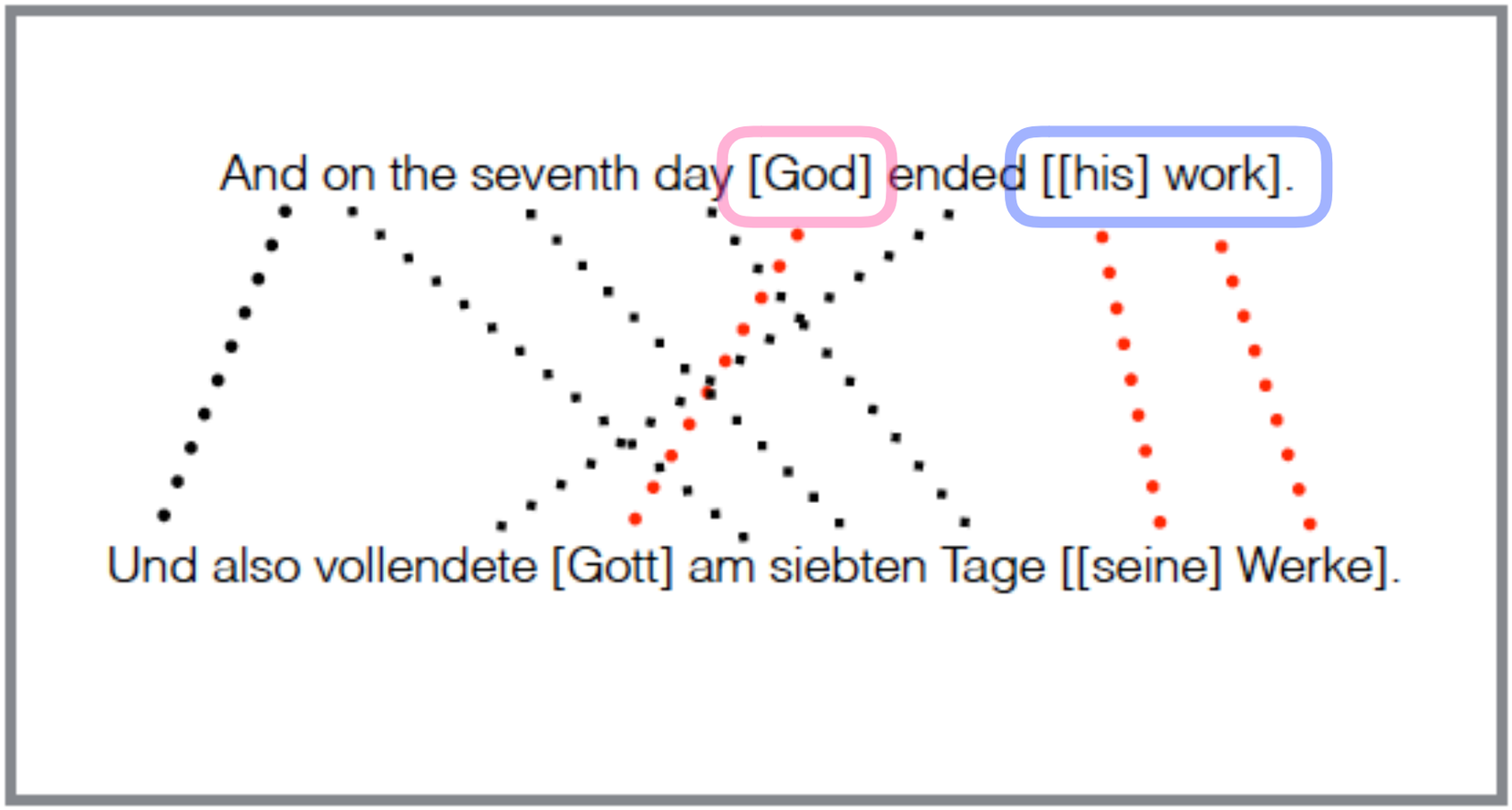
# Annotation projection

- **automatically transfer annotations from source to target**



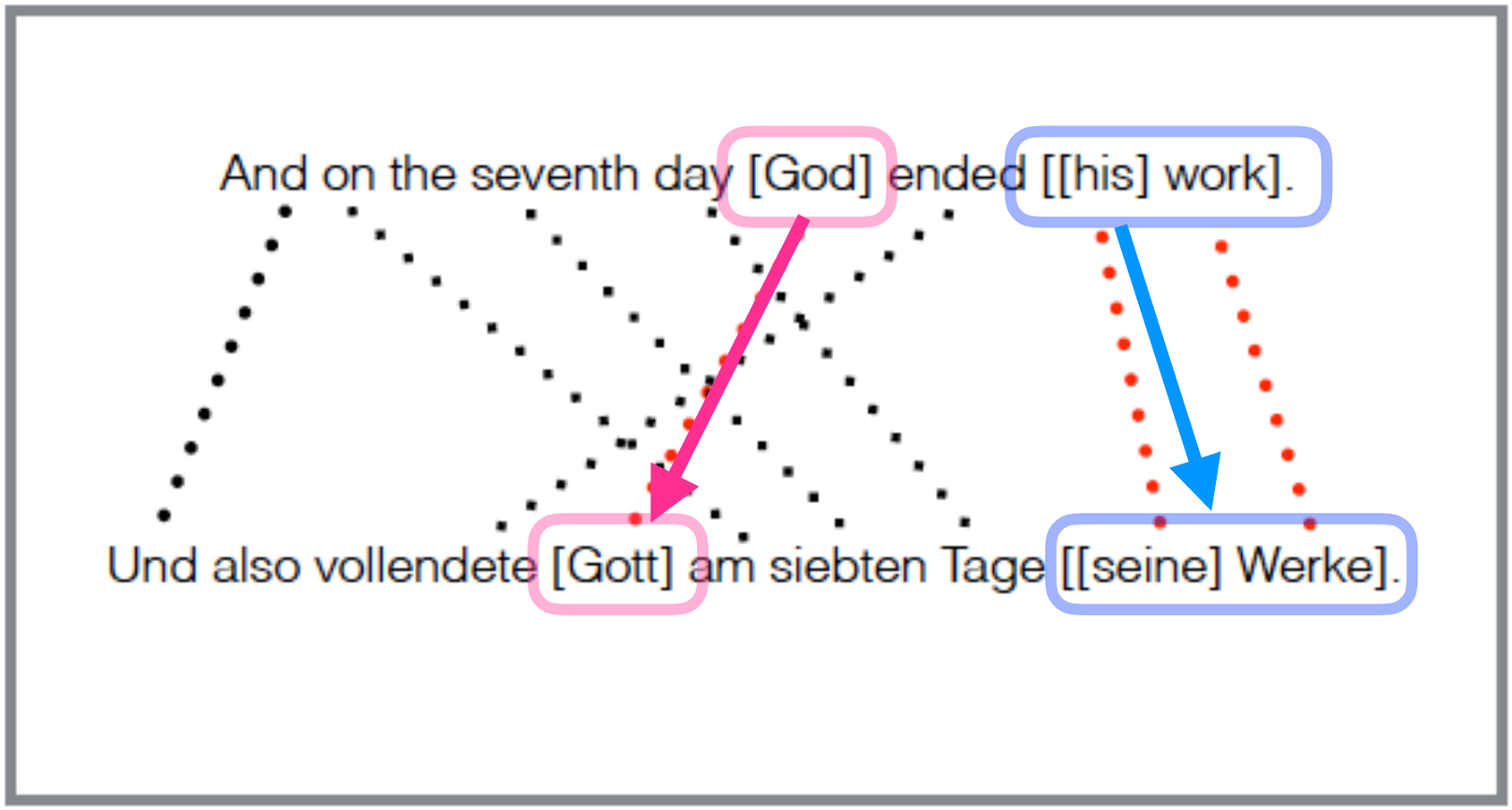
# Annotation projection

- **automatically transfer annotations from source to target**



# Annotation projection

- **automatically transfer annotations from source to target**



# Annotation projection

- POS tags (Yarowsky et al., 2001), (Agic et al., 2015)
- Named Entities (Ehrmann et al., 2011)
- syntactic trees (Tiedemann, 2014), (Johannsen, 2016)
- **coreference**

