# Inter-Annotator Agreement in Coreference Annotation of Polish

CORE

Mateusz Kopeć and Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

## POLISH COREFERENCE CORPUS

### Annotation layers

The Polish Coreference Corpus (PCC) contains following levels of manual annotation which were investigated for inter-annotator agreement:

- mentions — nominal groups referencing discourse-world objects,
- mention semantic heads — the most relevant word of the group in terms of meaning; typically equal to syntactic head, but different for numerals or elective expressions (cf. $one_{synh}$ *of the* $girls_{semh}$),
- identity clusters — groups of mentions having the same referent,
- near-identity links — associations between a pair of semi-identical mentions, carrying some of their properties (cf. *prewar Warsaw* and *Warsaw today*),
- dominating expressions — a mention in a cluster which carries the richest semantics or describes the referent with most precision.

### The manual annotation process

- Text in PCC is a random 250-350 token full-paragraph sample from the National Corpus of Polish.
- 210 texts (with 60,674 tokens) from PCC (full corpus has over 500,000 tokens) were annotated independently by two annotators (hence: annotators A and B, yet there were more than 2 persons involved).
- There were 15 texts from each of 14 PCC domains.
- The annotation was performed in a customized MMAX2 tool.
- Texts were automatically pre-annotated before manual annotation.

## MENTION-LEVEL AGREEMENT

### Boundaries

- Chance agreement was not taken into account, as no standard exists for estimation of the chance-based factor for the task of marking up mentions (which can be nested, discontinuous and overlapping).
- Table 1 presents observed agreement while regarding annotation A as gold and B as system. Exact boundaries mean matching only the mentions consisting of exactly the same tokens, while heads only compares only head token for each mention.

| Match type | A | B | A ∩ B | P | R | F1 |
|---|---|---|---|---|---|---|
| Exact boundaries | 20,420 | 20,560 | 17,530 | 85.26% | 85.85% | 85.55% |
| Heads only | 19,394 | 19,522 | 18,317 | 93.83% | 94.47% | 94.14% |

Table 1 : Mention boundaries agreement

### Heads

- Head agreement was investigated only for common mentions of A and B.
- For 17,363 shared mentions out of 17,530 the same heads were marked, which gives the observed agreement: $p_{A_O} \approx \mathbf{99.05\%}$.
- The chance agreement ($p_{A_E}$) was calculated on a basis that each annotator chooses random word of a mention as its head. Chance agreement yielded: $p_{A_E} \approx \mathbf{68.32\%}$. This value is high due to a high count of one-token mentions, having the chance agreement equal to $\mathbf{1}$.
- Chance corrected $S$ inter-annotator agreement measure was therefore equal to $S = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} \approx \mathbf{97.00\%}$.

### Dominating expressions

- Calculated for 6,162 non-singleton mentions annotated by both A and B.
- Chance agreement analysis was not carried out since apart from choosing a mention as the dominating expression the annotator could also enter an arbitrary text value, which makes chance agreement estimations impossible.
- 4,115 mentions ($\approx \mathbf{66.78\%}$) shared the same dominating expression.
- 1,146 out of 1,818 cluster representatives ($\approx \mathbf{63.04\%}$) had the same dominating expression in both annotations.

## RELATION-LEVEL AGREEMENT

### Near-identity

- Near-identity agreement was investigated for common mentions of A and B.
- For each mention pair in a text the annotator could decide on their linking.
- We calculated Cohen's $\kappa$ for coincidence tables of these linking decisions for each text separately and then averaged it.
- When text did not contain any links by any annotator, agreement value was $\mathbf{1}$. When one annotator did not mark any link while the second one did, the agreement was $\mathbf{0}$. Per-text-and-annotator probability distribution was assumed.
- Applying this procedure to all texts we have calculated the average $\kappa \approx \mathbf{22.20\%}$. The result is low which can be interpreted as a difficulty in linking mentions with near-identical relation. The notion seems vague — in 128 cases mention pairs were marked as near-identical by one annotator and at the same time as purely-identical (i.e. were clustered) by the other annotator.

### Coreference

- Coreference clustering agreement was investigated only for common mentions of A and B.
- The agreement of coreference annotation is equal to 79.08% when calculated using weighted Krippendorff $\alpha$. It reaches 59.54% according to the MASI metric (both measures applied as proposed by Passoneau).
- Coincidence matrix of A and B regarding the decision, whether given mention is a singleton or not, is presented in Table 2. This table yields observed agreement of $\approx \mathbf{87.46\%}$ and expected agreement of $\approx \mathbf{51.32\%}$ which results in Cohen's $\kappa \approx \mathbf{74.24\%}$. This approach to agreement is similar to Recasens work.

| | | Annotation B | |
|---|---|---|---|
| | | **Clustered** | **Singleton** |
| **Annotation A** | **Clustered** | 6,238 | 975 |
| | **Singleton** | 1,223 | 9,094 |

Table 2 : Inter-annotator agreement on singleton/non-singleton decision for mentions

- Newly proposed BLANC-type agreement measure uses coreferential and non-coreferential links for all mention pairs (as in BLANC metrics). Decisions about each pair are used to create a coincidence matrix for each text and calculate Cohen's $\kappa$ using a per-text-and-annotator probability distribution. This value is then averaged to get final chance corrected agreement evaluation of $\mathbf{77.50\%}$.

## CONCLUSIONS

### Conclusions

- Coreference is more of a semantic and conceptual phenomenon which cannot reach scores as high as those achieved in lower-level linguistic tasks such as segmentation or morphosyntactic annotation.
- The average coreference agreement result of $\mathbf{77.50\%}$ seems to show the upper limit of coreference resolution capabilities, currently being reached by the state-of-the art tools for Polish.
- Results of near-identity annotation prove the difficulty of its reliable annotation in the current understanding of this phenomenon which should be verified in the further coreference annotation projects.

### Acknowledgements

The work reported was carried out within the "Computer-based methods for coreference resolution in Polish texts" (CORE) project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40). The work was also co-funded by the European Union from the resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications".