Experiments on bridging across languages and genres

Yulia Grishina

Applied Computational Linguistics
University of Potsdam / Germany

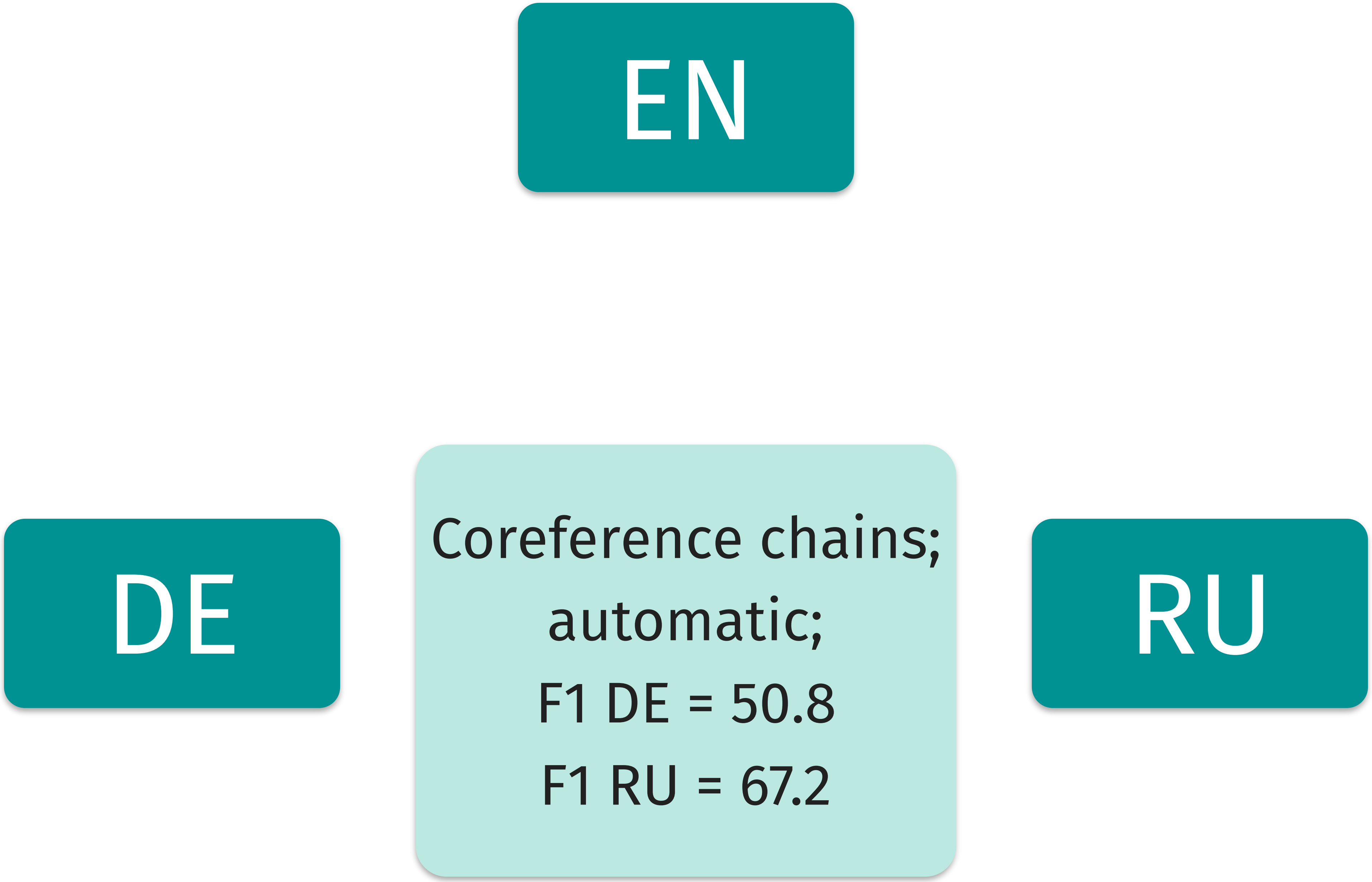
Goals

- 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

Annotation projection



EN

DE

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

RU

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)

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

- common coreference annotation guidelines (based on PoCoS (Krasavina & Chiarcos, 2007), OntoNotes (Hovy et al., 2006))
- uniform annotations in 3 languages
- NP coreference: full NPs, proper names, pronouns
- already annotated with identity coreference
- annotation tool: MMAX-2 (Müller & Strube, 2006), subsequently converted into CoNLL-2012 format

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 and Stede, 2015)

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*
Av. chain length	4.5	4.2	4.3	10.4	11.04	10.0	3.6	3.5	*5.2*	*5.1*	*12.3*

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

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*

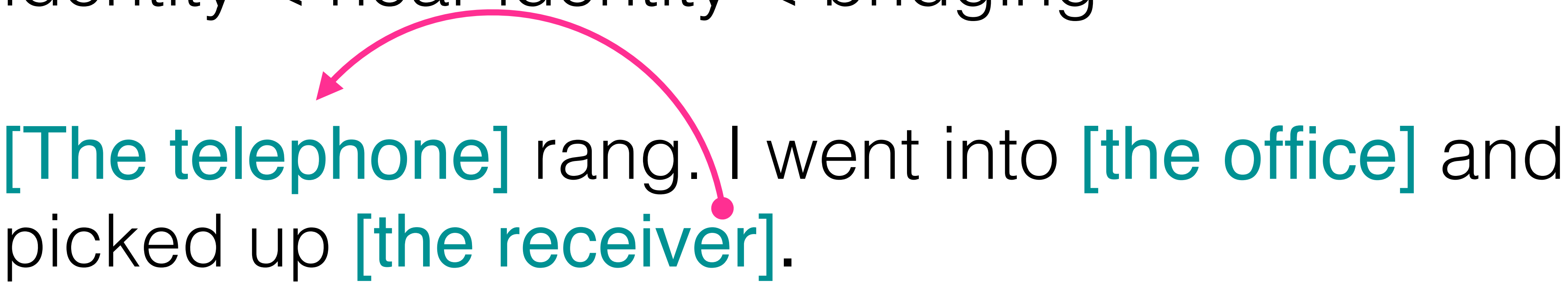
Annotation: bridging

- Annotation principles
  - ➡ semantic relatedness
  - ➡ proximity
  - ➡ identity < near-identity < bridging

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

Annotation: bridging

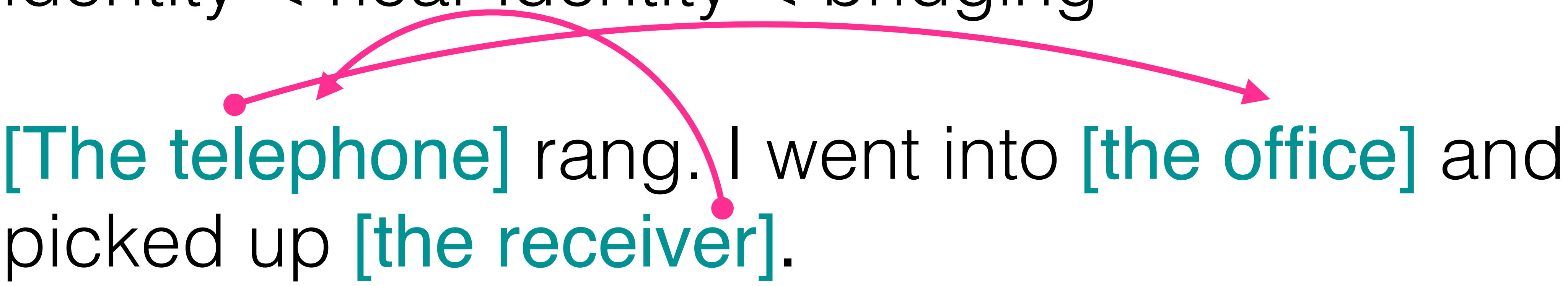
- Annotation principles
  - ➡ semantic relatedness
  - ➡ proximity
  - ➡ identity < near-identity < bridging