# Experimental setup

1. automatic sentence and word alignment

2. extraction of REs

3. transfer of coreference chains

# Alignment

- Sentence alignment: HunAlign (Varga et al., 2007) and its wrapper LF Aligner
- Word alignment: GIZA++ (Och & Ney, 2003)
  - training set: around 200 000 parallel sentences
  - intersective alignments

# Projection

- Direct projection algorithm (following (Postolache et al., 2006)):
  - for each word span $e_1 \ldots e_i$ we extract corresponding aligned words $f_1 \ldots f_j$
  - remove duplicates, reorder according to the surface order
  - target RE is the span between the 1st and the last word and it belongs to the same coreference set as the source RE

# Projection

It was a fat lady who wore a fur around her neck.

She said that she needs Daisy’s help and does not know what to do.

Es war eine dicke Dame mit einer Pelzstola. Sie hat gesagt, dass sie Daisys Hilfe braucht und dass sie nicht weiß, was sie tun soll.

# Projection

It was [a fat lady] [who] wore a fur around [her] neck. [She] said that [she] needs [Daisy's] help and does not know what to do.

Es war eine dicke Dame mit einer Pelzstola. Sie hat gesagt, dass sie Daisys Hilfe braucht und dass sie nicht weiß, was sie tun soll.

# Projection

It was **[a fat lady]** **[who]** wore a fur around **[her]** neck. **[She]** said that **[she]** needs **[Daisy's]** help and does not know what to do.

Es war **eine** dicke **Dame** mit einer Pelzstola. **Sie** hat gesagt, dass **sie Daisys** Hilfe braucht und dass sie nicht weiß, was sie tun soll.

# Projection

It was **[a fat lady]** **[who]** wore a fur around **[her]** neck. **[She]** said that **[she]** needs **[Daisy's]** help and does not know what to do.

Es war **[eine dicke Dame]** mit einer Pelzstola. **[Sie]** hat gesagt, dass **[sie]** **[Daisys]** Hilfe braucht und dass sie nicht weiß, was sie tun soll.

# Errors! Projection

It was **[a fat lady]** **[who]** wore a fur around **[her]** neck. **[She]** said that **[she]** needs **[Daisy's]** help and does not know what to do.

Es war **[eine dicke Dame]** mit einer Pelzstola **[Sie]** hat gesagt, dass **[sie]** **[Daisys]** Hilfe braucht und dass sie nicht weiß, was sie tun soll.

# Results

EN

DE

Coreference chains;
automatic;
F1 DE = 50.8
F1 RU = 67.2

RU

# Results

| | Mentions | | | MUC | | | CEAF | | | $B^3$ | | | Average (coref) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| *de*-News | 61.5 | 48.6 | 54.3 | 55.9 | 43.2 | 48.7 | 58.6 | 46.7 | 51.9 | 45.8 | 34.2 | 39.1 | 53.4 | 41.4 | 46.6 |
| *de*-Stories | 82.0 | 54.5 | 65.5 | 81.9 | 51.6 | 63.3 | 81.7 | 53. 7 | 64.8 | 71.6 | 32.5 | 44.7 | 78.4 | 45.9 | 57.6 |
| *de*-Medicine | 61.2 | 44.7 | 51.7 | 66.2 | 42.7 | 51.9 | 59.1 | 43.3 | 50.0 | 53.43 | 35.16 | 42.41 | 59.6 | 40.4 | 48.1 |
| *de*-News$_{min}$ | 89.9 | 71.2 | 79.4 | 87.3 | 66.2 | 75.3 | 85.5 | 67.5 | 75.5 | 80.4 | 58.1 | 67.5 | 84.4 | 63.9 | 72.8 |
| *de*-Stories$_{min}$ | 95.4 | 62.2 | 75.3 | 94.4 | 58.5 | 72.2 | 95.1 | 61.2 | 74.5 | 90.9 | 40.2 | 55.7 | 93.5 | 53.3 | 67.5 |
| *de*-Medicine$_{min}$ | 79.9 | 58.4 | 67.5 | 84.2 | 54.4 | 66.1 | 77.7 | 56.9 | 65.7 | 73.3 | 47.2 | 57.4 | 78.4 | 52.8 | 63.1 |
| *ru*-News | 79.3 | 64.5 | 71.2 | 76.3 | 60.7 | 67.6 | 76.3 | 62.0 | 68.4 | 69.0 | 52.2 | 59.4 | 73.9 | 58.3 | 65.1 |
| *ru*-Stories | 87.4 | 65.1 | 74.6 | 87.9 | 64.4 | 74.3 | 86.1 | 64.6 | 73.8 | 79.7 | 47.9 | 59.8 | 84.6 | 59.0 | 69.3 |
| *ru*-News$_{min}$ | 90.9 | 72.6 | 80.7 | 89.6 | 69.8 | 78.5 | 87.3 | 69.7 | 77.5 | 83.7 | 61.4 | 70.9 | 86.9 | 67.0 | 75.6 |
| *ru*-Stories$_{min}$ | 94.3 | 72.4 | 81.9 | 94.0 | 70.9 | 80.9 | 93.6 | 71.7 | 81.2 | 90.2 | 57.3 | 70.1 | 92.6 | 66.6 | 77.4 |

# Results

Results for Russian are better

| | Mentions | | | MUC | | | CEAF | | | $B^3$ | | | Average (coref) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| *de*-News | 61.5 | 48.6 | 54.3 | 55.9 | 43.2 | 48.7 | 58.6 | 46.7 | 51.9 | 45.8 | 34.2 | 39.1 | 53.4 | 41.4 | 46.6 |
| *de*-Stories | 82.0 | 54.5 | 65.5 | 81.9 | 51.6 | 63.3 | 81.7 | 53. 7 | 64.8 | 71.6 | 32.5 | 44.7 | 78.4 | 45.9 | 57.6 |
| *de*-Medicine | 61.2 | 44.7 | 51.7 | 66.2 | 42.7 | 51.9 | 59.1 | 43.3 | 50.0 | 53.43 | 35.16 | 42.41 | 59.6 | 40.4 | 48.1 |
| *de*-News$_{min}$ | 89.9 | 71.2 | 79.4 | 87.3 | 66.2 | 75.3 | 85.5 | 67.5 | 75.5 | 80.4 | 58.1 | 67.5 | 84.4 | 63.9 | 72.8 |
| *de*-Stories$_{min}$ | 95.4 | 62.2 | 75.3 | 94.4 | 58.5 | 72.2 | 95.1 | 61.2 | 74.5 | 90.9 | 40.2 | 55.7 | 93.5 | 53.3 | 67.5 |
| *de*-Medicine$_{min}$ | 79.9 | 58.4 | 67.5 | 84.2 | 54.4 | 66.1 | 77.7 | 56.9 | 65.7 | 73.3 | 47.2 | 57.4 | 78.4 | 52.8 | 63.1 |
| *ru*-News | 79.3 | 64.5 | 71.2 | 76.3 | 60.7 | 67.6 | 76.3 | 62.0 | 68.4 | 69.0 | 52.2 | 59.4 | 73.9 | 58.3 | 65.1 |
| *ru*-Stories | 87.4 | 65.1 | 74.6 | 87.9 | 64.4 | 74.3 | 86.1 | 64.6 | 73.8 | 79.7 | 47.9 | 59.8 | 84.6 | 59.0 | 69.3 |
| *ru*-News$_{min}$ | 90.9 | 72.6 | 80.7 | 89.6 | 69.8 | 78.5 | 87.3 | 69.7 | 77.5 | 83.7 | 61.4 | 70.9 | 86.9 | 67.0 | 75.6 |
| *ru*-Stories$_{min}$ | 94.3 | 72.4 | 81.9 | 94.0 | 70.9 | 80.9 | 93.6 | 71.7 | 81.2 | 90.2 | 57.3 | 70.1 | 92.6 | 66.6 | 77.4 |

