Polish Coreference Corpus*

Maciej Ogrodniczuk[1], Katarzyna Głowińska[2], Mateusz Kopeć[1], Agata Savary[3], and 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. The Polish Coreference Corpus (PCC) is a large corpus of Polish general nominal coreference built upon the National Corpus of Polish. With its 1900 documents from 14 text genres, containing about 540,000 tokens, 180,000 mentions and 128,000 coreference clusters, the PCC is among the largest coreference corpora in the international community. It has some novel features, such as the annotation of the quasi-identity relation, inspired by Recasens' near-identity, as well as the mark-up of semantic heads and dominant expressions. It shows a good inter-annotator agreement and is distributed in three formats under an open license. Its by-products include freely available annotation tools with custom features such as file distribution management and annotation adjudication.

Keywords: corpus, coreference, mention detection, anaphora

1 Introduction

One of the main challenges in linguistics is to understand how entities of the language refer to those of the discourse world. Modelling and studying this phenomenon – as many others – is frequently based on corpus annotation. Since discourse world referents are hard to represent, instead of representing reference phenomena directly, one usually builds coreference chains between linguistic entities and considers those chains (or clusters) abstract representatives of referents. Additionally, other coreference-related (non-transitive) relations, such as bridging anaphora, near-identity or quasi-identity (introduced here), find other specific representations in coreference annotation schemas.

Coreference-annotated corpora of considerable size have an increasingly rich bibliography and concern about a dozen languages from several language families (cf. [12, Chapter 3]). In this paper we present one of these resources, the Polish Coreference Corpus (PCC) [12], a large manually annotated corpus of general Polish coreference, encoded in an extended format of the National Corpus of

* 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).

Polish – NKJP [20]. 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 approx. 540,000 tokens, the PCC is among the largest coreference corpora in the international community, together with Tüba/DZ [5] for German, NAIST Text [6] for Japanese, OntoNotes 2.0 [18] for English, Arabic and Chinese, the Prague Dependency Treebank [10] for Czech and ANCOR [9] for French.

We describe the composition of this (largely balanced) corpus, its annotation process and results, as well as its availability and future work.

2 Text base of the corpus

The PCC consists of two subcorpora:

- 1773 "short" texts, i.e. containing 250-350 segments 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.

The subcorpus contains 1773 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) [19] – 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 1000 to 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 the metadata present in the original HTML (DZIAL attribute) 7 most common text domains in RC were determined and 3 texts representing each domain have

Table 1. Short text types in PCC

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	5492	1.08
Unclassified written	19	5423	1.07
Any	1773	503,981	100.00

been included into PCC. Number of texts and their size in segments are shown in Table 2.

The subcorpus contains 21 texts, 1996 sentences, 36,234 segments, which makes approx. 1725 segments/text and 95 sentences/text. The average sentence length is 18 segments.

Table 2. Long text types in PCC

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	36,234	100.00

3 Annotation

3.1 Annotation levels

Extracted texts were automatically annotated with Morfeusz, a morphosyntactic analyser [25], Pantera, a sentence- and token-level segmenter and morphosyn-

tactic tagger [1] and prepared for manual annotation (by means of automatic pre-annotation) with Ruler – a mention and coreference cluster detector [14]. 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 mention heads
- mark quasi-identity relations
- cluster coreferential mentions
- indicate dominant expressions in each cluster (see Section 3.3 for details).

For the large majority of the corpus the annotation methodology followed the so-called *series approach* in which each document was first reviewed by one human annotator, and his/her results were further corrected and validated by an adjudicator. This approach is non-standard for the NKJP corpus, where each previously performed annotation task followed a *parallel approach* with two independent annotators reviewing each document and an adjudicator comparing their decisions and solving discrepancies. We performed an annotation experiment [12, Chapter 6.2], which showed that, with equivalent human resources (two annotators and one adjudicator), the series annotation mode yields better results than the parallel annotation mode, as far as mention detection and coreference cluster markup is concerned. Conversely, the annotation of dominant expressions and of quasi-identity links was of a higher quality in the parallel mode. Since the two latter annotation aspects are of lower importance than the two former ones, we believe that the series mode is more appropriate for coreference annotation (but due to budgetary constraints, only one annotator instead of two, and one adjudicator, worked on each text). The final annotation statistics are shown in Table 3 (please see also [13] for a detailed analysis of various linguistic constructs in PCC).

