Polish Coreference Corpus*

Maciej Ogrodniczuk[1], Katarzyna Głowińska[2], Mateusz Kopeć[1], Agata Savary[3], Magdalena Zawisławska[4]

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Lingventa
[3] François Rabelais University Tours, Laboratoire d'informatique
[4] Institute of Polish Language, Warsaw University

Abstract

This article describes the composition, annotation process and availability of the newly constructed Polish Coreference Corpus – a large Polish corpus of general nominal coreference. The tools used in the process and final linguistic representation formats are also presented.

Keywords: corpus, coreference, mention detection, anaphora

1. Introduction

The Polish Coreference Corpus (PCC) is a large manually annotated corpus of general Polish coreference, encoded in the extended format of the National Corpus of Polish – NKJP (Przepiórkowski et al., 2012). Its size is comparable to the anaphora annotation layer of the Polish KPWr corpus (Broda et al., 2012) but its scope is broader (e.g. coreference links are not restricted to named entities and markables are not limited to heads) and its development methodology includes revision of annotations. With a total number of approx. 540,000 tokens, the PCC is among the largest coreference corpora in the international community, together with Tüba/DZ (Hinrichs et al., 2005a) for German, NAIST Text (Iida et al., 2007) for Japanese, OntoNotes 2.0 (Pradhan et al., 2007) for English, Arabic and Chinese, the Prague Dependency Treebank (Nedoluzhko et al., 2009) for Czech and ANCOR (Muzerelle et al., 2013) for French.

This paper presents the composition of this (largely balanced) corpus, its annotation process and its availability.

2. Text base of the corpus

The PCC consists of two subcorpora:

- 1,773 "short" texts, i.e. containing 250-350 segments[1] in length, constituting fragments of longer documents (but always full consecutive paragraphs),
- 21 "long" texts – complete documents.

We believe that this composition allows for testing the correlation between length and completeness of Polish text and the nature of its coreferential links.

2.1. Short texts

"Short texts" are plain text fragments of randomly selected documents (of certain types, to create a balanced representation) from NKJP. For each document, paragraph sequences were also extracted randomly..

Short text types in PCC correspond to NKJP text types and text type representation is similarly balanced, matching the 1-million-word manually annotated subcorpus of NKJP. The number of texts, their size and the distribution of text genres is shown in Table 1.

Type of text	Texts	Segments	%
Dailies	459	127,840	25.36
Magazines	406	117,694	23.35
Fiction literature (prose, poetry, drama)	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 (static pages, Wikipedia)	63	17,734	3.51
Internet interactive (blogs, forums, usenet)	63	17,694	3.51
Misc. written (legal, ads, manuals, letters)	55	15,190	3.01
Spoken from the media	44	12,806	2.54
Quasi-spoken (parliamentary transcripts)	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

Table 1. Short text types in PCC

* The work reported here 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 paper is also co-founded by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications".

[1] Segments are word-like units established to reflect Polish morphosyntactic properties (and certain arbitrary decisions); e.g. agglutinants (*długo+śmy*), *-że* particle-adverb (*znasz+że*) etc. are all distinguished as separate segments. For details see Section 6.2.2 in (Przepiórkowski et al. 2012)

The subcorpus contains 1,773 short texts, 31,136 sentences and 503,981 segments, i.e. approx. 284 segments/text and 18 sentences/text. The average sentence length is 16 segments.

2.2. Long texts

"Long texts" are complete texts from the so-called Rzeczpospolita Corpus (RC; Weiss, 2002) – press articles retrieved in HTML from the online edition of Rzeczpospolita, one of the most prominent daily newspapers in Poland. The length of the selected texts varies from 1,000 to 4,000 segments. Collection of data, ultimately converted to plain text, has been performed semi-randomly (with interviews or documents combining a series of short press notes removed from the selection). Based on metadata present in the original HTML (DZIAL attribute) 7 most common text domains in RC were determined and 3 texts representing each domain have been included into PCC. Number of texts and their size in segments are shown in Table 2.

The subcorpus contains 21 texts, 1,996 sentences, 36,234 segments, which makes approx. 1,725 segments/text and 95 sentences/text. The average sentence length is 18 segments.