# Results

Results for Russian are better

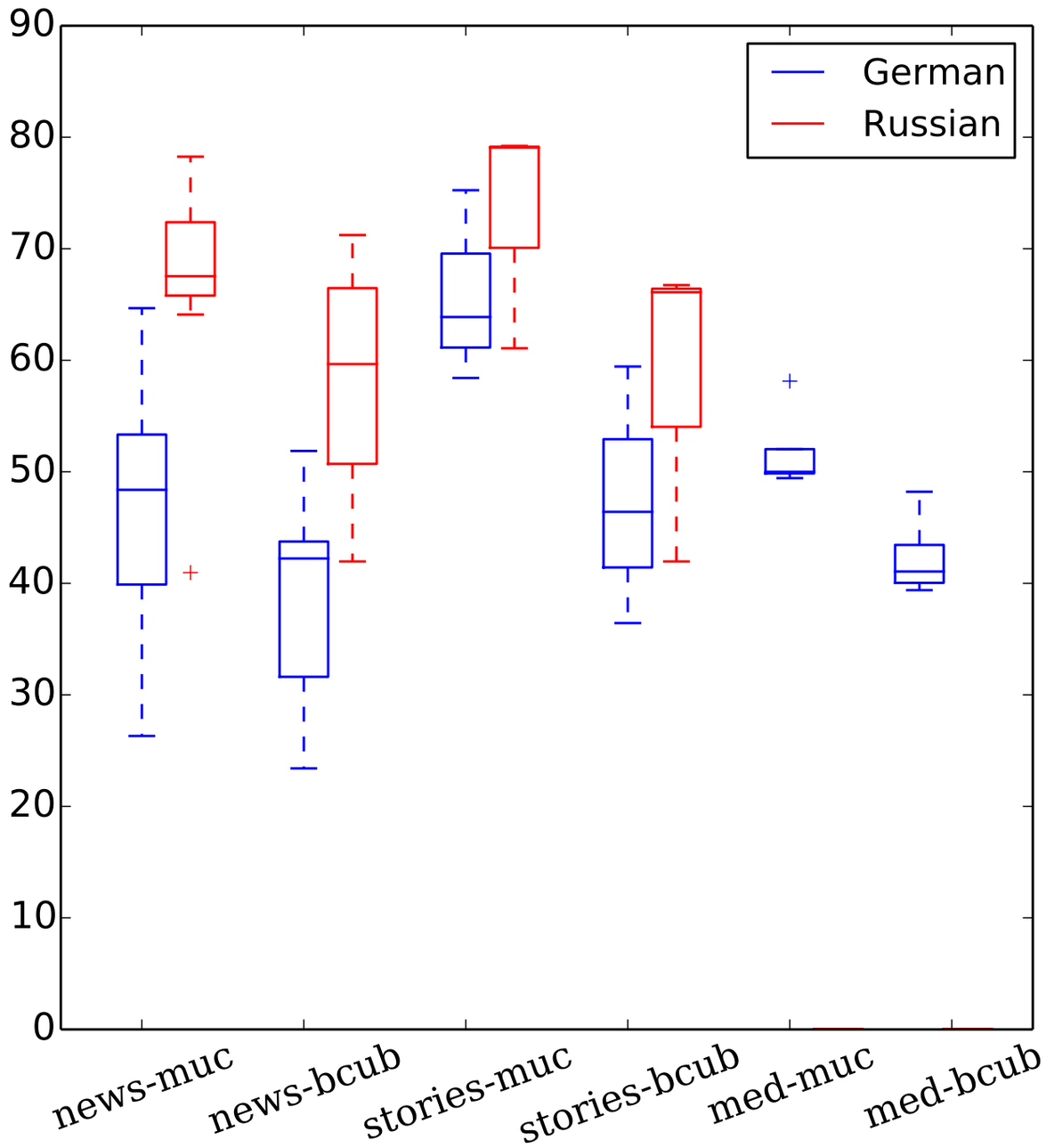
| | Mentions | | | MUC | | | CEAF | | | $B^3$ | | | Average (coref) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| *de*-News | 61.5 | 48.6 | 54.3 | 55.9 | 43.2 | 48.7 | 58.6 | 46.7 | 51.9 | 45.8 | 34.2 | 39.1 | 53.4 | 41.4 | 46.6 |
| *de*-Stories | 82.0 | 54.5 | 65.5 | 81.9 | 51.6 | 63.3 | 81.7 | 53. 7 | 64.8 | 71.6 | 32.5 | 44.7 | 78.4 | 45.9 | 57.6 |
| *de*-Medicine | 61.2 | 44.7 | 51.7 | 66.2 | 42.7 | 51.9 | 59.1 | 43.3 | 50.0 | 53.43 | 35.16 | 42.41 | 59.6 | 40.4 | 48.1 |
| *de*-News$_{min}$ | 89.9 | 71.2 | 79.4 | 87.3 | 66.2 | 75.3 | 85.5 | 67.5 | 75.5 | 80.4 | 58.1 | 67.5 | 84.4 | 63.9 | 72.8 |
| *de*-Stories$_{min}$ | 95.4 | 62.2 | 75.3 | 94.4 | 58.5 | 72.2 | 95.1 | 61.2 | 74.5 | 90.9 | 40.2 | 55.7 | 93.5 | 53.3 | 67.5 |
| *de*-Medicine$_{min}$ | 79.9 | 58.4 | 67.5 | 84.2 | 54.4 | 66.1 | 77.7 | 56.9 | 65.7 | 73.3 | 47.2 | 57.4 | 78.4 | 52.8 | 63.1 |
| *ru*-News | 79.3 | 64.5 | 71.2 | 76.3 | 60.7 | 67.6 | 76.3 | 62.0 | 68.4 | 69.0 | 52.2 | 59.4 | 73.9 | 58.3 | 65.1 |
| *ru*-Stories | 87.4 | 65.1 | 74.6 | 87.9 | 64.4 | 74.3 | 86.1 | 64.6 | 73.8 | 79.7 | 47.9 | 59.8 | 84.6 | 59.0 | 69.3 |
| *ru*-News$_{min}$ | 90.9 | 72.6 | 80.7 | 89.6 | 69.8 | 78.5 | 87.3 | 69.7 | 77.5 | 83.7 | 61.4 | 70.9 | 86.9 | 67.0 | 75.6 |
| *ru*-Stories$_{min}$ | 94.3 | 72.4 | 81.9 | 94.0 | 70.9 | 80.9 | 93.6 | 71.7 | 81.2 | 90.2 | 57.3 | 70.1 | 92.6 | 66.6 | 77.4 |

