Polish Coreference Corpus – version 0.85

October 1, 2013

Mateusz Kopeć

Institute of Computer Science, Polish Academy of Sciences
m.kopec@ipipan.waw.pl

1 Introduction

The Polish Coreference Corpus (PCC) is a large manually annotated corpus of general Polish coreference. Its size is comparable to the anaphora annotation layer of the Polish KPWr corpus [2] 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 540,000 tokens, the PCC is among the largest coreference corpora in the international community, together with Tüba/DZ [3] for German, NAIST Text [5] for Japanese, OntoNotes 2.0 [16] for English, Arabic and Chinese, the Prague Dependency Treebank [11] for Czech and ANCOR [10] for French.

PCC is 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. Its detailed annotation guidelines present a detailed study of coreference-related phenomena ready for use in future coreference annotation projects. The corpus 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.

This document 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:

– 1773 "short" texts, i.e. consisting of 250-350 segments [1] in length, constituting fragments of longer documents (but always full consecutive paragraphs),

[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 [17].

– 21 "long" texts – complete documents of 1000-4000 segments in length.

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 fragments of randomly selected documents (of certain types, to create a balanced representation) from the National Corpus of Polish – NKJP [17]. For each document, paragraph sequences were also extracted randomly. Texts which had already contributed to PCC were omitted in subsequent draws, continued until the intended corpus size was achieved. Texts fragments were extracted in plain text form.

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 (1MNKJP). The number of texts, their size and the distribution of text genres is shown in Table 1.

Type of text	# texts	# segments	%
Dailies	459	127840	25,36%
Magazines	406	117694	23,35%
Fiction literature (prose, poetry, drama)	288	80263	15,92%
Non-fiction literature	96	27743	5,50%
Instructive writing and textbooks	100	27732	5,50%
Spoken – conversational	83	25336	5,02%
Internet non-interactive (static pages, Wikipedia)	63	17734	3.51%
Internet interactive (blogs, forums, usenet)	63	17694	3.51%
Misc. written (legal, ads, manuals, letters)	55	15190	3,01%
Spoken from the media	44	12806	2,54%
Quasi-spoken (parlia-mentary transcripts)	43	12783	2,53%
Academic writing and textbooks	35	10255	2,03%
Journalistic books	19	5492	1,08%
Unclassified written	19	5423	1,07%
Any	1773	503985	100,00%

Table 1. Short text types in PCC

The subcorpus contains 1 773 short texts, 31 145 sentences and 503 985 segments, i.e. approx. 284 segments/text and 18 sentences/text. Average sentence length is 16 segments.

2.2 Long texts

"Long texts" are complete texts from the so-called Rzeczpospolita Corpus (RC; [20]) – press articles retrieved in HTML from the online edition of Rzeczpospolita,

one of the most prominent daily newspapers in Poland. Text length is 1000-4000 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.

Domain	# texts	# segments	%
Journalism	3	7078	19,53%
Law	3	5915	16,32%
Economics	3	5843	16,13%
Domestic news	3	5172	14,27%
Sport	3	4324	11,93%
Culture	3	4113	11,35%
Science and technology	3	3789	10,46%
Any	21	36234	100,00%

Table 2. Long text types in PCC

The subcorpus contains 21 texts, 2 078 sentences, 36 234 segments, which makes approx. 1 725 segments/text and 99 sentences/text. Average sentence length is 17 segments.

3 Annotation

3.1 Annotation levels

Extracted texts were automatically annotated with Morfeusz, a morphosyntactic analyser [23], Pantera, a sentence- and token-level segmenter and morphosyntactic tagger [1] and prepared for manual annotation (by means of automatic pre-annotation) with Ruler – a mention and coreference cluster detector [13]. Segmentation errors were manually corrected only when errors introduced by automatic tool 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 super-annotator solving problems in cases of disagreement.

Annotation statistics are shown in Table 3.

Type of text	# mentions	# near-identity links	# singleton clusters	# non-singleton clusters
short	167765	4288	103938	17326
long	12442	421	7075	1256
any	180207	4709	111013	18582

Table 3. Annotation statistics

3.3 Annotation guidelines

In [12], we have presented the major aspects of the PCC annotation scope, schema and strategies. 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 [8] and [11], 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 [4], [11] and [19], we mark split NGs as unitary mentions (To był delikatny, że tak powiem, temat. 'It was a touchy, so to speak, subject.'). Finally, like [14,16,5], and [19], 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 [18]. 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, as discussed in [12], 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 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 [4,11,7,15], ellipses (with the exception of zero anaphora), predicative and bound relations [6], split antecedent [4], identity of sense [5], etc.