Domain	# 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

Table 2. Long text types in PCC

3. Annotation

3.1. Annotation levels

Extracted texts were automatically annotated with Morfeusz, a morphosyntactic analyser (Woliński, 2006), Pantera, a sentence- and token-level segmenter and morphosyntactic tagger (Acedański, 2010) and prepared for manual annotation (by means of automatic pre-annotation) with Ruler – a mention and coreference cluster detector (Ogrodniczuk and Kopeć, 2011). Segmentation and tagging errors were manually corrected only when errors introduced by the automatic tools would make coreference annotation impossible.

3.2. Annotation procedure

Pre-annotated texts have been evaluated by human annotators. Wherever the automatic annotation was wrong or unavailable, their task was to:

- mark mention borders,
- indicate semantic heads of mentions,
- mark near-identity relations,
- cluster coreferential mentions,
- indicate dominant expressions in each cluster.

The annotation procedure had two levels, with an expert adjudicator (super-annotator) verifying the process.

To calculate inter-annotator agreement, 210 short texts were processed in a different manner, with two annotators independently marking up the same text and the super-annotator solving problems in cases of disagreement.

Annotation statistics are shown in Table 3.

Type of text	# mentions	# near-identity[2] links	# singleton clusters	# non-singleton clusters
short	167,871	4,699	102,218	17,630
long	12,561	407	7,166	1,259
any	180,432	5,106	109,384	18,889

Table 3. Annotation statistics

3.2.2. Annotation guidelines

The PCC annotation schema and strategies conform with (Ogrodniczuk et al., 2013). The scope of annotation covers all nominal groups (NGs) including pronouns, since we consider the difference between an NG and a mention too controversial to be reliably decided in a general case. As far as introducing coreference links is considered, we limit ourselves to those semantic relations which cannot be deduced directly from syntax. Firstly, nominal predicates (Helena jest dyrektorką. 'Helena is the principal.') are never linked with their subjects (although, as all other NGs, they are considered mentions). Secondly, unlike in (Linguistic-Data-Consortium, 2006) and (Nedoluzhko et al., 2009), an apposition is not viewed as a sequence of coreferent mentions but as one mention only (*Oskarżony, mąż ofiary, ojciec trojga dzieci został dowieziony do sądu.' The accused, husband of the victim, father of three children was brought into court.'*). Thirdly, like (Hinrichs et al., 2005b), (Nedoluzhko et al., 2009) and (Recasens & Martí, 2010), we mark split NGs as unitary mentions (*To był delikatny, że tak powiem, temat. 'It was a touchy, so to speak, subject.'*). Finally, like (Osenova & Simov, 2004), (Pradhan et al., 2007), (Iida et al., 2007), and (Recasens & Martí, 2010), we take special care in annotating zero subjects, pervasive in Polish.

We take two coreferential relations into account: the **identity** (leading to splitting the set of mentions into clusters, i.e. equivalence classes) and – experimentally – the **near-identity** proposed by (Recasens et al., 2011). The definition of near-identity is interesting in that it allows us to see coreference in terms of a degree of identity rather than as a binary relation. Nevertheless the frequency of near-identity links introduced by our annotators, and the inter-annotator agreement are too low in our corpus to consider this relation as reliably

[2] See Section 3.2.2 for the explanation of the near-identity.

annotated. Due to the novel (wrt. Polish) character of our project, all relations different from identity and near-identity are outside the scope of annotation: indirect (bridging or associative) anaphora and discourse deixis (Hinrichs et al., 2005b; Poesio & Artstein, 2008; Nedoluzhko et al., 2009; Korzen & Buch-Kromann, 2011), ellipses (with the exception of zero anaphora), predicative and bound relations (Hendrickx et al., 2008), split antecedent (Hinrichs et al., 2005b), identity of sense (Iida et al., 2007), etc.

Besides annotating near-identity, other original aspects of our annotation schema are: (i) indicating the *dominant expression*, i.e. the expression that carries the richest semantics or describes the referent the most precisely, (ii) indicating the semantic (rather than syntactic) head.