Results for stories are better

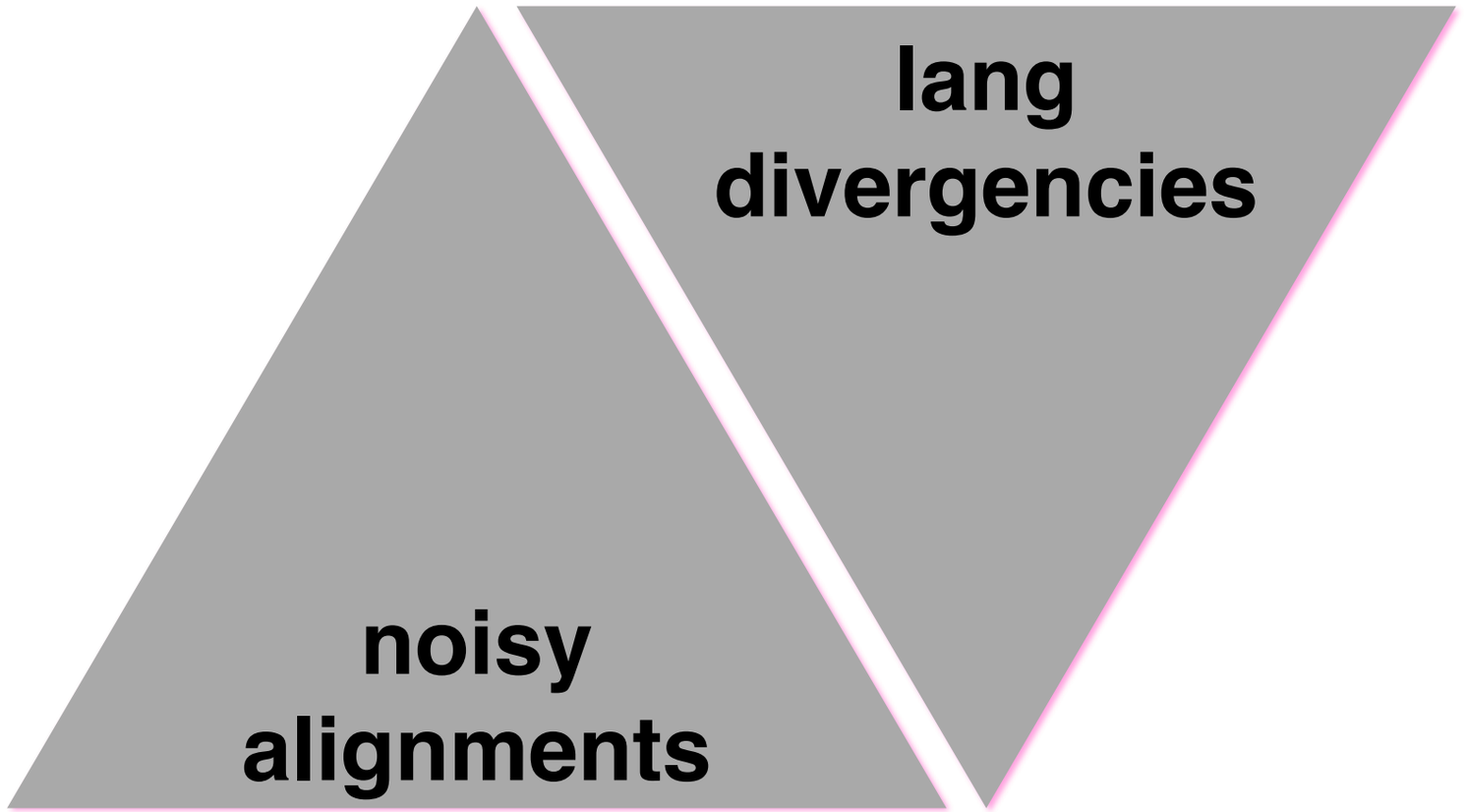
# Results

| | approach | F1 |
|---|---|---|
| **Postolache et al.** | POS tagging, syntactic heads, language-dependent word aligner | 63.9 |
| **EN-DE** | statistical word alignment | 54.6 |
| **EN-RU** | | 71.0 |