Besides annotating near-identity, other original aspects of our annotation schema, as discussed in [12] are:

- 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.

3.4 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 [22]. 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 [9], 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 Figure 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 web page:
`http://zil.ipipan.waw.pl/PolishCoreferenceTools`.



Fig. 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 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).

4.1 Brat format

Brat is an online collaborative annotation environment, [21] 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:

- file with `.txt` extension – containing raw text,
- file with `.ann` extension with information about mentions (marked as spans of characters in the former file) and relations between them (both coreference and near-identity).

The file with `.ann` extension contains 4 types of rows, each one being tab-separated:

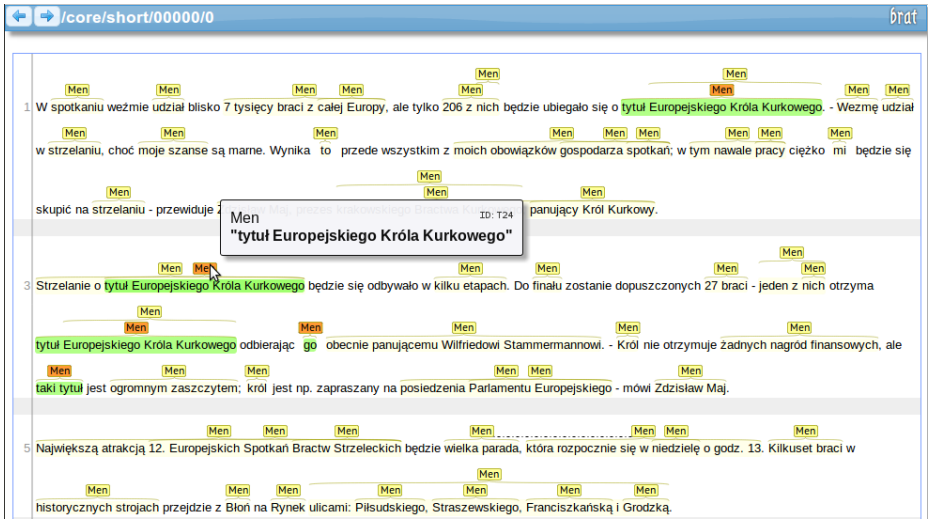


Fig. 2. Online corpus visualisation

1. Rows starting with letter `T`, representing mentions. Example:

   ```
   T24 Mention 370 398 Jeorjos Aleksandros Mangakis
   ```

   Tab-separated fields are:
   - `T24` is the mention's identifier.
   - `Mention 370 398` describes the offset of the first and last characters of the mention (in the file with `.txt` extension). In case of a discontinuous mention a semicolon will occur here, separating multiple continuous spans of the mention (for example `370 398;410 420`).
   - `Jeorjos Aleksandros Mangakis` is the text form of the mention.

2. Rows starting with the letter `R`, marking near-identity links. Example:

   ```
   R4 Quasi Arg1:T337 Arg2:T628
   ```

   Tab-separated fields are:
   - `R4` – link identifier.
   - `Quasi Arg1:T337 Arg2:T628` – describes the source (mention with id `T377`) and the target (mention with id `T628`) of the link.

3. Rows starting with the character `*`, marking non-singleton clusters. Example:

   ```
   * Coref T361 T362
   ```

   Tab-separated fields are:

   - `*` – the identity relation symbol.
   - `Coref T361 T362 T451` – `T361 T362 T451` are the identifiers of the mentions in the coreference cluster. There may be any number of mentions in a cluster (but at least two).
4. Rows starting with the character `#`, marking heads of mentions or their dominant expressions (where applicable – for non-singleton mentions). Examples:

   ```
   #2 empty T2 Head: "braci"
   #3 empty T2 Dominant: "7 tysięcy braci z całej Europy"
   ```

   Tab-separated fields are:
   - `#2` or `#3` – the identifier of annotation.
   - `empty T2` – `T2` is the identifier of the mention, which head or dominant expression is described.
   - `Head: "braci"` or `Dominant: "7 tysięcy braci z całej Europy"` – head of mention or dominant expression in quotes.

