CORE

Interesting Linguistic Features in Coreference Annotation of Polish

Maciej Ogrodniczuk Katarzyna Głowińska Mateusz Kopeć Agata Savary Magdalena Zawisławska

CORE project

General information

Computer-based methods for coreference resolution in Polish project (CORE) financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40).

Project time frame: 2011–2014.

Project mission

Create methods and tools for **automated coreference resolution of Polish** (identity of reference) by preparation of:

- coreference annotation guidelines and methodology
- Polish Coreference Corpus (PCC) – a subset of the National Corpus of Polish (NKJP) manually annotated with coreferential chains
- IT tools for coreference resolution (rule-based, statistical, hybrid) and their evaluation.

ANNOTATION GUIDELINES

Potential vs. actual referentiality

Mentions are defined as nominal groups (NGs) taking into account their **potential referentiality**:

Nie wahał się włożyć kij w mrowisko. Mrowisko to, czyli cały senat uniwersytecki, pozostawało zwykle niewzruszone.

'He didn't hesitate to put a stick into an anthill (i.e. to provoke a disturbance). This anthill, i.e. the whole university senate, usually didn't care.'

Mention types

1. nouns, nominal phrases, personal pronouns
2. numeral groups (*trzy rowery* = ‘*three bicycles*’)
3. adjectival phrases with elided nouns (‘*bukiet z czerwonych kwiatów i z tych niebieskich*’ = ‘*a bouquet of the red flowers and these blue ones*’)
4. date/time expressions of various syntactic structures
5. coordinated nominal phrases, including conjoining commas (*krzesło, stół i fotel* = ‘*a chair, a table, and an armchair*’).

Allowed components

1. adjectives and adjectival participles in agreement (with respect to case, gender and number) with superior noun (*duży czerwony tramwaj* = ‘*big red tram*’)
2. subordinate nouns in the genitive case (*kolega brata* = ‘*my brother's colleague*’)
3. nouns in apposition (*malarz pejzażysta* = ‘*landscape painter*’, pol. ‘*painter landscapist*’)
4. subordinate prepositional-nominal phrases (*koncert na skrzypce i fortepian* = ‘*a concerto for violin and piano*’)
5. relative clauses (*dziewczyna, o której rozmawiamy* = ‘*the girl that we talk about*’).

Continuity

Discontinuous phrases and compounds are also marked:

To był delikatny, że tak powiem, temat.

‘It was a touchy, so to speak, subject.’

Nesting

The deep structure of phrases is marked as separate mentions when they don't contain finite verb forms having semantic heads other than those of the superior phrase (which reference different entities), are annotated:

dyrektor departamentu firmy

‘manager of a company department’

Coordination

For coordination both the individual constituents and the resulting compound are annotated (because they can be both referred to):

Jan z Marią przyszli na obiad. Oni są przemili, zwłaszcza Maria.

‘Jan and Maria have come to dinner. They are charming, especially Maria.’

EVALUATION

Inter-Annotator Agreement

Mentions	Dominant expressions	Semantic heads	Identity clusters
85.55%	66.78%	97.00%	79.08%

INTERESTING FEATURES

Zero subjects

- very frequent in most Slavic languages due to rich inflection of verbs
- null referent indicated by the morphology of the verb
- marked by including verbs (whose pronominal subjects are elided) into coreference clusters
- not considered for objects and complements.

Maria wróciła już z Francji. ØSpędziła tam miesiąc.

'Maria came back from France. ØHad$_{singular:feminine}$ spent a month there.'

Statistics: 4678 coreference clusters with at least one zero subject
26.89% of the total number of non-singleton clusters

Near-identity

Near-identity is a novel concept, in PCC including two phenomena:

1. two mentions refer to the *same* entity but text suggests the opposite (refocusing, e.g. *pre-war Warsaw* vs. *today's Warsaw*),
2. two mentions refer to *different* entities but the text suggests the opposite (neutralization, e.g. *wine* as a bottle vs. its contents).

- inspiring idea, but uncertain applicability or typology: $\kappa = 0.222$ (inter-annotator agreement in untyped near-identity links)
- the concept might be a result of mixing two different levels of language: the meaning of a word and its reference.

Dominant expressions

Dominant expression carries the richest semantics of the whole cluster (or describes the referent the most precisely):

- named entities
- periphrastic phrases that denote a particular object.

Cluster: David Beckham, rozgrywający Realu ‘David Beckham, Real playmaker’
Dominant expr.: David Beckham

Statistics: 62% selected from among NGs contained in the cluster
77% taken without any changes (= base form clustered)
38% given by the annotator

Dominant expressions can be annotated with a much higher reliability (66.78%) than near-identity.

Semantic heads

Semantic head is the most relevant word in the mention group in terms of meaning, typically equal to the syntactic head. Annotation reliability: 97.00%.

EXCEPTIONS: in numeral groups (e.g., *a lot of money*, *three of you*) the numeral is the syntactic head, and the noun is the semantic head.

CONCLUSION: coreference is a phenomenon on the level of semantics and discourse more than syntax.

Clustered indefinite pronouns

Originally indefinite, negative, reflexive, interrogative pronouns were excluded from annotation. Surprisingly, they can frequently form coreferential chains:

Jak ktoś jest zazdrosny, znaczy, że Ø naprawdę kocha.

‘If someone is jealous, it means, that (he/she) really loves.’

Open issues

- coreference in citations — meta- or standard text? preserved reference?
- links to superior phrase (*umowa zawarta przez zainteresowane nią strony* = ‘*contract concluded by parties interested in signing it*’) — traditionally excluded
- alternative coordination (*a or b*) — annotated as a whole?

POLISH COREFERENCE CORPUS

General information

- 540K tokens (one of the largest in the world)
- 1773 short texts (250-350 segments), 21 long texts
- shorts: 31 145 sentences (18 sentences / text, 16 segments / sentence)
- balanced representation of text genres.

Annotation statistics

	# mentions	# near-identity links	# singleton clusters	# non-singleton clusters
Short texts	167 765	4 288	103 938	17 326
Long texts	12 442	421	7 075	1 256
Total	180 207	4 709	111 013	18 582