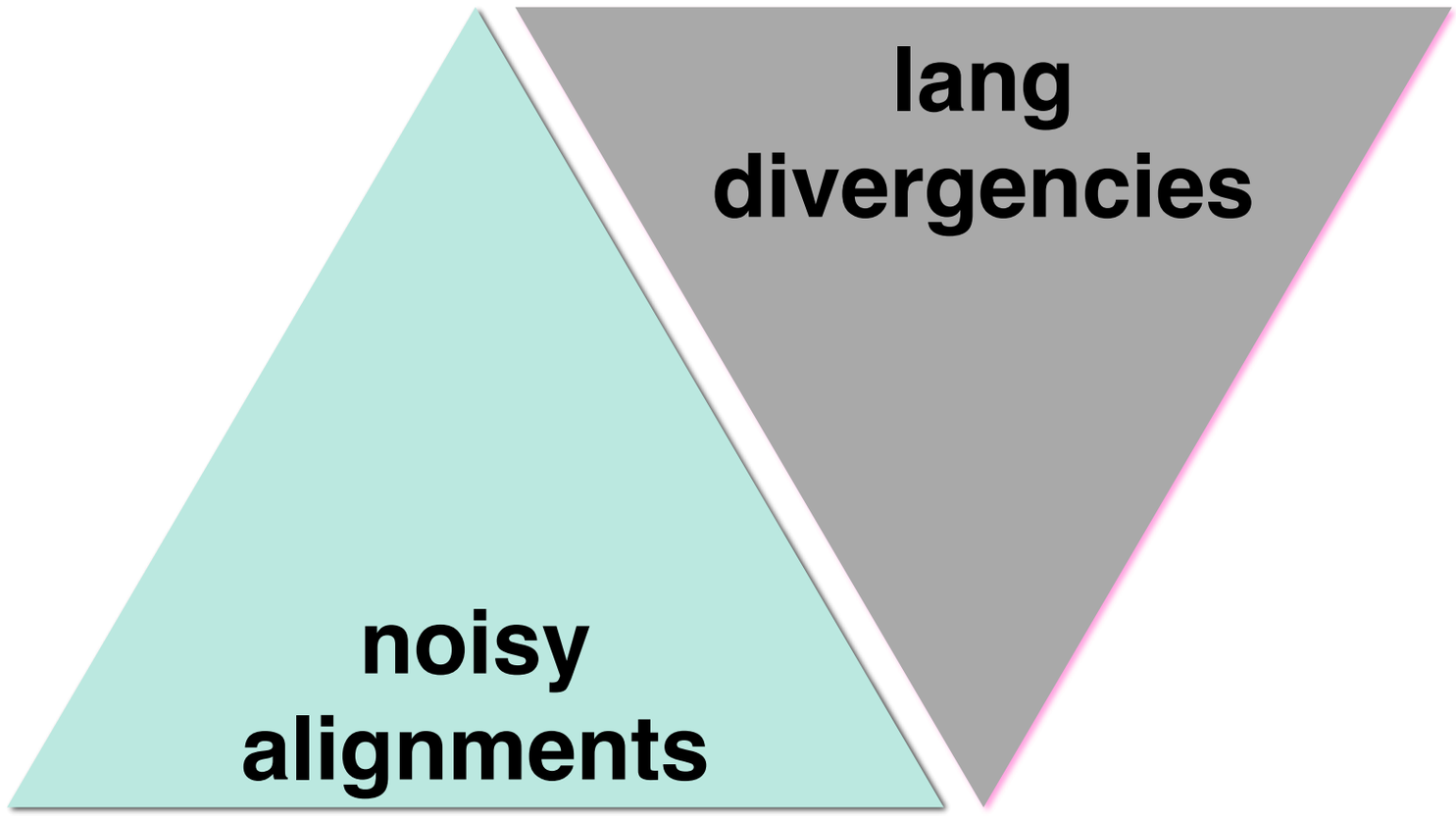
# F1-scores for different genres



# Error analysis: two-fold



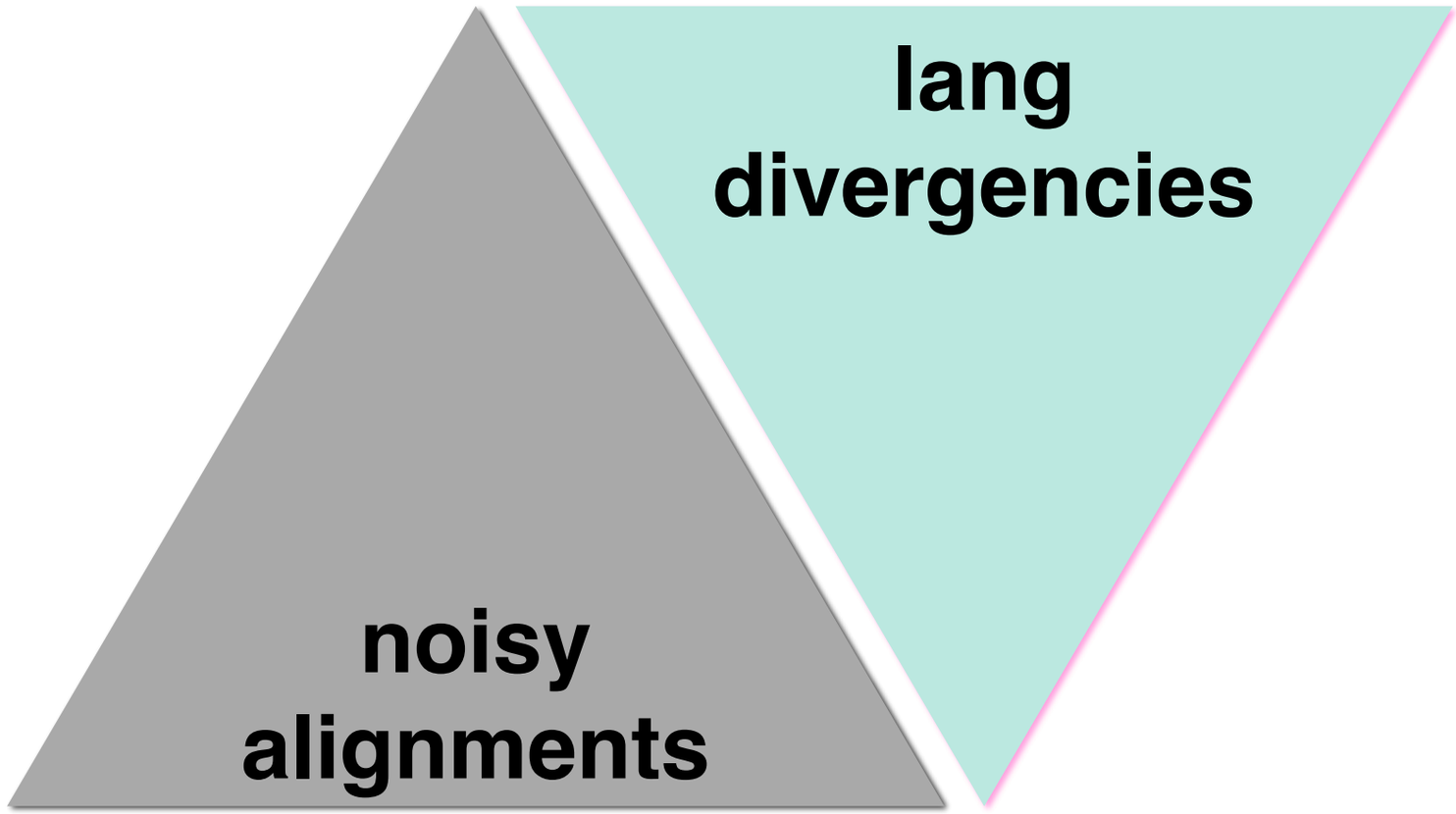
# Error analysis: two-fold



# Word alignment quality

| | Bisents | P | R | F1 |
|---|---|---|---|---|
| **Padó (2007)** | 1,029,400 | 98.6 | 52.9 | 68.86 |
| **Spreyer (2011)** | 1,314,944 | 94.88 | 62.04 | 75.02 |
| **Our approach** | 205,208 | 92.95 | 51.23 | 66.05 |

# Error analysis: two-fold



# Typology of errors

**(1) Morphological differences**

(1) contractions

(2) compounds

**(2) Differences in NP syntax**

(1) the use of articles

(2) pre- & post-modification

**(3) Non-equivalences in translation**

(1) personal & indefinite pronouns

(2) relative clauses & participial constructions

# (1.1) Compounds

- Europeans, however, are prompt to criticise the US for any failure in their policy toward [minorities].

- Europäer sind allerdings schnell bereit, der USA jeden Fehler in ihrer Minderheitenpolitik vorzuwerfen.

# (2.1) The use of articles

- English allows the use of nouns with zero articles more frequently than German (Kunz, 2010)

  A. Lastly, the G-20 could also help drive momentum on climate change.

  B. Schließlich könnten die G-20 auch für neue Impulse im Bereich [des Klimawandels] sorgen.

- Named Entities

  C. Hamas - [die Hamas]

# (2.2) Pre- & post-modification

- German NPs allow more complicated pre-modification

A. Pakistan needs international help to bring hope to [the young people] [who] live there.

B. Pakistan braucht internationale Hilfe, um [den dort lebenden jungen Menschen] Hoffnung zu bringen.

C. Пакистан нуждается в международной помощи, чтобы дать надежду [молодым людям], [которые] там живут.

# (3.1) Personal & indefinite pronouns

- German indefinite pronoun *man*

A. [It] was pursing a two-pronged strategy.

B. Man verfolgte eine Doppelstrategie.

C. [Она] преследовала двойную стратегию.

# -> Outcomes II

- applied annotation projection to 2 language pairs
- knowledge-lean approach
- F1 54.6 (EN-DE), 71.0 (EN-RU)
- errors in alignments and lang divergencies

# III.

# indirect coreference in EN, DE, RU (bridging & near-identity)

# Goals

- introduce a typology for bridging relations
- use an existing one for near-identity & apply it to German
- validate on a corpus of different languages and genres

# Experiments

- Design for German - Apply on German - Transfer to English and Russian
- manual transfer
- aiming at automatic projection via parallel corpora

# Bridging & near-identity

- **Bridging:** indirect relations that can only be inferred based on the common knowledge shared by the speaker and the listener (e.g. part-whole, set-membership)

- **Near-identity:** two NPs are almost identical, but differ in one crucial dimension (e.g. time)

# Example: bridging

Daisy walked into **[the office]** and saw a bunch of flowers on **[the windowsill]**.

# Example: bridging

Daisy walked into **[the office]** and saw a bunch of flowers on **[the windowsill]**.

**-> [the windowsill]** ***PartOf*** ((window) *PartOf)* **[the office]**

# Example: bridging

*Daisy walked into **[the office]** and saw a bunch of flowers on **the road**.

??

# Example: near-identity

In the afternoon, **[the temperature]** rose to 20C. This morning **[it]** was 12C.

# Example: near-identity

In the afternoon, **[the temperature]** rose to 20C. This morning **[it]** was 12C.