4.2 MMAX format

MMAX format is described in the MMAX2 manual (see `http://mmax2.net`). In this (slightly modified) format, each text is stored in 3 XML files:

- file with `.mmax` extension, storing the text source (original NKJP or RC text id) and text type,
- file with `_words.xml` ending, containing the text segmented into words, enriched with morphological annotation,
- file with `_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.

File with `mmax` extension stores the text source in the `title` tag, in `val` attribute. For example:

```
Paragraphs: p-57,p-58,p-59,p-60
from NKJP text with id: IPIPAN_1301919980826
```

means, that the text comes from NKJP text with `IPIPAN_1301919980826` identifier, more precisely it is formed from 4 its paragraphs, with identifiers: `p-57`, `p-58`, `p-59` and `p-60`. Text type, conforming to the tables 1and 2, is in the `val` attribute of the `catRef` tag.

File with `_words.xml` ending contains the text layer, with segmentation into words (`word` tags).

Each word has its identificator in `id` attribute, base form in `base` attribute, part of speech and morphosyntactic information in `ctag` and `msd` attributes.

Optional attributes are `hasNps`, `lastInSent`, `lastInPar`, occuring with `true` value when there is no space preceeding the word, the word is last in a sentence, the word is last in a paragraph, respectively.

File with `_mentions.xml` ending stores information about mentions (denoted as `markable` tags).

Each mention has `span` attribute, describing in terms of word identifiers from the file with `_words.xml` ending, which words the given mention contain. Identifiers are comma-separated, and two dots between identifiers represent all the words between the words, whose identifiers are before and after the dots.

The `mention_head` attribute contains semantic head of the mention.

The `near_identity` attribute has a value `empty` when the mention is not a source of near identity relation, or the identifier of the mention, connected with given mention with the near-identity relation.

The `mention_group` attribute can also be `empty`, which means that given mention is a singleton, or contain a value, which is a cluster identifier (for example `set_5`). All mentions with the same cluster identifier are in coreference relation.

The `dominant` attribute contains the dominant expression of a mention, which is in a non-sigleton cluster. It is the same for all mentions in a cluster.

4.3 TEI format

PCC TEI format is an extension of the TEI format of the National Corpus of Polish [17]. Each text is in separate directory, named with its identifier, and contains standard TEI files:

- `text_structure.xml`,
- `ann_segmentation.xml`,
- `ann_morphosyntax.xml`,
- `header.xml`,

and additionaly:

- `ann_mentions.xml`,
- `ann_coreference.xml`.

The `text_structure.xml` file contains the text divided into paragraphs. The `ann_segmentation.xml` file contains sentence and segment segmentation, the `ann_morphosyntax.xml` file has the morphosyntactic information about the segments. The `header.xml` file stores the information abour the type of the text (`catRef` tag, `target` attribute, conforming to the tables 1 and 2) and the text source – identificator of the original text and paragraphs (`title` tag).

The `ann_mentions.xml` file contains all the mentions (marked by `<seg>` tags), 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">`). Each mention is preceeded with the XML comment with the text form of the mention.

The `ann_mentions.xml` has also the segmentation of text into sentences and paragraphs in `<s>` and `<p>` tags.

The `ann_coreference.xml` file provides the coreference and near-identity cluster information as groups (in `<seg>` tags) of mentions from the `ann_mentions.xml` file. Each group is preceeded with a comment containing orthographical form of the mentions in the group. Fig. 4 presents the encoding of two relations: coreference identity cluster (containing `mention_8` and `mention_14`, with `ident` type) and near-identity relation (between `mention_30` and `mention_5`, with `near-ident` type). In the case of identity, encoding has also information about the dominant expression (`<f>` element with `"dominant"` `name` attribute). In case of near-identity, the group has always two elements and one of them is indicated to be the source of the relation (`type` attribute in `<ptr>` tag, value: `source`).

