Polish Coreference Corpus in Numbers

Maciej Ogrodniczuk[1], **Mateusz Kopeć**[1], **Agata Savary**[2]
[1]Institute of Computer Science, Polish Academy of Sciences,
[2]François Rabelais University Tours, Laboratoire d'informatique
[1]Jana Kazimierza 5, 01-248 Warsaw, Poland
[2]3 place Jean-Jaurès, 41029 Blois, France
m.ogrodniczuk@ipipan.waw.pl, m.kopec@ipipan.waw.pl, agata.savary@univ-tours.fr

Abstract

This paper attempts a preliminary interpretation of the occurrence of different types of linguistic constructs in the manually-annotated Polish Coreference Corpus by providing analyses of various statistical properties related to mentions, clusters and near-identity links. Among others, frequency of mentions, zero subjects and singleton clusters is presented, as well as the average mention and cluster size. We also show that some coreference clustering constraints, such as gender or number agreement, are frequently not valid in case of Polish. The need for lemmatization for automatic coreference resolution is supported by an empirical study. Correlation between cluster and mention count within a text is investigated, with short characteristics of outlier cases. We also examine this correlation in each of the 14 text domains present in the corpus and show that none of them has abnormal frequency of outlier texts regarding the cluster/mention ratio. Finally, we report on our negative experiences concerning the annotation of the near-identity relation. In the conclusion we put forward some guidelines for the future research in the area.

Keywords: coreference, corpus, Polish, annotation

1. Introduction

The analysis of the statistical properties of linguistic constructs in the Polish Coreference Corpus (Ogrodniczuk et al., 2013a) presented in this paper aims at two targets. One of them is the assessment of how the frequency of certain linguistic phenomena compares to efforts put in their processing and successes achieved. As some of them are not yet considered in automatic processing of Polish coreference, proving their high frequency can bring new motivation to the development of their resolution techniques. Other important questions we would like to answer are about the inter-annotator agreement of different linguistic phenomena and current efficiency of automatic methods to discover them to estimate whether their careful annotation brings real value to the task and whether the task is sufficiently clear.

The core[1] of the Polish Coreference Corpus (PCC, see (Ogrodniczuk et al., 2013b)) is 1,773 texts, 250-350 tokens each, constituting fragments of longer documents (but always full consecutive paragraphs) randomly selected from the National Corpus of Polish (NKJP) (Przepiórkowski et al., 2012). The texts have been manually annotated with general nominal coreference. Average counts of corpus building blocks are presented in Table 1.

Indicator	Average count
Paragraphs per text	7.38
Sentences per paragraph	2.38
Tokens per sentence	16.19

Table 1: Average counts of units in PCC

[1]PCC contains also 21 "long" documents, for data homogeneity omitted in this study.

2. Mentions

The total number of mentions identified in PCC is 167,871, which amounts to 5.39 mentions per sentence on average. This number seems high, taking into account the 16-token sentences, but can be explained with our annotation strategy: mentions are nominal groups (NGs) which can be nested based on different potential referents (as in *the CEO of Microsoft*). These nominal groups include pronouns, named entities and zero subjects (annotated at verbs missing explicit subjects).

Mention boundaries in PCC were maximized by including e.g. adjectives and adjectival participles in agreement with superior noun, subordinate nouns in the genitive, subordinate prepositional-nominal phrases and relative clauses. The reasoning behind this extension was to provide precise reference (cf. 'the astronaut' vs. 'the astronaut who stayed in the command module while Armstrong and Aldrin walked on the Moon'). This assumption could result in extended mention size — and it partially did, creating mentions as long as 147 segments. Nevertheless 90% of the mentions are of size 5 and below. This can be explained with the fact that even potentially lengthy definitions are constrained by text authors to maintain their understandability and 5 tokens seem enough to convey the complete nature of the mention. The average mention size is 2.66 (for singleton mentions: 3.19, for non-singletons: 1.85); see Table 2 for reference.