**-> same referent [temperature], different numerical values**

# Bridging: 2 viewpoints

- Information Status
  - an IS subcategory, along with given, new, etc.(Gardent et al., 2003), (Nissim et al, 2004), (Ritz et al., 2008), (Riester et al., 2010), (Markert et al., 2012)
- **Coreference**
  - a separate coreference relation, e.g. part-whole, set-membership (Poesio et al., 2004), (Poesio and Artstein, 2008), (Nedoluzhko et al., 2009)

# Annotation

- bridging (Clark, 1975) and near-identity (Recasens et al., 2010)

- German side of the corpus

- 2 annotators (half of the corpus)

- bridging: examine all definite NPs that are not linked to anything

- near-identity: check all NPs

# Parallel corpus

| | #EN | #DE | #RU |
|---|---|---|---|
| Documents | 14 | 14 | 10 |
| Sentences | 589 | 598 | 431 |
| Tokens | 11908 | 11894 | 8106 |
| REs | 1350 | 1395 | 1085 |
| Coreference chains | 259 | 273 | 188 |
| Bridging markables | 188 | 432 | 188 |

(Grishina, 2016)

# Example annotation

[Daisy Hamilton] was a private detective. [She] was thirty years old and [she] has been a detective for the past two years. Every morning [Daisy] went to **[[her] office]_B1** to wait for phone calls or open **[[the door]_B1]** to clients needing [her] services. One day somebody knocked on [the door].

# Annotation: bridging

- **PART-WHOLE**
  - *the telephone - the receiver*
- **SET-MEMBERSHIP**
  - *the European Union - the least developed countries*
- **ENTITY-ATTRIBUTE/FUNCTION**
  - *Kosovo - the current policy of rejection*
- **EVENT-ATTRIBUTE**
  - *the regional conflict - the trained fighters*
- **LOCATION-ATTRIBUTE**
  - *Germany - in the south*

# Annotation: bridging

- Annotation principles

  ➡ semantic relatedness

  ➡ proximity

  ➡ identity < near-identity < bridging

  [The telephone] rang. I went into [the office] and picked up [the receiver].

# Annotation: bridging

- Annotation principles

  ➡ semantic relatedness

  ➡ proximity

  ➡ identity < near-identity < bridging

  [The telephone] rang. I went into [the office] and picked up [the receiver].

# Annotation: bridging

- Annotation principles

  ➡ semantic relatedness

  ➡ proximity

  ➡ identity < near-identity < bridging

[The telephone] rang. I went into [the office] and picked up [the receiver].

# Results: bridging

| | Poesio (2004) | Nedoluzhko et al. (2009) | **This work** |
|---|---|---|---|
| Anaphor selection (F-1) | 0.22 | 0.5 | **0.64** |
| Antecedent selection (F-1) | N/A | N/A | **0.79** |
| Relation assignment (Cohen's kappa) | N/A | 0.9 | **0.98** |

# Annotation: near-identity

- **NAME METONYMY**
  - *the US (geographical entity) - the US (the government)*
- **MERONYMY**
  - *the president - the US (=the president)*
- **SPATIO-TEMPORAL FUNCTION**
  - *Budapest - the medieval Budapest*

(from Recasens et al., 2010)

# Results: near-identity

- small amount of near-identity links in the corpus, insufficient to compute the IAA

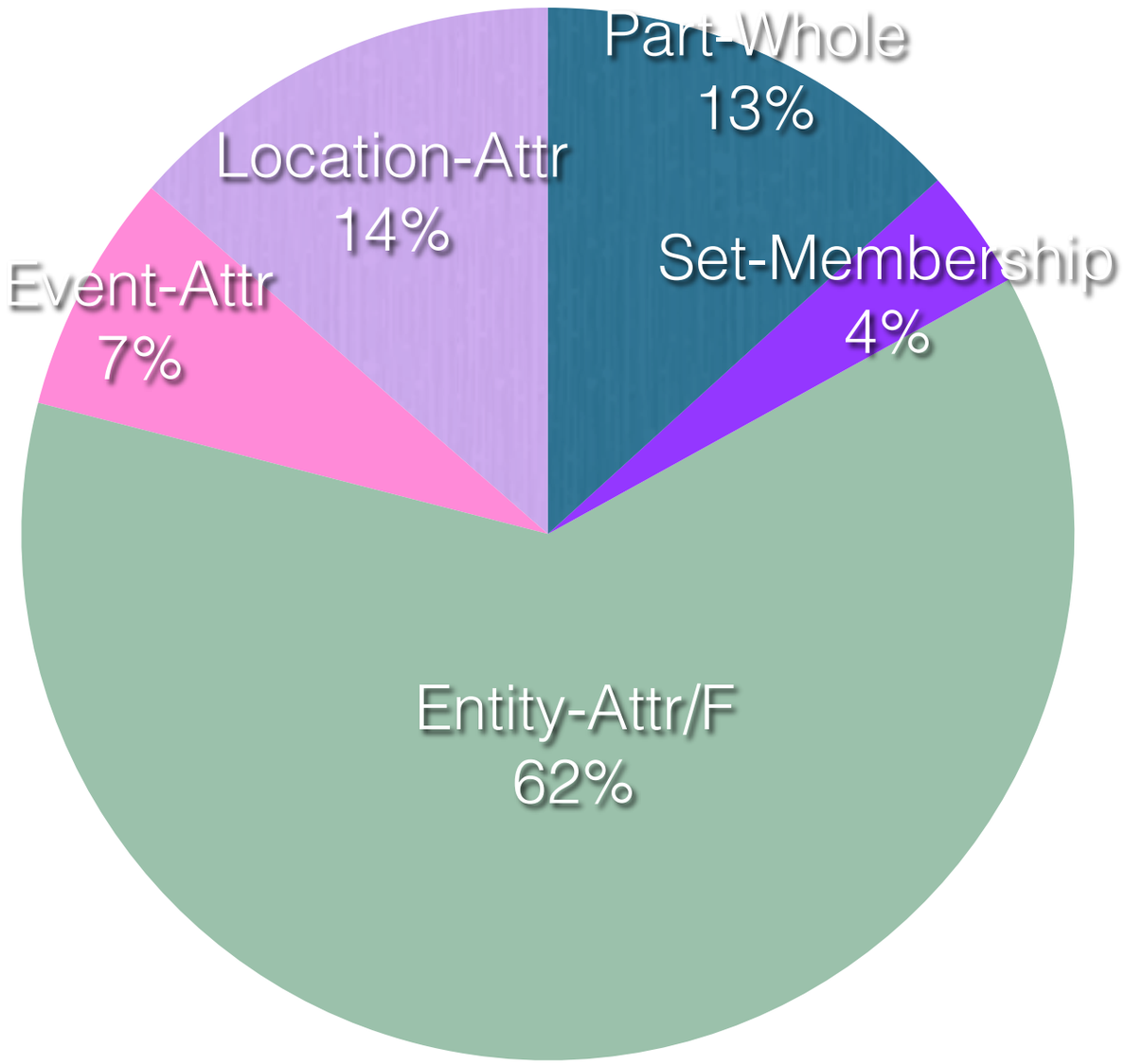
- for German, it conforms to the results of (Recasens et al., 2012)