[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)

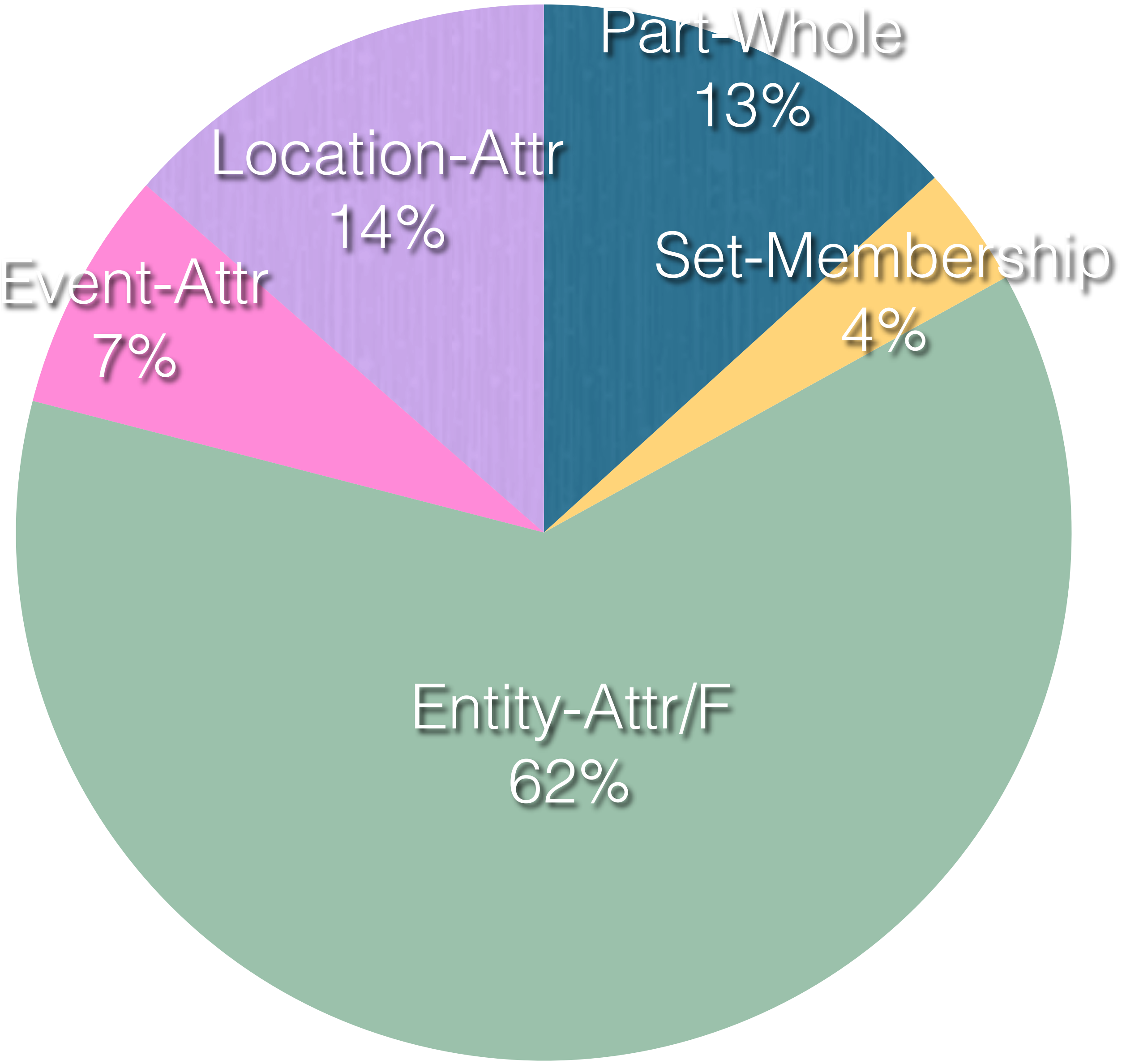
Results: near-identity

- 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