3. Coreference Clusters

In PCC we mark identity of reference in its strict form (direct reference) with an extension to the so called near-identity (explained later). Direct reference, which is an equivalence relation, clusters group mentions referring to the same discourse-world entity. Table 3 shows basic cluster statistics. Non surprisingly, the large majority (85%) of

	Count	Avg. count per text	Avg. count per paragraph	Avg. count per sentence
All clusters	119,848	67.60	9.16	3.85
Singleton clusters	102,218	57.70	7.82	3.29
Non-singleton clusters (ns-clusters)	17,630	9.94	1.35	0.57
Single paragraph ns-clusters	8,364	4.72	0.64	0.27

Table 3: Basic cluster statistics

Length in tokens	Count	% of all mentions	% singleton
1	83405	49.68	44.68
2	36976	22.03	72.16
3	15966	9.51	78.08
4	8805	5.25	81.18
5	5887	3.51	82.33
6	3795	2.26	81.79
7	2682	1.60	83.71
8	2067	1.23	83.02
9	1570	0.94	81.08
10	1234	0.74	81.85
11-20	4,466	2.66	81.61
21-30	724	0.43	79.14
31-40	176	0.10	82.39
41-50	57	0.00	84.74
51-98	57	0.00	85.96
103-147	4	0.00	100.00
1-147	167,871	100.00	60.86

Table 2: Mention length

all 119,848 clusters corresponds to singleton clusters (containing one mention only). The remaining 17,630 non-singleton clusters (referred to as ns-clusters) are mostly composed of 2 (54%), 3 (18%) or 4 (8%) mentions. The longest cluster, occurring in a spoken dialogue, contains 41 mentions (mostly *ja* 'I' pronouns and 1st person zero subjects). The total average cluster size is 1.40 mention, while for non-singleton clusters only it is 3.72 mentions. Less than the half (8,364) of the non-singleton clusters are included in single paragraphs. This validates our choice of text fragments for the PCC construction, which are larger than single paragraphs (unlike the manually annotated 1-million word NKJP subcorpus).

3.1. Pronouns and Zero Subjects

Two particularly interesting types of mentions are pronouns and zero subjects. The former are prototypical anaphoric techniques while the latter are marked at finite verb tokens in case of non-subject sentences. As shown in Table 4 they account for about 14% of all mentions. Most (over 92%) mentions of these types appear in non-singleton clusters which is often due to the presence of a more specific mention in the same cluster which introduces the referent. Notable exceptions include impersonal use of the second person plural pronoun *my, nas*, etc. 'we' (see Example 1), improper verbs[2] (Examples 1–2), impersonal use of second and third person plural verbs (Examples 3–4), and verbs contained in titles of works (Example 5).

(1) *należy [. . .] dać nową szanse każdorazowo, gdy nas o nią proszą*
'(lit.) should$_3 pers.sing$ give a new chance each time we are asked for it' = 'one should [. . .] give a new chance each time one is asked for it'

(2) *Ale jeśli chodzi o Halinę . . .*
'(lit.) But if goes about Halina . . . = As far as Halina is concerned . . . '

(3) *Jest to wszystko zrozumiałe, gdy weźmiemy pod uwagę, że . . .*
'All that is understandable if (we) take$_2 pers.pl$ into account that . . . '

(4) *Już pana wyleli?*
'(lit.) Already fired$_3 pers.pl$ you ? = Have they fired you already?'

(5) *druga kompozycja Czesława – "Czy mnie jeszcze pamiętasz?"*
'another song by Czesław – (lit.) "Still remember$_{2 pers.sing}$ me?" = "Do you still remember me?"'

Note also that over 30% of all verbal forms in the corpus are marked as mentions, which confirms the importance of the zero anaphora phenomenon in Polish.

	Count	% singleton
Verbs	50,134	–
Pronouns	8,794	–
Mentions	167,871	60.89
Verb mentions	15,398	7.61
Pronoun mentions	7,547	7.38

Table 4: Singleton vs. non-singleton mentions

Table 5 presents sizes of clusters. The average cluster size is 1.40, for nonsingleton clusters only it is 3.72.