- -> it is difficult to annotate near-identity explicitly

# Results: near-identity

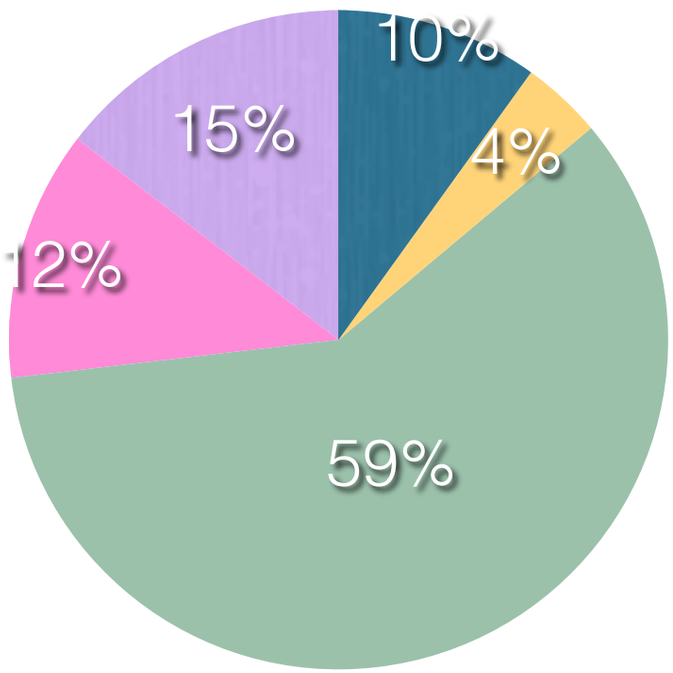
| Relation | News | Narrative | Medicine |
|---|---|---|---|
| Metonymy | 15.79 | 100.0 | 0.0 |
| Meronymy | 76.32 | 0.0 | 28.57 |
| Spatio-temporal function | 7.89 | 0.0 | 71.43 |
| Other | 0.0 | 0.0 | 0.0 |