Table 3. Annotation statistics

Type of text	**# mentions**	**# quasi-identity links**	**# singleton clusters**	**# non-singleton clusters**
Short	167,679	4699	102,160	17,636
Long	12,562	407	7167	1259
Any	180,241	5106	109,327	18,895

3.3 Annotation guidelines

The PCC annotation schema and strategies conform with [11]. 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 [10] and [3], 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 [5], [10] and [22], we mark split NGs as unitary mentions (*To był delikatny, że tak powiem, temat.* 'It was a touchy, so to speak, subject.'). Finally, like [6], [18], [22] and [16], 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 quasi-identity inspired by the concept of near-identity proposed in [21]. The four specific types of quasi-identity relation are: (i) a relation between a pair of mentions of which the second one distorts properties of the object, so that both of them begin to refer to the meta-object, e.g.: *Nie widziała „Przeminęło z wiatrem", ale czytała je.* 'She hasn't seen "Gone with the wind", but she has read it.'; (ii) a relation between a pair of mentions of which the second one is created by distinguishing a given property of the object called by the first reference, e.g.: *Warszawa jest pięknym miastem, ale przedwojenna Warszawa była jeszcze piękniejsza.* 'Warsaw is a beautiful city, but pre-war Warsaw was even more beautiful.; (these expressions refer to the same city, but from different periods); (iii) a relation between the name of the substance and the container in which the substance remains, e.g.: *Zdjął z półki wino i włożył je do koszyka.* 'He put the wine down from the shelf and put it into the basket.'; (iv) a relation between the set (described by pluralia tantum, collectiva, nouns in plural with numerals) and its distinguished element, e.g.: *Rodzice przyszli na zebranie. Jeden z nich poruszył problem agresji wśród dzieci.* 'Parents came to the meeting. One of them touched upon the problem of aggression among children.'. However, the annotators were instructed to mark with this notion other close-to-identity relations, which are not characterised by identity or non-identity.

The definition of quasi-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 quasi-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 quasi-identity are outside the scope of annotation: indirect (bridging or associative) anaphora and discourse deixis [5], [10], [17] and [7], ellipses (with the exception of zero anaphora), predicative and bound relations [4], split antecedent [5], identity of sense [6], etc.

Besides annotating quasi-identity, other original aspects of our annotation schema are: indicating the dominant expressions in each cluster and marking semantic (rather than syntactic) mention heads.

Dominant expression is the expression that carries the richest semantics or describes the referent the most precisely. The best candidates for dominant expressions are proper names, descriptions of unequivocal reference, or expressions with richest semantics (hyponyms), most likely originally present in annotated texts, but often in inflected form (cluster: *Prezesa PKP* '(of the) CEO of PKP'; *go* 'him'; dominant expression: *Prezes PKP* 'CEO of PKP').

Semantic heads (i.e. the most important word from the point of view of mention's sense) were identified because of the prevalence of semantic information over the mention's structure (cf. *jedna*$_{synh}$ *z dziewcząt* 'one$_{synh}$ of the girls$_{semh}$').

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 the NKJP annotation [24]. 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 MMAX4CORE, a heavily modified version of MMAX2 [8], which was used for the annotation task of a single text (when it was acquired by the annotator via DistSys). For the sake of simplicity and annotation speed, many options were removed from the original version of the application while some new features were added, as requested by the annotators (for example the possibility of undoing the last change).

The modifications included a superannotation plugin, which allowed 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, and the third column shows the decision of annotator B. In the second row, we can see that annotator A marked the mention *gorzką czekoladę* 'dark chocolate' (plus) while annotator B didn't (minus) since he decided to mark the complete (discontinued) mention *gorzką czekoladę, ... której połykam od 2–10 kostek dziennie* 'dark chocolate ... 2–10 squares of which I gulp down every day'. With such information, adjudicator needed to double-click the plus or minus depending on the version he agrees with to resolve to difference (B's decision in this case, according to the broad understanding of mention borders as clarified in the annotation guidelines).

Fig. 2 presents a similar interface for adjudicating cluster contents. Mentions from each cluster are assigned the same cluster number and colour. For example,

two occurrences of mention *gorzką czekoladę* 'dark chocolate' are in the same cluster according to annotation A, and are singletons according to annotation B. A single click on any of these decisions displays their textual context in the main application window while a double click selects the clicked version as the adjudicated one and updates the remaining clustering to match it.

Both DistSys and MMAX4CORE are available for free download at the `http://zil.ipipan.waw.pl/PolishCoreferenceTools` web page.