3.2. Agreement in Clusters

Number and gender agreement between mentions, as well as identity of headwords, are among the features frequently

[2]Improper verbs in Polish are verbs occurring only in the 3rd person singular.

Cluster size in mentions	Count	% of all clusters	% of all mentions in clusters of that size
1	102,218	85.29	60.89
2	9,446	7.88	11.25
3	3,229	2.69	5.77
4	1,498	1.25	3.57
5	918	0.77	2.73
6	572	0.48	2.04
7	395	0.33	1.65
8	335	0.28	1.60
9	205	0.17	1.10
10	174	0.15	1.04
11-20	547	0.45	4.93
21-30	140	0.12	1.98
31-41	24	0.01	0.47
Any	119,848	100.00	100.00

Table 5: Cluster size

Ns-cluster type	Count	% ns-clusters
Same head number	14,088	79.91
Same head gender-relaxed	11,016	62.48
Same head gender	10,286	58.34
Same head base	7,048	39.98
Same head orth	3,707	21.03
Any	17,630	100.00

Table 6: Cluster agreement types

Mention type	Cluster count	% of all ns-clusters
Indefinite pronoun	44	0.25
Negative pronoun	11	0.06
Universal pronoun	46	0.26
Any pronoun	4,183	23.73
Any	17,630	100.00

Table 7: Ns-clusters with at least one mention of specific type

taken into account in automatic, both rule-based and probabilistic, coreference resolution tools. Table 6 presents the agreement statistics for non-singleton clusters in PCC. The agreement counts in the table represent the situations when all mentions in a cluster have the same value of the given parameter assigned by the PANTERA tagger (Acedański, 2010). Relaxed gender stands for the three Polish agglomerated masculine genders: masculine human, masculine animate and masculine inanimate. Clusters containing at least one mention with non agreeing gender are very frequent (41% of all clusters, 37% in case of relaxed gender agreement). This fact can be justified e.g. by the use of synonyms or hyper-/hyponyms to describe the same referent (e.g. *to Volvo* 'this Volvo$_{neut.sing}$', *ten samochód* 'this car$_{masc.sing}$'). The frequent disagreement in number (20% of clusters) is more surprising. Interesting cases of this type concern generic noun groups as in Example 6.

(6) *wyleczenie z antysemityzmu* 'curation of antisemitism'$_{neut.sing}$
takie zmiany 'such changes'$_{femin.plur}$

These statistics show that using traditional strict gender/number agreement constrains for coreference clustering in an automated tool is not a good solution in the case of Polish.
Note also that there are only about 40% of clusters with the same head base form in all mentions and almost half of them (21%) contain graphically different inflected forms of the head. This fact confirms the necessity of using lemmatization or stemming techniques for automatic coreference resolution in highly inflected languages, notably those admitting declension (inflection for case of nouns, adjectives and pronouns).

3.3. Clusters with Indefinite Mentions

Traditionally indefinite pronouns (*ktokolwiek* 'anyone', *ktoś* 'someone', *cokolwiek* 'anything', *coś* 'something'), negative pronouns (*nic* 'nothing', *nikt* 'nobody') or universal pronouns (*wszystko* 'everything', *wszyscy* 'everybody') are not regarded as coreferential since they do not carry direct reference information. However, analysis of the corpus showed that they can frequently form coreferential chains, as in Example 7.

(7) *Jak ktoś jest zazdrosny, znaczy, że naprawdę kocha.*
'If someone is jealous, it means, that $_{h}e/she$ really loves.'

Table 7 presents statistics of non-singleton clusters containing at least one pronoun of these types.

4. Cluster and Mention Count Correlation

The distribution of cluster and mention count in texts is depicted in Figure 1. Each dot corresponds to one text from

the corpus; the dot numbers indicate text identifiers to enable tracking the outliers. The horizontal axis shows the number of mentions in a given text, divided by the length of text in tokens for normalization. The vertical axis shows the number of clusters in a given text, again divided by the number of tokens.

