PCC: Polish Coreference Corpus

Maciej Ogrodniczuk Mateusz Kopeć	Institute of Computer Science Polish Academy of Sciences
Katarzyna Głowińska	Lingventa
Agata Savary	Université François Rabelais Tours, France, Laboratoire d'informatique
Magdalena Zawisławska	Institute of Polish Language, Warsaw University

6th Language and Technology Conference
Poznań, December 9, 2013

The CORE project

Project factsheet:

- Computer-based methods for coreference resolution in Polish texts
- A National Science Centre grant 6505/B/T02/2011/40
- Duration: 2011-2014
- Principal investigator: Maciej Ogrodniczuk

Project summary:

1. Create innovative methods and tools for automated anaphora and coreference resolution in Polish texts
2. Create a corpus of Polish annotated with coreferential chains
3. Test various coreference resolution approaches on the annotated data (rule-based, statistical, hybrid etc.)

Text base of the corpus

CORE

Short texts:

- 1773 plain texts (31,136 sentences, 503,981 segments)
- 250-350 segments each (284 segments and 18 sentences/text)
- randomly selected from the National Corpus of Polish (NKJP)
- fragments of longer documents
  (but always full consecutive paragraphs)

Long texts:

- 21 complete documents (1,996 sentences, 36,234 segments)
- 1000-4000 segments each (1,725 segments and 95 sentences/text)
- selected from the Rzeczpospolita Corpus (KR)

Annotation

CORE

Automatic pre-annotation stages:

- morphosyntactic layer: Morfeusz SGJP
- sentence- and token-level segmentation and tagging: Pantera
- detection of mentions and coreference clusters: Ruler (using information from NERF and Spejd)

Manual annotation:

Pre-annotated:

- mention borders
- semantic heads of mentions
- coreferential mentions clustering

Not pre-annotated:

- near-identity relations
- dominant expressions in each cluster

Annotation statistics

CORE

Text type	# mentions	# near-identity links
short	167,871	4,699
long	12,561	407
any	180,432	5,106

Text type	# singleton clusters	# non-singleton clusters
short	102,218	17,630
long	7,166	1,259
any	109,384	18,889

Annotation scope

Mentions:

- all nominal groups (NGs) including pronouns
- zero subjects (marked at the verb)

Near-identity:

- coreference as a continuous relation rather than a discreet one
- inter-annotator agreement too low to consider this relation reliably annotated

Annotation scope

Mentions – examples:

- *Obecnie rząd pracuje nad zmianą ustawy, która przewiduje wykup mieszkań w towarzystwach budownictwa społecznego.*

  Currently, the government is working on amendments to the act, which provides purchase of flats in social housing associations.

- *. . . zwiększeniu o 25 mln zł wydatków przeznaczonych na staże i specjalizacje medyczne . . .*
  . . . the increase by 25 million zł spendings for internships and medical specialties . . .

- *Omotała mojego chłopca i porzuciła.*
  [She] entangled my boy and [she] abandoned [him].

Annotation scope

Near-identity – example:

- *Uroczysta inauguracja nowej Filharmonii Łódzkiej odbędzie się w budynku z surowym, niedokończonym frontonem. [. . .] Filharmonia nie zdąży z procedurami i wykonaniem szklanej tafli.*

  Inauguration of the new Philharmonic Orchestra will take place in a building with raw, unfinished pediment. Philharmonic won't manage to cope with procedures and implementation of the glass pane.

Annotation scope

Coreference clusters:

Only relations going beyond syntactic level are marked,
i.e. we do not cluster:

- nominal predicates
- appositions

Identity of reference only, i.e. we do not annotate:

- indirect (bridging or associative) anaphora
- discourse deixis
- ellipses (with the exception of zero anaphora)
- predicative and bound relations
- split antecedent
- identity of sense etc.

