Inter-Annotator Agreement in Coreference Annotation of Polish*

Mateusz Kopeć and Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences

Abstract. This paper discusses different methods of estimating the inter-annotator agreement in manual annotation of Polish coreference and proposes a new BLANC-based annotation agreement metric. The commonly used agreement indicators are calculated for mention detection, semantic head annotation, near-identity markup and coreference resolution.

1 Introduction

A substantial annotation of Polish coreference has been recently carried out in the course of creation of the Polish Coreference Corpus (PCC) — a large manually annotated resource of general Polish coreference. The annotation procedure consisted of marking up the following entities:

- mentions — all nominal groups constituting reference to discourse-world objects
- mention semantic heads — the most relevant word of the group in terms of meaning; typically equal to syntactic head, but different for numerals or elective expressions (cf. one_{synh} *of the* $girls_{semh}$)
- identity clusters — groups of mentions having the same referent
- near-identity links — associations between a pair of semi-identical mentions, carrying some of their properties (cf. *prewar Warsaw* and *Warsaw today*[1])
- dominating expressions — a mention in a cluster which carries the richest semantics or describes the referent with most precision.

For each of the above-mentioned subtasks inter-annotator agreement can be evaluated to show difficulty of each assignments individually. In this paper we present conclusions stemming from the investigation of 210 texts (60,674 segments) from 14 domains (15 texts per domain).

* 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 was also co-funded by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications".

[1] See [1] for general introduction to the concept and section 4 of [2] for details of Polish annotation of near-identity in PCC.

2 The Annotation Process

The texts for annotation, 250-350 token each, had been randomly selected from the National Corpus of Polish [3]. The samples were automatically processed to detect mention candidates with a newly implemented software based on existing language processing tools for Polish: Spejd shallow parser [4], Pantera tagger and Nerf named entity recognizer [5]. Baseline coreference resolution tool Ruler [6] was used for initial mention clustering.

Each pre-processed text from the sample selected for the experiment has been annotated independently by two annotators (hence: A and B) from the team of eight linguists co-operating with the project. The annotation was performed in a customized MMAX2 tool [7]. The annotators were instructed (with detailed guidelines[2]) to correct the pre-annotation results by removing existing markup, changing it or adding new entities or associations. As a result of the process, 420 annotated texts were produced with a total of 41,006 mentions, 4,410 clusters and 1,009 near-identity links.

It is worth pointing out that the quantity of the annotated content is relatively large and surpasses the previous attempts of evaluation of inter-annotator agreement in coreference annotation, while the number of annotators per text is minimal. For example, the authors of AnCora-CO-Es corpus [9] evaluated agreement of 8 annotators processing only 2 texts (approx. 1100 segments altogether).

3 Mentions

According to [10], estimation of the inter-annotator agreement including the chance-based factor for the task of marking up mentions (which can be nested, discontinuous and overlapping) has not been yet investigated. Since it is difficult to estimate the probability of a random markup of a mention, we present the observed agreement only.

In our sample of 210 texts the annotation of the annotator A contained 20,420 mentions while for annotator B — 20,560 mentions. 17,530 mentions were shared which means they had exactly the same borders (including the inner borders for discontinuous mentions). Regarding annotation A as gold and B as system, precision is 85.26%, recall = 85.85% which gives $F_1 = 85.55\%$.

When comparing mention heads only, the annotation A contained 19,394 mentions (after excluding mentions having the same heads), annotation B — 19,522 mentions; 18,317 mentions were shared. For this setting precision = 93.83%, recall = 94.47%, $F_1 = 94.14\%$.

4 Semantic Heads

Agreement in annotation of mention heads can only be investigated for shared mentions. For 17,363 mentions out of 17,530 the same heads were marked, which gives the observed agreement: $p_{A_O} \approx 0.9905$.

[2] For the PCC annotation schema and strategies see [8].

The chance agreement (p_{A_E}) was calculated in the following way. For each mention its head was selected by pointing out one segment from all mention tokens. We assumed that the probability of choosing the same head by chance for given mention was equal to the inverse of its token count. Chance agreement was therefore calculated as an average of chance agreement probabilities for individual mentions and yielded: $p_{A_E} \approx 0.6832$. This value is high due to a high count of one-token mentions, having the chance agreement equal to 1.

Having computed both the above-mentioned values the S inter-agreement measure [11] (as the chance probability distribution is uniform) could be calculated, yielding satisfactory result:

$$S = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} \approx 0.9700$$

5 Near-Identity

As in the case of semantic heads, the agreement of near-identity linking was investigated only for mentions present in both annotations.

For each mention pair in a text the annotator could decide on their linking. Combined agreement results for each mention pair are presented in Table 1.

		Annotation B	
		Near-identical	**Non-near-identical**
Annotation A	**Near-identical**	67	306
	Non-near-identical	367	741,584

Table 1. Near-identity link agreement in all texts

For this table the Cohen κ [12] could be calculated, but this approach would be wrong since annotators cannot add a near-identity link between mentions in two different texts. Therefore we calculated κ for each text separately and then averaged it. This time we assumed per-text-and-annotator probability distribution (details about κ calculation are presented in section 7.1).

When text did not contain any links, agreement value of 1 was assumed. When one annotator did not mark any link while the second one did, the agreement was assumed as 0 (with predicted value equal to the observed one).

Applying this procedure to all texts we have calculated the average $\kappa \approx 0.2220$. The result is low which can be interpreted as a difficulty in linking mentions with near-identical relation. The notion seems vague — in 128 cases mention pairs were marked as near-identical by one annotator and at the same time as purely-identical (i.e. were clustered) by the other annotator.