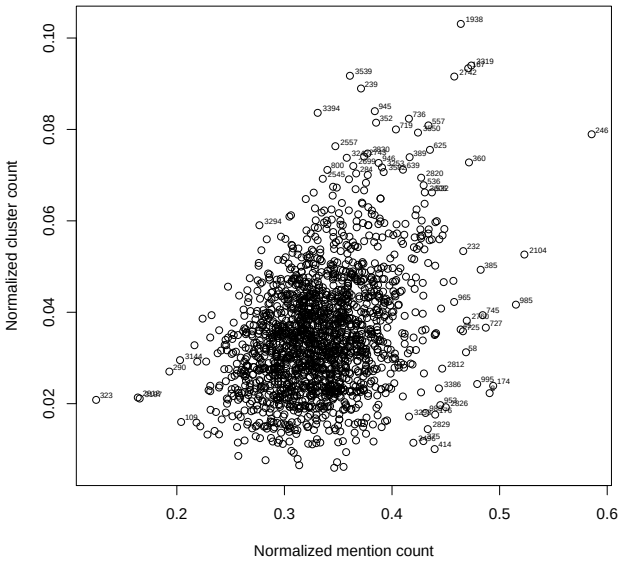

Figure 1: Normalized cluster/mention ratio

Figure 1 shows that both normalized mention count and normalized cluster count are similar for most texts, with certain outliers:

- text 246 (the rightmost dot) is a training programme with extended number of mentions due to the nominal phrase-based character of such document types (with training places, instructor names and topics being all nominal phrases),
- text 1938 (the topmost dot) is a quasi-spoken parliament session transcript with an extensive list of clusters resulting from the discussed topics (multiple references to parliamentary committees, voting procedure, bills etc.),
- text 323 (the leftmost dot) is a short and highly fragmented relation from a book promotional event, with a book excerpt, book title, information about the event time and place; it features low number of clusters and hardly any nested mentions.

We have also calculated the Mahalanobis distance from the mean point for this data to sort the texts from the most "atypical" regarding the mention/cluster proportion (number 0) to the most "typical" one (number 1772). Then, we have drawn a boxplot presented in Figure 2, which shows the positions of texts from different domains on that list. We have 14 different text domains in the corpus and there is no strong evidence that any of them is having abnormal frequency of "atypical" texts regarding the cluster/mention ratio. Text type numbers are explained in table 8.

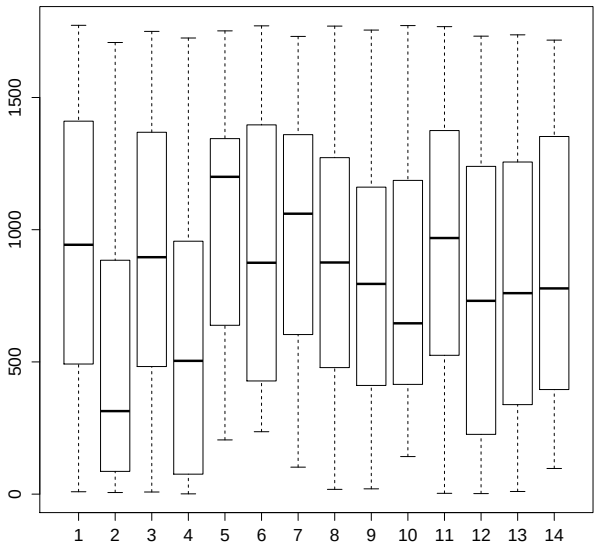

Figure 2: "Typicality" of texts from different domains

Id	Text type
1	Dailies
2	Misc. written (legal, ads, manuals, letters)
3	Internet interactive (blogs, forums, usenet)
4	Internet non-interactive (static pages, Wikipedia)
5	Unclassified written
6	Journalistic books
7	Non-fiction literature
8	Fiction literature (prose, poetry, drama)
9	Spoken – conversational
10	Spoken from the media
11	Magazines
12	Quasi-spoken (parliamentary transcripts)
13	Instructive writing and textbooks
14	Academic writing and textbooks