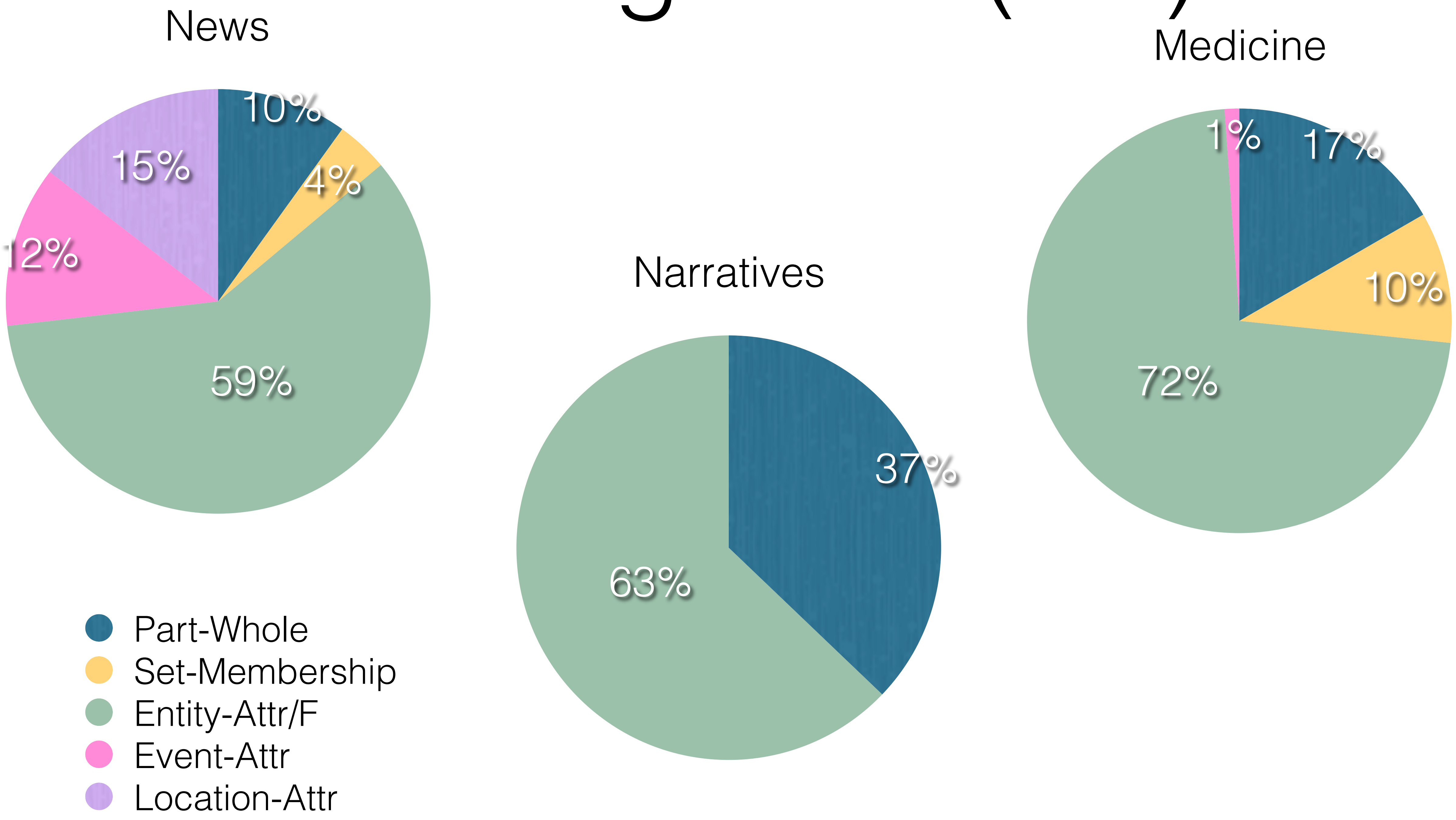
Results: near-identity

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)