3.2.3. Annotation tools

For the purpose of manual text annotation, two tools were used. The first one was DistSys – an application for managing the distribution process of texts among annotators and adjudicators inspired by the design of a similar tool created for NKJP annotation (Waszczuk et al. 2013). It is a general purpose tool, not focused on any specific type of annotation. It may serve any project if only the annotation task involves distributing text fragments from a central server among a number of annotators, annotating them locally (using some other application) and uploading them back to the central repository.

The second tool used is MMAX, a heavily modified version of the MMAX2 annotation tool by (Müller and Strube 2006), which was used for the annotation task of a single text (when it was acquired by the annotator via DistSys). As MMAX2 is a general annotation tool, for the sake of simplicity and annotation speed, many options were removed from the application. Some features were added, as requested by the annotators (for example the possibility of undoing the last change). The modifications include a superannotation plugin, which allows to see the annotation differences between two versions of the same text and easily merge them into one final version. Differences at each level are shown separately: an example of superannotating mention boundaries is depicted in Fig. 1. Each row represents one difference between annotators A and B: the first column describes which mention is relevant to the difference, the second column shows the decision of annotator A, the third column shows the decision of annotator B. In the first row, we can see that annotator A marked the mention "to" (plus), while the other one did not (minus). We should double click the plus or minus depending on the version we agree with to resolve to difference.

Both tools are available at the http://zil.ipipan.waw.pl/PolishCoreferenceTools web page.

Figure 1. Superannotation window in MMAX

4. Corpus availability

Polish Coreference Corpus is freely available for download under the Creative Commons Attribution 3.0 Unported License at the following address: http://zil.ipipan.waw.pl/PolishCoreferenceCorpus. There are 3 download formats, described briefly below. PCC is also available for browsing online (see Fig. 2) in a modified version of the Brat annotation tool (visualization tweaks were needed for the readability of long coreference chains). For a detailed description, visit the web page.

4.1. Brat

Brat is an online collaborative annotation environment, (Stenetorp et al., 2012) which uses a simple standoff annotation format described at http://brat.nlplab.org/standoff.html. Each text in this format is represented by two files: one containing raw text, the other one with information about mentions (marked as spans of characters in the former file) and relations between them (both coreference and near-identity).

4.2. MMAX

MMAX format is described in the MMAX2 manual (see mmax2.net). In this format, each text is stored in 3 files:

- a file with the ".mmax" extension, storing the text source (original NKJP text id) and text type,
- a file with the "_words.xml" ending, containing the text segmented into words, enriched with morphological annotation,
- a file with the "_mentions.xml" ending with information about mentions (represented as spans of words from the previous file), together with identity and near-identity relations between them.

4.3. TEI

PCC TEI format is an extension of the TEI format of the National Corpus of Polish. In addition to standard files:

- text_structure.xml
- ann_segmentation.xml
- ann_morphosyntax.xml
- header.xml

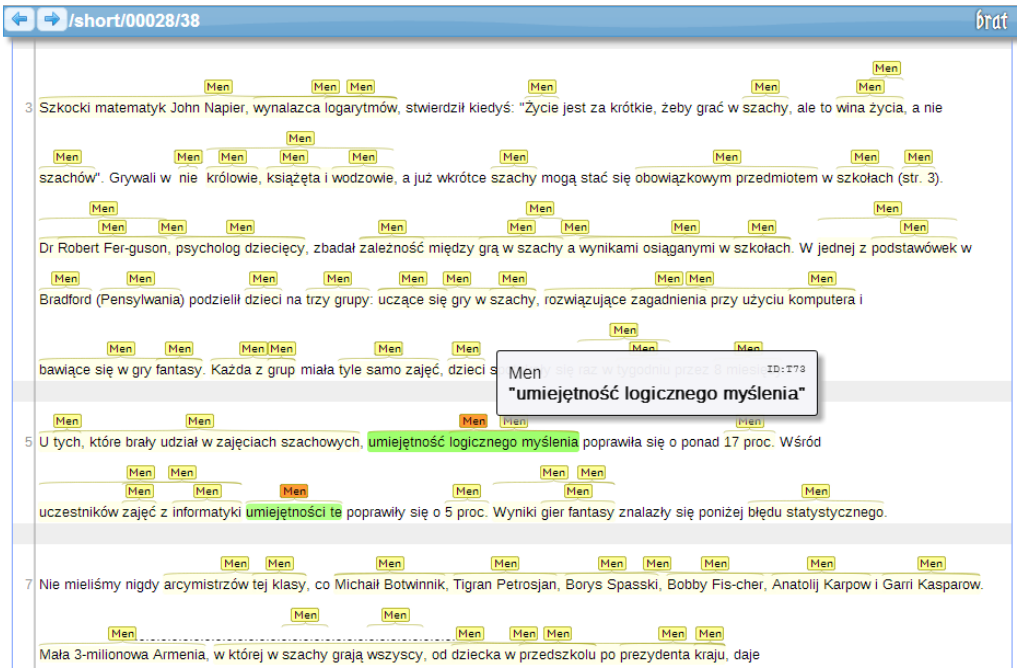

Fig. 2 Online corpus visualisation

each text in the corpus also has:

- ann_mentions.xml
- ann_coreference.xml.

The first file contains all the mentions, annotated as sets of segments from the ann_morphosyntax.xml file (similar to the named entity annotation in NKJP). In Fig. 3. we can see mention *"umiejętność logicznego myślenia"* ("*logical thinking ability"*) marked as a list of 3 pointers to segments in ann_morphosyntax.xml, out of which one is the head of the mention (as marked by the feature `<f name="semh">` in the feature structure `<fs name="mention">`).

```
<!-- 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>
```

Figure 3. Mention encoding in ann_mentions.xml

The second file provides the coreference and near-identity cluster information as groups of mentions from the former file. Fig. 4. presents the encoding of two relations: coreference identity cluster (containing mention_8 and mention_14) and near-identity relation (between mention_30 and mention_5). In the case of identity, encoding also contains the information about the dominant expression (`<f>` element with "dominant" name attribute).

```
<!-- 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>
...
<!-- filharmonia; nowa filharmonia -->
<seg xml:id="coreference_2">
  <fs type="coreference">
    <f name="type" fVal="near-ident"/>
  </fs>
  <ptr target="ann_mentions.xml#mention_5"/>
  <ptr type="source"
target="ann_mentions.xml#mention_30"/>
</seg>
```

Figure 4. Identity and near-identity encoding in ann_coreference.xml

5. Conclusions and perspectives

The Polish Coreference Corpus is a large, manually validated resource intended to boost linguistic studies on coreference phenomena, as well as the development of advanced text analysis tools for Polish, most prominently, computer coreference resolvers. It evaluates concepts of near-identity, dominant expressions and semantic approach to identity-of-reference which may contribute to a high-quality methodology for constructing similar corpora, particularly for other richly inflected languages.

The corpus can be further extended with other types of anaphoric and coreferential relations, such as identity-of-sense, bridging or bound anaphora as well as different

types of clustered mentions (e.g. verbal or adverbial constructs, references to relative clauses etc.)

References

Acedański, Sz. (2010). *A Morphosyntactic Brill Tagger for Inflectional Languages*. In Advances in Natural Language Processing, volume 6233 of Lecture Notes in Computer Science, pp. 3-14.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A ., Wardyński, A. (2012). *KPWr: Towards a Free Corpus of Polish*. In: Proceedings of LREC 2012.

Hendrickx, I., Gosse B., Daelemans W., Hoste V., Kloosterman G., Mineur A.-M., Van J., Vloet D. and Verschelde J.-L. (2008). *A Coreference Corpus and Resolution System for Dutch*. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), pp. 144–149. Marrakech, Morocco. European Language Resources Association (ELRA).

Hinrichs, E., Kübler S., Naumann K. and Zinsmeister H. (2005a). *Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank*. In Proceedings of the 27th Annual Meeting of the German Linguistic Association, Cologne, Germany.