Annotation

CORE

Other original aspects of annotation:

- indicating the dominant expression, i.e. the expression that carries the richest semantics or describes the referent the most precisely
- indicating the semantic (rather than syntactic) head

Semantic head different than syntactic head – example:

- *czternaście tysięcy studentów*
  (fourteen thousand students)

Annotation

Dominant expression – example:

Forms in text:

- *pomocy zbożowej związanej ze stratami poniesionymi podczas powodzi (grain supply due to the flood damage)*
- *przekazanie rolnikom zboża (grain handover to farmers)*
- *pomocy poszkodowanym (help for victims)*

Dominant expression:

- *pomoc zbożowa związana ze stratami poniesionymi podczas powodzi (grain supply due to the flood damage)*

Annotation tools

CORE

DistSys:

- application for managing the distribution process of texts among annotators and adjudicators
- text fragments distributed from a central server
- local annotation
- upload of texts to the central repository.

MMAX:

- used for the annotation task of a single text
- a heavily modified version of the MMAX2 annotation tool
- new features: undo, adjudication plugin, etc.

`http://zil.ipipan.waw.pl/PolishCoreferenceTools`

Corpus availability

Various formats available:

- `TEI` – NKJP format extension. Each text has two more files:
  - file with mention boundaries and heads
  - file with coreference clusters, near-identities and dominant expressions
- `MMAX` – extended MMAX format. Each text is stored 3 files:
  - file with source text information
  - file with sentence, paragraph and word segmentation
  - file with mentions and coreference relations
- `BRAT` – extended BRAT format. Each text in two files:
  - file with raw text data
  - file with all annotation

`http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`

Corpus availability

CORE

BRAT version:

Summary of the PCC

Polish Coreference Corpus:

- is among the largest coreference corpora in the international community
- is manually validated
- is extensible
- evaluates concepts of near-identity, dominant expressions and semantic approach to identity-of-reference
- intends to boost linguistic studies on coreference phenomena, as well as the development of advanced text analysis tools for Polish

Thank you!

CORE

It's question time!

?

Representation of short text types

CORE

Type of text	Texts	Segments	%
Dailies	459	127,840	25.36
Magazines	406	117,694	23.35
Fiction literature	288	80,263	15.92
Non-fiction literature	96	27,743	5.50
Instructive writing and textbooks	100	27,728	5.50
Spoken – conversational	83	25,336	5.02
Internet non-interactive	63	17,734	3.51
Internet interactive	63	17,694	3.51
Misc. written	55	15,190	3.01
Spoken from the media	44	12,806	2.54
Quasi-spoken	43	12,783	2.53
Academic writing and textbooks	35	10,255	2.03
Journalistic books	19	5,492	1.08
Unclassified written	19	5,423	1.07
Any	1,773	503,981	100.00

Representation of long text types

Type of text	Texts	Segments	%
Journalism	3	7,078	19.53
Law	3	5,915	16.32
Economics	3	5,843	16.13
Domestic news	3	5,172	14.27
Sport	3	4,324	11.93
Culture	3	4,113	11.35
Science and technology	3	3,789	10.46
Any	21	36,234	100.00

Corpus availability

TEI — a NKJP-based format:

```
<!-- umiejętność logicznego myślenia -->
<seg xml:id="mention_8">
  <fs type="mention">
    <f name="semh" fVal="ann_morphosyntax.xml
                         #morph_1.1.23-seg"/>
  </fs>
  <ptr target="ann_morphosyntax.xml
               #morph_1.1.23-seg"/>
  <ptr target="ann_morphosyntax.xml
               #morph_1.1.24-seg"/>
  <ptr target="ann_morphosyntax.xml
               #morph_1.1.25-seg"/>
</seg>
```

Corpus availability

TEI — a NKJP-based format:

```
<!-- umiejętność logicznego myślenia;
     umiejętności te -->
<seg xml:id="coreference_0">
  <fs type="coreference">
    <f name="type" fVal="ident"/>
    <f name="dominant"
       fVal="umiejętność logicznego myślenia"/>
  </fs>
  <ptr target="ann_mentions.xml#mention_8"/>
  <ptr target="ann_mentions.xml#mention_14"/>
</seg>
```