Coreference & bridging

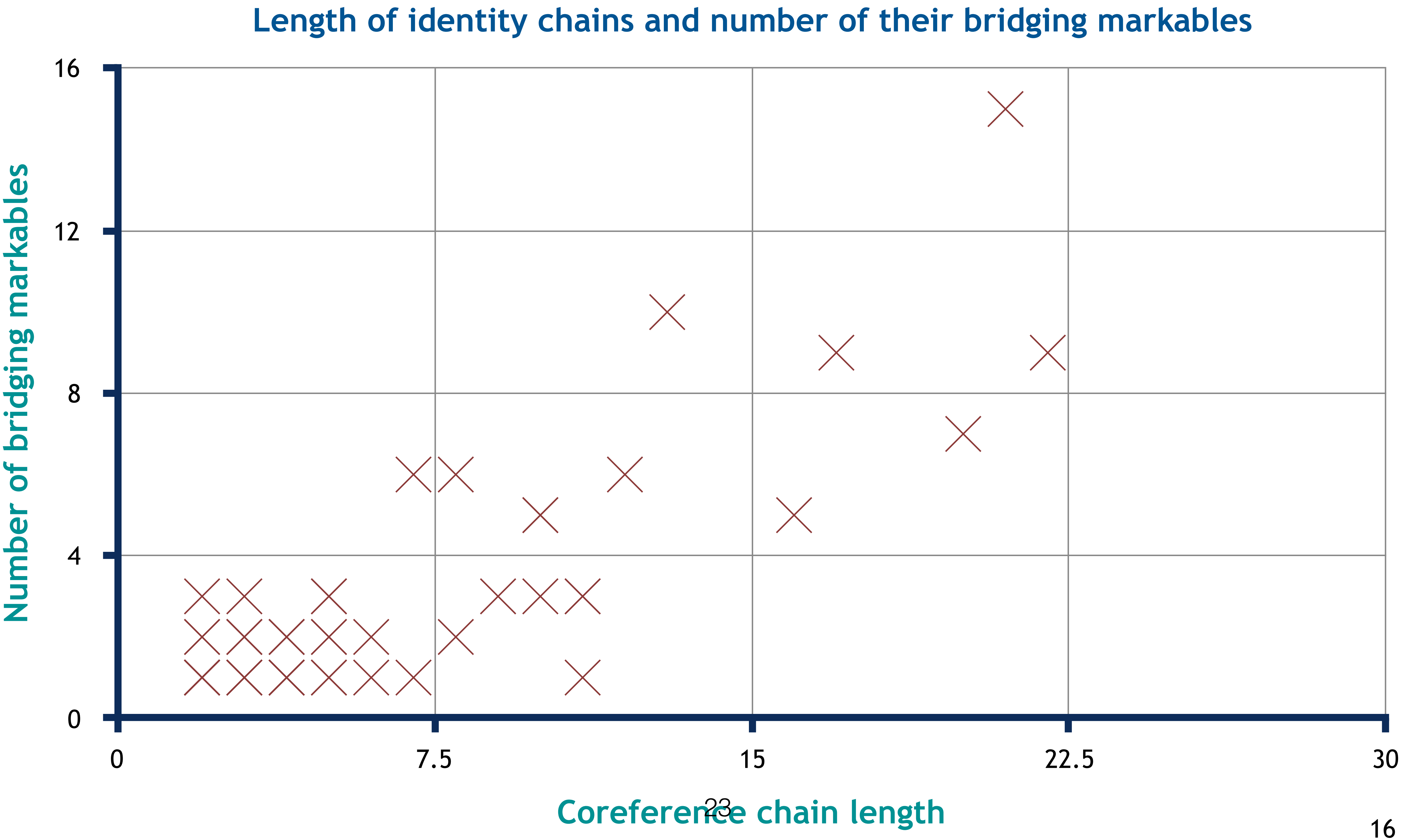
- 17% bridging markables that start a coreference chain