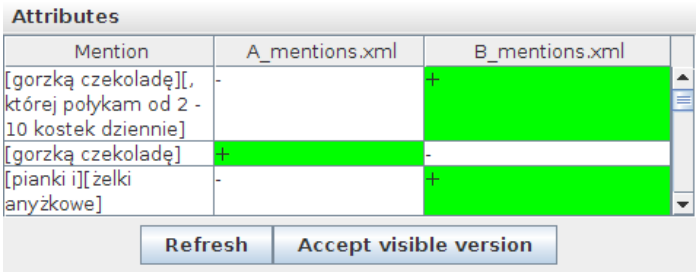

Fig. 1. MMAX4CORE superannotation window — mention adjudication

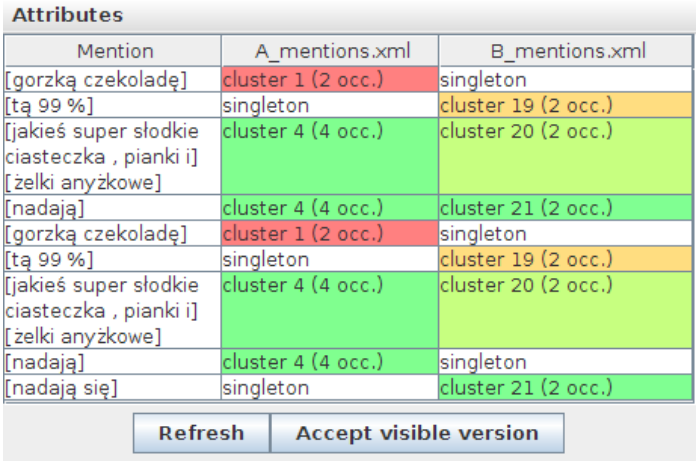

Fig. 2. MMAX4CORE superannotation window — cluster adjudication

4 Corpus availability

Polish Coreference Corpus is freely available for download under the Creative Commons Attribution 3.0 Unported License at: `http://zil.ipipan.waw.pl/`

PolishCoreferenceCorpus. There are 3 download formats, described briefly below. PCC is also available for browsing online (see Fig. 3) 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 [23] is an online collaborative annotation environment 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 quasi-identity).

4.2 MMAX

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

- a file with the .mmax extension, storing the text source (named with the original NKJP text identifier) 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 quasi-identity relations between them.

4.3 TEI

The 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 and header.xml each text in the corpus also has two additional files: ann_mentions.xml and 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. 4 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 <fname="semh"> in the feature structure <fsname="mention">).

The second file provides the coreference and quasi-identity cluster information as groups of mentions from the former file. Fig. 5 presents the encoding of two relations: coreference identity cluster (containing mention_8 and mention_14) and quasi-identity relation (between mention_30 and mention_5). In the case of identity, the encoding also contains the information about the dominant expression (<f> element with "dominant" name attribute).

Fig. 3. Online corpus visualisation

```
<!-- 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>
```

Fig. 4. Mention encoding in `ann_mentions.xml`

```
<!-- 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="quasi-ident"/>
  </fs>
  <ptr target="ann_mentions.xml#mention_5"/>
  <ptr type="source" target="ann_mentions.xml#mention_30"/>
</seg>
```

Fig. 5. Identity and quasi-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. By referring to concepts of quasi-identity, dominant expressions and semantic approach to identity-of-reference it may contribute to a high-quality methodology for constructing similar corpora, particularly for other richly inflected languages. Our ongoing work based on corpus data and tools includes experiments with improvement of extractive summarization algorithms by incorporating coreference information into sentence selection procedure and using mention detectors and coreference resolvers in a linguistic chaining environment (see [15]).

Since our current efforts were limited to direct nominal coreference, an obvious direction for further improvements of the corpus is its extension 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.). Another underexplored topic seems the notion of a dominant expression. We believe that dominant expressions could facilitate cross-document annotation, as well as facilitate the creation of a semantic framework covering different expressions/descriptions of the same object.

As far as coreference-related tools and resources are concerned, their results are much dependent on further development of lower-level tools for Polish such as morphosyntactic analysers (still skipping certain abbreviations, diminu-

tives, slang words etc.), or taggers, directly influencing further processing. New data extraction sources for interpretation of periphrastic (e.g. *Kraj Wschodzącego Słońca = Japonia* 'Land of the Rising Sun = Japan') or knowledge-based expressions (e.g. *mąż Celiny Szymanowskiej = Adam Mickiewicz* 'the husband of Celina Szymanowska = Adam Mickiewicz') seem also urgently needed.