Table 8: Identifiers of text types in Figure 2

5. Near Identity

Near-identity is a novel coreference relation defined in (Recasens et al., 2011) and further extended in (Recasens et al., 2010). It is based on observation that in certain cases clear distinction between identity and non-identity is difficult. Two popular examples of near-identity are achieved with refocusing (focusing on a certain property of an object, thus quasi-splitting it into several facets, e.g. "a child" vs. "an adult" while speaking of different aspects of the same person) and neutralization (merging two distinct objects into a 'meta-object', as in reading "a book" and watching "a movie" while referencing the same content).

The total number of such quasi-identity links in the annotated corpus was relatively high, amounting to 4,699 (2.65 per text) while the value obtained for inter-annotator agreement, in terms of Cohen's κ, is extremely low — only 0.222.

Some annotators never even used the quasi-identity link.

On the other hand annotators obviously had no problem with deciding if there is coreference between phrases — there are only 73 cases where one annotator linked phrases with the identity link and another — with the quasi-identity link. Many times the differences arose due to insufficient knowledge of an annotator, for example an annotator did not link the phrases *All Saint's Day* and *the 1st of November* with the identity link although they refer to the same day.

These results bring strong doubts not only about the utility of the mentioned typology but also of the near-identity as such. Near-identity links seem rather hard to establish and no repeatable pattern in the near-identity annotation has occurred. Therefore in our opinion the utility of the near-identity concept for coreference annotation is questionable.

6. Conclusion and Further Work

We have shown several observations resulting from statistical properties of the Polish Coreference Corpus. Some of them are currently being further investigated in related research tasks such as measuring text readability based on the number and character of coreferential links (negative correlation is shown between e.g. the average size of a coreference cluster in a document and traditional readability measures, such as the Gunning FOG index).

One of the next steps will be a comparison of "NG coreference density" between Polish and other languages, based on data extracted from existing coreference-enabled corpora such as the Tübingen treebank of German newspaper texts (TüBa-D/Z) annotated with a set of coreference relations (Hinrichs et al., 2005), a coreference-annotated corpus of Dutch newspaper texts, transcribed speech and medical encyclopaedia entries (Hendrickx et al., 2008), NAIST, a Japanese corpus annotated for coreference and predicate-argument relations (Iida et al., 2007), AnCora-CO, coreferentially annotated corpora for Spanish and Catalan (Recasens and Martí, 2010) and many others. It seems that there exists no systematic evaluation of statistical properties of such corpora going beyond a simple mention and cluster count. This results probably from the differences in annotation guidelines and approaches to certain coreference-related linguistic properties such as appositions, predicates or relative clauses, hindering unifications and comparisons.

Acknowledgements

The work reported here was cofunded by the "Computer-based methods for coreference resolution in Polish texts", a project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40) and by the European Union from the resources of the European Social Fund (PO KL project: "Information technologies: Research and their interdisciplinary applications").

7. References

Acedański, S. (2010). A Morphosyntactic Brill Tagger for Inflectional Languages. In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S., editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer.

Hendrickx, I., Bouma, G., Daelemans, W., Hoste, V., Kloosterman, G., marie Mineur, A., 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)*, pages 144–149, Marrakech, Morocco. European Language Resources Association (ELRA).

Hinrichs, E. W., Kübler, S., and Naumann, K. (2005). 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*, pages 13–20, Ann Arbor, Michigan, USA.

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

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (2013a). Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In et al., M. S., editor, *CCL and NLP-NABD 2013*, volume 8202 of *Lecture Notes in Computer Science*, pages 97–108. Springer-Verlag, Berlin, Heidelberg.

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (2013b). Polish Coreference Corpus. In Vetulani, Z., editor, *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 494–498, Poznań, Poland. Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

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

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

Recasens, M., Hovy, E., and Marti, M. A. (2010). A Typology of Near-Identity Relations for Coreference (NIDENT). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 149–156, Valletta, Malta. European Language Resources Association.

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