- -> 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



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

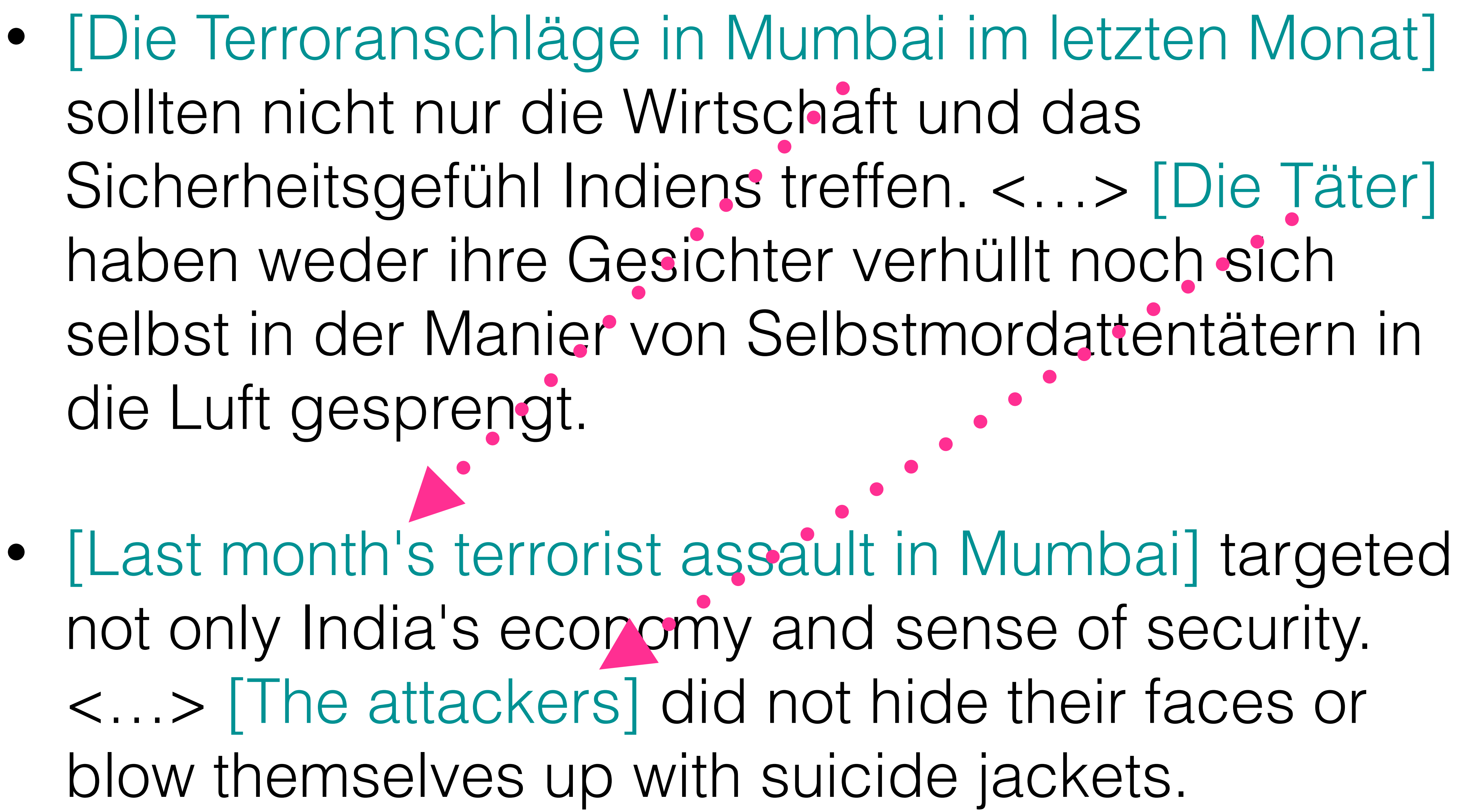
- looking at German, we annotated English and Russian
- 44% of the German markables transferred
- -> newswire was the most problematic genre

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.