References

1. Szymon Acedański. A Morphosyntactic Brill Tagger for Inflectional Languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer, 2010.
2. Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*, Istanbul, Turkey, 2012. ELRA.
3. Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. Recent developments in linguistic annotations of the tüba-d/z treebank, 2004.
4. Erhard W. Hinrichs, Sandra Kübler, and Karin Naumann. A unified representation for morphological, syntactic, semantic, and referential annotations. In *IN ACL WORKSHOP ON FRONTIERS IN CORPUS ANNOTATION II: PIE IN THE SKY, ANN ARBOR*, 2005.
5. Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 132–139, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

```
<?xml version="1.0" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
 xmlns:nkjp="http://www.nkjp.pl/ns/1.0">
<TEI>
  <teiHeader/>
  <text>
    <body>
      <p xml:id="mentions_p-1" corresp="morph_1-p">
        <s xml:id="mentions_p-1.1-s" corresp="morph_1.1-s">
          ...
          <!-- 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>
          ...
          <!-- Europejskiego Króla Kurkowego  -->
          <seg xml:id="mention_9">
            <ptr type="nonhead" target="morph_1.1.46-seg"/>
            <ptr type="head" target="morph_1.1.47-seg"/>
            <ptr type="nonhead" target="morph_1.1.48-seg"/>
          </seg>
          ...
        </s>
        <s xml:id="mentions_p-1.2-s" corresp="morph_1.2-s">
          ...
          <!-- to -->
          <seg xml:id="mention_11">
            <ptr type="head" target="morph_1.2.2-seg"/>
          </seg>
          ...
        </s>
      </p>
      <p xml:id="mentions_p-2" corresp="morph_2-p">
      ...
      </p>
      ...
    </body>
  </text>
</TEI>
</teiCorpus>
```

Fig. 3. Mention encoding in `ann_mentions.xml`

```
<?xml version="1.0" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:xi="http://www.w3.org/2001/XInclude"
  xmlns:nkjp="http://www.nkjp.pl/ns/1.0">
  <xi:include href="NKJP_1M_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <body>
        <p>
          <!-- 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>
          ...
        </p>
      </body>
    </text>
  </TEI>
</teiCorpus>
```

Fig. 4. Identity and near-identity encoding in `ann_coreference.xml`

6. Frederik Coppens Walter Daelemans Veronique Hoste Geert Kloosterman Anne-Marie Mineur Joeri Van Der Vloet Iris Hendrickx, Gosse Bouma and Jean-Luc Verschelde. A coreference corpus and resolution system for dutch. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.
7. I. Korzen and Buch-Kromann M. Anaphoric relations in the copenhagen dependency treebanks. In *Proceedings of DGFS workshop*, pages 83–98, 2011.
8. Linguistic-Data-Consortium.
9. Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.
10. Judith Muzerelle, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, and Iris Eshkol. 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)*, pages 555–563, Les Sables d'Olonne, France, 2013.
11. A. Nedoluzhko, Mírovský J., Ocelák R., and Pergler J. Extended coreferential relations and bridging anaphora in the prague dependency treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, page 1–16.
12. Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In Maosong Sun, Min Zhang, Dekang Lin, and Haifeng Wang, editors, *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)*, volume 8202 of *Lecture Notes in Computer Science*, pages 97–108. Springer-Verlag, Berlin, Heidelberg, 2013.
13. Maciej Ogrodniczuk and Mateusz Kopeć. End-to-end coreference resolution baseline system for Polish. In Zygmunt Vetulani, editor, *Proceedings of the Fifth Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 167–171, Poznań, Poland, 2011.
14. P. Osenova and Simov K. Btb-tr05: Bultreebank stylebook. bultreebank version 1.0., 2004.
15. Massimo Poesio and Ron Artstein. Anaphoric annotation in the arrau corpus. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.
16. Sameer S. Pradhan, Lance Ramshaw, Ralph M. Weischedel, Jessica MacBride, and Linnea Micciulla. Unrestricted coreference: Identifying entities and events in ontonotes. In *ICSC*, pages 446–453. IEEE Computer Society, 2007.
17. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw, 2012.

18. Marta Recasens, Eduard Hovy, and M. Antònia Martí. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138 – 1152, 2011.
19. Marta Recasens and M.Antònia Martí. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44(4):315–345, 2010.
20. Korpus rzeczpospolitej. [on-line] `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.
21. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April 2012. Association for Computational Linguistics.
22. Jakub Waszczuk, Katarzyna Glowinska, Agata Savary, Adam Przepiórkowski, and Michal Lenart. Annotation tools for syntax and named entities in the national corpus of polish. *IJDMMM*, 5(2):103–122, 2013.
23. Marcin Woliński. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag, Berlin, 2006.