References

1. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) Advances in Natural Language Processing. Lecture Notes in Computer Science, vol. 6233, pp. 3–14. Springer (2010)
2. Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., Wardyński, A.: KPWr: Towards a Free Corpus of Polish. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012. pp. 3218–3222. ELRA, ELRA, Istanbul, Turkey (2012)
3. Linguistic Data Consortium: ACE (Automatic Content Extraction) Spanish Annotation Guidelines for Entities (2006), `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/spanish-entities-guidelines-v1.6.pdf`, accessed on Aug. 28, 2015
4. Hendrickx, I., Bouma, G., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Van, J., Vloet, D., Verschelde, J.L.: 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. European Language Resources Association (ELRA), Marrakech, Morocco (2008)
5. Hinrichs, E.W., Kübler, S., Naumann, K.: 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 (2005)
6. Iida, R., Komachi, M., Inui, K., Matsumoto, Y.: Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In: Proceedings of the Linguistic Annotation Workshop (LAW 2007). pp. 132–139. Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
7. Korzen, I., Buch-Kromann, M.: Anaphoric relations in the Copenhagen Dependency Treebanks. In: Proceedings of DGfS Workshop. pp. 83–98. Göttingen, Germany (2011)
8. Müller, C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (eds.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197–214. Peter Lang, Frankfurt a.M., Germany (2006)
9. Muzerelle, J., Lefeuvre, A., Antoine, J.Y., Schang, E., Maurel, D., Villaneau, J., Eshkol, I.: ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In: Proceedings of the 20th Conference Traitement Automatique des Langues Naturelles (TALN 2013). pp. 555–563. Les Sables d'Olonne, France (2013)
10. Nedoluzhko, A., Mírovský, J., Ocelák, R., 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). pp. 1–16. AU-KBC Research Centre, Anna University, Chennai, AU-KBC Research Centre, Anna University, Chennai, Goa, India (2009)
11. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Interesting linguistic features in coreference annotation of an inflectional language. In: et al., M.S. (ed.) CCL and NLP-NABD 2013, Lecture Notes in Computer Science, vol. 8202, pp. 97–108. Springer-Verlag, Berlin, Heidelberg (2013)
12. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Coreference in Polish: Annotation, Resolution and Evaluation. Walter De Gruyter (2015), http://www.degruyter.com/view/product/428667, accessed on Aug. 28, 2015
13. Ogrodniczuk, M., Kopeć, M., Savary, A.: Polish Coreference Corpus in Numbers. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). pp. 3234–3238. European Language Resources Association, Reykjavík, Iceland (2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/1088_Paper.pdf, accessed on Aug. 28, 2015
14. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for Polish. In: Vetulani, Z. (ed.) Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. pp. 167–171. Poznań, Poland (2011)
15. Ogrodniczuk, M., Lenart, M.: Web Service integration platform for Polish linguistic resources. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012. pp. 1164–1168. ELRA, Istanbul, Turkey (2012)
16. Osenova, P., Simov, K.: BTB-TR05: BulTreeBank Stylebook. BulTreeBank Version 1.0. Tech. Rep. BTB-TR05, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia, Bulgaria (2004)
17. Poesio, M., Artstein, R.: Anaphoric Annotation in the ARRAU Corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). ELRA, European Language Resources Association, Marrakech, Morocco (2008)
18. Pradhan, S.S., Ramshaw, L., Weischedel, R., MacBride, J., Micciulla, L.: Unrestricted coreference: Identifying entities and events in ontonotes. In: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007). pp. 446–453. IEEE Computer Society, Washington, DC, USA (2007)
19. Presspublica: Rzeczpospolita corpus (2013), http://www.cs.put.poznan.pl/dweiss/rzeczpospolita, accessed on Aug. 28, 2015
20. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.): Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. Wydawnictwo Naukowe PWN, Warsaw (2012), http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf, accessed on Aug. 28, 2015
21. Recasens, M., Hovy, E., Martí, M.A.: Identity, non-identity, and near-identity: Addressing the complexity of coreference. Lingua 121(6), 1138–1152 (2011)
22. Recasens, M., Martí, M.A.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation 44(4), 315–345 (2010)
23. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: 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. pp. 102–107. EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)

24. Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A., Lenart, M.: Annotation tools for syntax and named entities in the National Corpus of Polish. International Journal of Data Mining, Modelling and Management 5(2), 103–122 (2013)
25. Woliński, M.: 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 (Jun 2006)