# Distribution of bridging relations (DE)

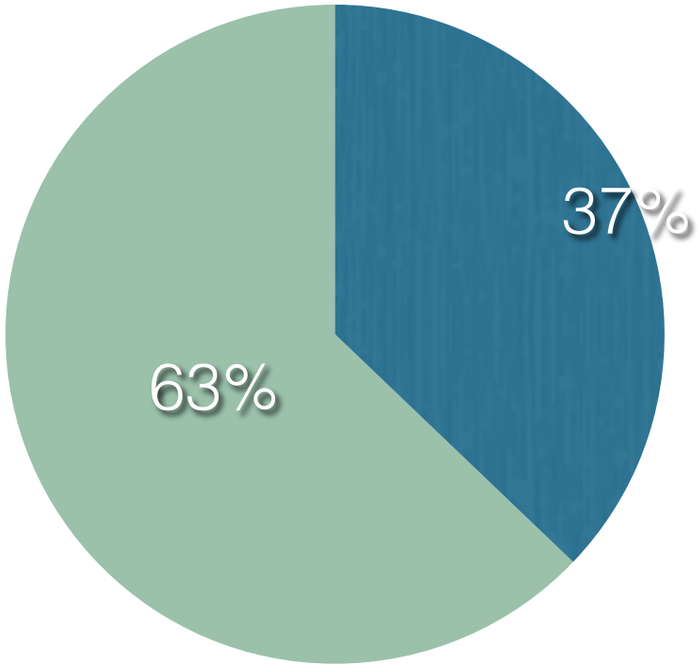


# Distribution of bridging across genres (DE)

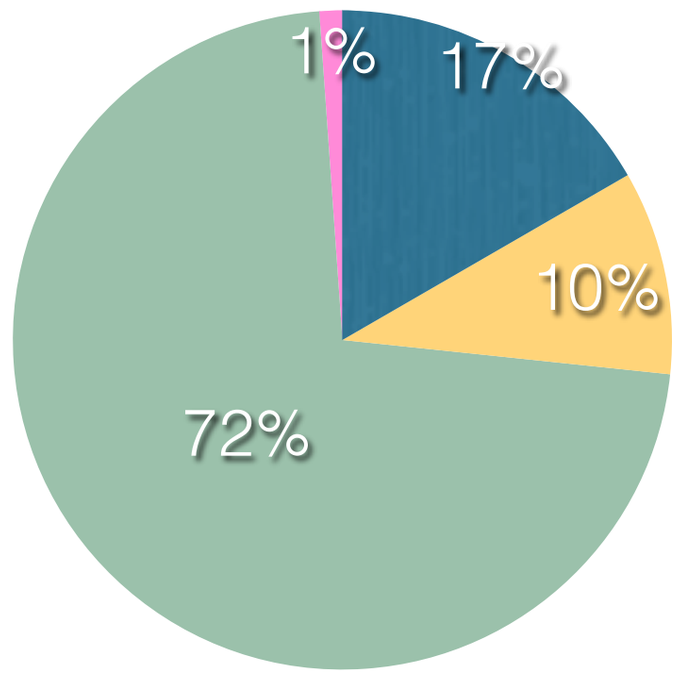
News




Narratives




Medicine



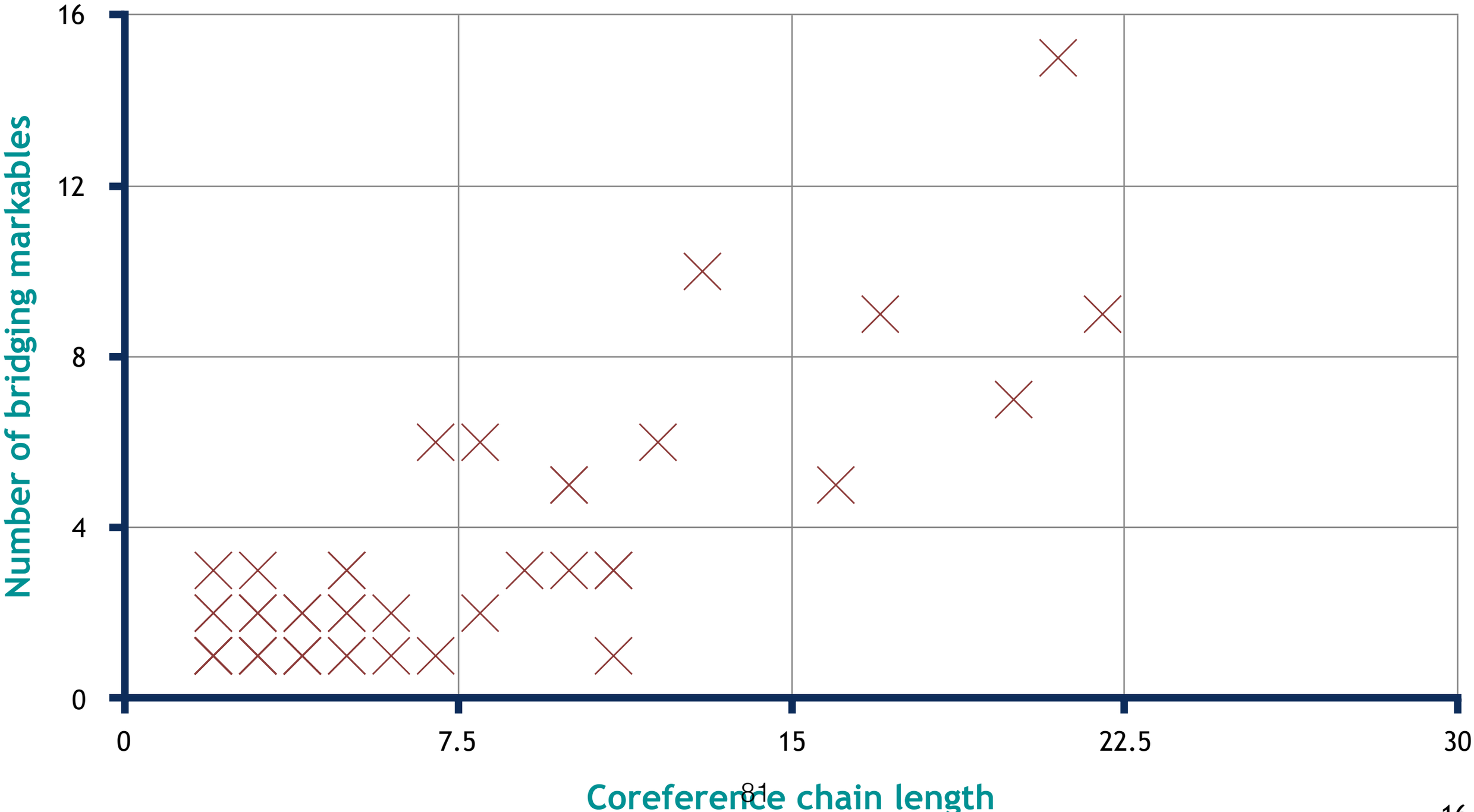

- Part-Whole
- Set-Membership
- Entity-Attr/F
- Event-Attr
- Location-Attr

# Coreference & bridging

- 17% bridging markables that start a coreference chain
- -> bridging entities are not as important on their own in the text
- 56% coreference chains that have bridging markables connected to them
- -> bridging markables are important for coreference entities

# Coreference & bridging

## Length of identity chains and number of their bridging markables



# Bridging distance

- anaphora+cataphora: 20.55 tokens (av. sentence length = 24.87 tokens)
- cataphora: -3.6 tokens
- anaphora: 30.96 tokens
- distance does not correlate with prominence

# Transfer

- looking at German, we annotated English and Russian
- 44% of the German markables transferred
- -> newswire was the most problematic genre

# Transfer

- [Die Terroranschläge in Mumbai im letzten Monat] sollten nicht nur die Wirtschaft und das Sicherheitsgefühl Indiens treffen. <...> [Die Täter] haben weder ihre Gesichter verhüllt noch sich selbst in der Manier von Selbstmordattentätern in die Luft gesprengt.

- [Last month's terrorist assault in Mumbai] targeted not only India's economy and sense of security. <...> [The attackers] did not hide their faces or blow themselves up with suicide jackets.

# Transfer

- [Die Terroranschläge in Mumbai im letzten Monat] sollten nicht nur die Wirtschaft und das Sicherheitsgefühl Indiens treffen. <...> [Die Täter] haben weder ihre Gesichter verhüllt noch sich selbst in der Manier von Selbstmordattentätern in die Luft gesprengt.

- [Last month's terrorist assault in Mumbai] targeted not only India's economy and sense of security. <...> [The attackers] did not hide their faces or blow themselves up with suicide jackets.

# Russian?

- Strategy: Genitive test

Daisy was in [the office] when someone knocked on [the door].

✓ [the door] == [the door of the office]

# -> Outcomes III

- a typology of bridging relations
- annotation of bridging with high inter-annotator reliability in 3 languages and 3 text domains
- near-identity: application to German
- strong correlation between bridging and coreference

# Conclusions

- manually annotated multilingual parallel coreference corpus including near-identity and bridging
- applied a knowledge-lean projection approach & manually transferred bridging pairs
- our projection results are competitive as compared to recent work
- our approach is generalisable to other languages and datasets

# Future work

- Corpus extension & refining the typology of relations

- Multi-source annotation projection

- Annotated data & guidelines will be available in summer 2017:

  ang-cl.ling.uni-potsdam.de/resources

# thank you!

# References


- Agić, Željko, Dirk Hovy, and Anders Søgaard. "If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages." The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015). 2015.
- Agić, Željko, et al. "Multilingual projection for parsing truly low-resource languages." *Transactions of the Association for Computational Linguistics* 4 (2016): 301-312
- Clark, Herbert H. "Bridging." Proceedings of the 1975 workshop on Theoretical issues in natural language processing. Association for Computational Linguistics, 1975.
- Ehrmann, Maud, Marco Turchi, and Ralf Steinberger. "Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection." *RANLP*. 2011.
- Gardent, Claire, Helene Manuelian, and Eric Kow. "Which bridges for bridging definite descriptions?." *4th International Workshop on Linguistically Interpreted Corpora-LINC'03*. 2003.
- Grishina, Yulia and Manfred Stede. “Knowledge-lean projection of coreference chains across languages”. *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, Beijing, China , page 14. Association for Computational Linguistics, 2015.
- Guillou, Liane, et al. "ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT." 9th International Conference on Language Resources and Evaluation (LREC), MAY 26-31, 2014, Reykjavik, ICELAND. European Language Resources Association, 2014.
- Hovy, Eduard, et al. "OntoNotes: the 90% solution." Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics, 2006.
- Krasavina, Olga, and Christian Chiarcos. "PoCoS: Potsdam coreference scheme." *Proceedings of the Linguistic Annotation Workshop*. Association for Computational Linguistics, 2007.
- Kunz, Kerstin Anna. 2010. Variation in English and German Nominal Coreference. A Study of Political Essays. Frankfurt am Main: Peter Lang.
- Markert, Katja, Yufang Hou, and Michael Strube. "Collective classification for fine-grained information status." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.
- Müller, Christoph, and Michael Strube. "Multi-level annotation of linguistic data with MMAX2." *Corpus technology and language pedagogy: New resources, new tools, new methods* 3 (2006): 197-214.
- Nedoluzhko, Anna, et al. "Extended coreferential relations and bridging anaphora in the prague dependency treebank." *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*. 2009.

# References


- Nissim, Malvina, et al. "An Annotation Scheme for Information Status in Dialogue." *LREC*. 2004.
- Padó, Sebastián. "Cross-Lingual Annotation Projection Models for Role-Semantic Information. Saarland University dissertation. Published as Volume 21, Saarbrücken Dissertations in Computational Linguistics and Language Technology." *German Research Center for Artificial Intelligence (DFKI) and Saarland University* (2007).
- Poesio, Massimo, and Ron Artstein. "Anaphoric Annotation in the ARRAU Corpus." *LREC*. 2008.
- Poesio, Massimo, et al. "Learning to resolve bridging references." Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.
- Postolache, Oana, Dan Cristea, and Constantin Orasan. "Transferring coreference chains through word alignment." *Proceedings of LREC-2006*. 2006.
- Recasens, Marta, Eduard H. Hovy, and Maria Antònia Martí. "A Typology of Near-Identity Relations for Coreference (NIDENT)." *LREC*. 2010.
- Recasens, Marta, Maria Antònia Martí, and Constantin Orasan. "Annotating Near-Identity from Coreference Disagreements." *LREC*. 2012.
- Riester, Arndt, David Lorenz, and Nina Seemann. "A Recursive Annotation Scheme for Referential Information Status." *LREC*. 2010.
- Ritz, Julia, Stefanie Dipper, and Michael Götze. "Annotation of Information Structure: an Evaluation across different Types of Texts." *LREC*. 2008.
- Spreyer, Kathrin. "Does it have to be trees?: Data-driven dependency parsing with incomplete and noisy training data." PhD Thesis. 2011.
- Tiedemann, Jörg. "News from OPUS-A collection of multilingual parallel corpora with tools and interfaces." *Recent advances in natural language processing*. Vol. 5. 2009.
- Tiedemann, Jörg. "Rediscovering Annotation Projection for Cross-Lingual Parser Induction." *COLING*. 2014.
- Yarowsky, David, Grace Ngai, and Richard Wicentowski. "Inducing multilingual text analysis tools via robust projection across aligned corpora." *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001.
- Zhekova, Desislava. *Towards Multilingual Coreference Resolution*. Diss. Bremen, Universität Bremen, Diss., 2013, 2013.