Hinrichs, E. Kübler S. and Naumann K. (2005b). *A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations*. In Proceedings of the ACL Workshop on Frontiers In Corpus Annotation II: Pie In The Sky, pp. 13–20. Ann Arbor, Michigan, USA.

Iida, R., Mamoru K., Kentaro I. and Yuji M. (2007). *Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations*. In Proceedings of the Linguistic Annotation Workshop (LAW 2007), pp 132–139. Stroudsburg, PA, USA. Association for Computational Linguistics.

Korzen, I. and Buch-Kromann M. (2011). *Anaphoric relations in the Copenhagen Dependency Treebanks*. In Proceedings of DGFS workshop, pp. 83–98. Göttingen, Germany.

Linguistic-Data-Consortium. (2006). *ACE (Automatic Content Extraction) Spanish Annotation Guidelines for Entities*. Available at http://projects.ldc.upenn.edu/ace/docs/Spanish-Entities-Guidelines_v1.6.pdf (accessed on Feb. 18, 2013).

Muzerelle, J., Lefeuvre A., Antoine J.-Y., Schang E., Maurel D., Villaneau J. and Eshkol I. (2013). *ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement*. In Actes de la 20e conférence sur le traitement automatique des langues naturelles (TALN 2013), pp. 555–563. Les Sables d'Olonne, France.

Müller, Ch., Strube, M. (2006). *Multi-Level Annotation of Linguistic Data with MMAX2*. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.) Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. English Corpus Linguistics, vol. 3.

Nedoluzhko, A., Mírovský J., Ocelák R. and Pergler J. (2009). *Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank*. In Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), pp. 1–16. AU-KBC Research Centre, Anna University, Chennai Goa, India: AU-KBC Research Centre, Anna University, Chennai.

Ogrodniczuk M., Głowińska K., Kopeć M., Savary A., Zawisławska M. (2013) *Interesting Linguistic Features in Coreference Annotation of an Inflectional Language*. In M. Sun, M. Zhang, D. Lin, H. Wang (Eds.): 12th China National Conference on Computational Linguistics (12th CCL) and the 1st International Symposium on Natural Language Processing based on Naturally Annotated Big Data (1st NLP-NABD), Lecture Notes in Artificial Intelligence vol. 8202, pp. 97-108. Springer, Berlin-Heidelberg.

Ogrodniczuk, M. and Kopeć, M. (2011) *End-to-end coreference resolution baseline system for Polish*. In Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pp. 167-171. Wydawnictwo Poznańskie i Fundacja Uniwersytetu im. A. Mickiewicza.

Osenova, P. and Simov K. (2004). *BTB-TR05: BulTreeBank Stylebook. BulTreeBank Version 1.0*. Tech. Rep. BTB-TR05 Linguistic Modelling Laboratory, Bulgarian Academy of Sciences Sofia, Bulgaria.

Poesio, M. and Artstein R.. (2008). *Anaphoric Annotation in the ARRAU Corpus*. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco: European Language Resources Association.

Pradhan, S. S., Ramshaw L., Weischedel R., MacBride J. and Micciulla L. (2007). *Unrestricted Coreference: Identifying Entities and Events in OntoNotes*. In Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 446–453. Washington, DC, USA: IEEE Computer Society.

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B. (Eds.) (2012). *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*, Wydawnictwo Naukowe PWN, Warsaw.

Recasens, M., Hovy E. and Antònia Martí M. (2011). *Identity, non-identity, and near-identity: Addressing the complexity of coreference*. Lingua 121(6).

Recasens, M. and Antònia Martí M. (2010). *AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan*. Language Resources and Evaluation 44(4), pp. 315–345.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A., Lenart, M. (2013). *Annotation tools for syntax and named entities in the National Corpus of Polish*. In International Journal of Data Mining, Modelling and Management, Vol. 5, No. 2, Inderscience Publishers, pp. 103-122.

Weiss, D. (2002). *Korpus Rzeczpospolitej*. Retrieved from: http://www.cs.put.poznan.pl/dweiss/.

Woliński M. (2006). *Morfeusz — a practical tool for the morphological analysis of Polish*. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference, pp. 511–520. Wisła, Poland.