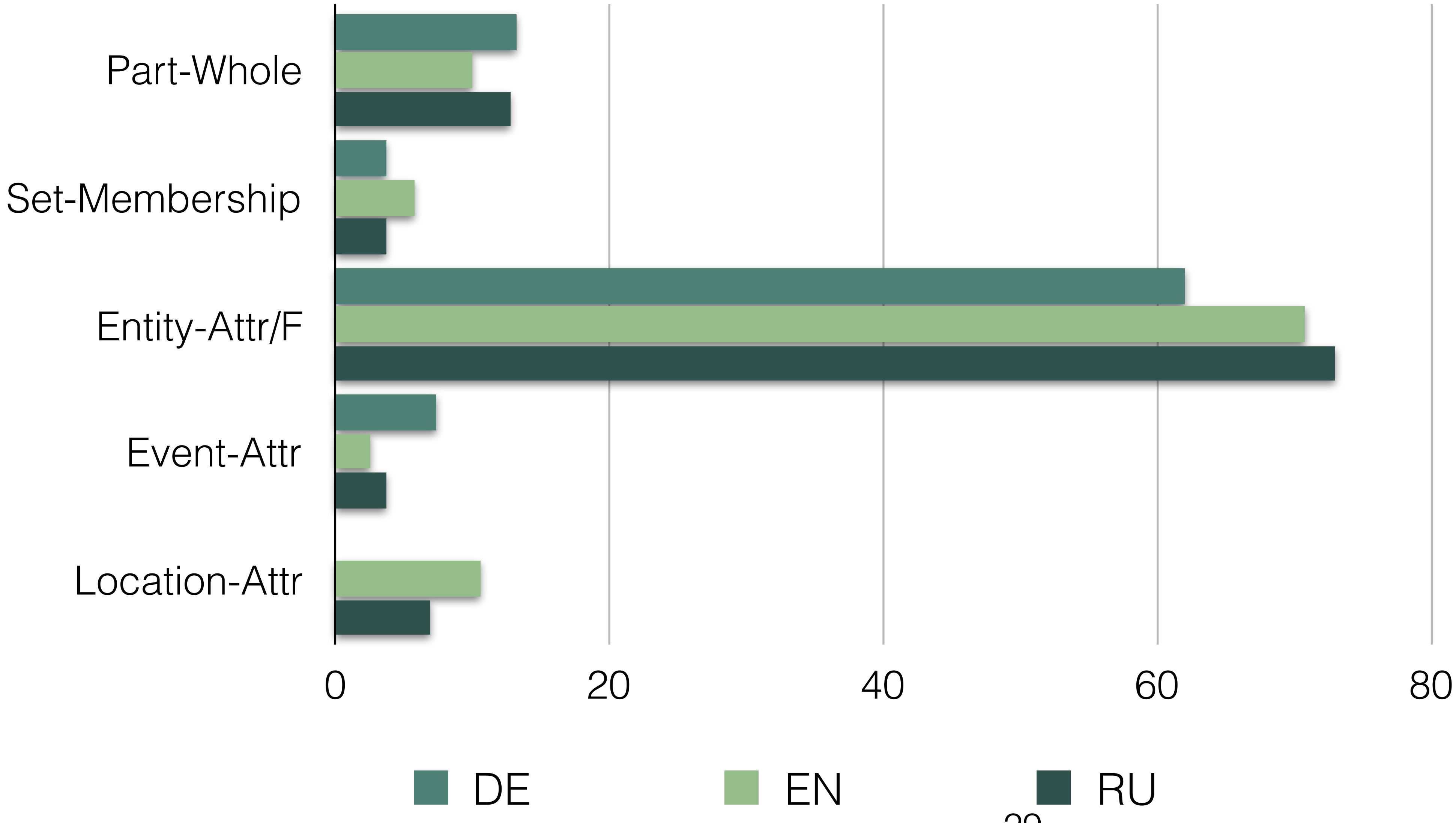
Russian?

- Strategy: Genitive test

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

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

Distribution of relations across languages



Outcomes

- 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

Thank you!