6 Dominating Expressions

As with previous cases, this type of agreement concerns only mentions annotated by both annotators.

Dominating expressions were marked for non-singleton clusters only. The number of common mentions with a marked dominating expression was 6,162, with 4,115 mentions sharing the same dominating expression ($\approx 66.78\%$). When only one representative of each cluster is investigated (which makes sense since each element of a cluster has the same dominating expression) 1,146 out of 1,818 cluster representatives has the same dominating expression in both annotations ($\approx 63.04\%$).

Chance agreement analysis is not carried out since apart from choosing a cluster element as the dominating expression the annotators could also enter arbitrary text value, which makes good chance agreement estimations impossible.

7 Coreference

According to [10], coreference resolution is a very specific task, dealing with clustering and not classification, atypically for the whole field of computational linguistics. Moreover, selection of the best evaluation method for the new resource is difficult since there is no consensus in the scientific world about 'the best' metric. In this extensive section we analyse the most popular ones to present their properties and results for our annotation.

7.1 Cohen's κ

In [13] Passonneau describes two inter-annotator agreement metrics: Cohen's κ [12] and original Krippendorff α [14].

$$\kappa = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} \qquad \alpha = 1 - \frac{p_{D_O}}{p_{D_E}}$$

κ is defined as the difference of observed agreement (p_{A_O}) and chance agreement (p_{A_E}) while α involves the observed non-agreement (p_{D_O}) and chance non-agreement (p_{D_E}). Passonneau shows that $\alpha = \kappa$, so in the further part of the text we concentrate on the procedure of calculating κ.

		Annotator A		
		Label X	**Label Y**	Σ
Annotator B	**Label X**	47	14	61
	Label Y	10	29	39
	Σ	57	43	100

Table 2. A sample coincidence matrix

α and κ can be calculated when coincidence matrix is available for multiple annotators' decisions (with data how frequent each of the annotators were choosing each label). For example, when it is defined as in Table 2, we can calculate:

$$p_{A_O} = \frac{47 + 29}{100} = 0.76$$

$$p_{A_E} = \frac{57}{100} * \frac{61}{100} + \frac{43}{100} * \frac{39}{100} = 0.5154$$

Observed agreement shows diagonally in the matrix. Expected agreement is based on the probability of selection of each label by each annotator. For instance, when annotator A selected label X in 57% decisions and annotator B in 61% decisions, the chance that they accidentally chose the same label A is $(57/100) * (61/100)$. The sum of probabilities for all labels gives the expected agreement. Finally for Table 2 we have:

$$\kappa = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} = 0.5$$

For the coreference annotation agreement assessment, crucial decision is to choose how to represent coreference annotation in coincidence matrix similar to Table 2. We present some approaches in the following sections. For the reason described in the near-identity section, we suggest calculating the agreement for each text separately and then averaging it.

7.2 MUC-based metrics

In [13] the coincidence table for agreement is calculated similarly to MUC metrics. The matrix similar to Table 7.2 is created, where $Link+$ denotes annotation of association between mentions (minimal) and $Link-$ — no association. The details of calculation of MUC metrics can be found in [15].

Unfortunately, because of certain properties of MUC metrics (e.g. not taking singletons into account) it was not widely accepted standard and is usually used as a supporting metrics only.

		Annotator A		
		$Link+$	$Link-$	Σ
Annotator B	$Link+$	47	14	61
	$Link-$	10	29	39
	Σ	57	43	100

Table 3. Coincidence matrix for MUC metrics

7.3 Weighted Krippendorff α

Passonneau in [16] presented a different approach, making use of the weighted version of Krippendorff α. To calculate it, annotator's decision for a given mention should be understood as assignment of a set of mentions from the same cluster (apart from this mention).

For instance, if for five mentions with labels $1, 2, 3, 4, 5$ annotator A created clusters: $\{1, 3\}, \{2, 4, 5\}$, and annotator B clusters: $\{1\}, \{2, 4\}, \{3, 5\}$, their annotation will be represented as in Table 4. It shows that e.g. according to the annotator A the mention 1 is clustered only with 3 and it is a singleton according to the annotator B.

Mention number	1	2	3	4	5
Annotator A	{3}	{4, 5}	{1}	{2, 5}	{2, 4}
Annotator B	{}	{4}	{5}	{2}	{3}

Table 4. Representation of a sample cluster annotation

Passonneau's idea was to punish differences between annotators subject to the degree of difference between the clusters assigned to the same mention. To apply this technique, Krippendorff α was used with weights assigned to each error, according to the selected distance metrics. Let's define distance between e_1 and e_2 as $\delta(e_1, e_2)$. For the first mention in our example the weight of an error is $\delta(\{3\}, \{\})$ — and analogically for all other mentions. Let's mark the set of all clusters used by the annotators as E (in our example it contains 9 elements which is the number of rows and columns in the coincidence matrix).

The α equation for two annotators is the following:

$$p_{D_O} = \frac{1}{n} \sum_{e_1 \in E} \sum_{e_2 \in E} o_{e_1 e_2} \delta(e_1, e_2)$$

$$p_{D_E} = \frac{1}{n(n-1)} \sum_{e_1 \in E} \sum_{e_2 \in E} n_{e_1} n_{e_2} \delta(e_1, e_2)$$

$$\alpha = 1 - \frac{p_{D_O}}{p_{D_E}} = 1 - (n-1) \frac{\sum_{e_1, e_2 \in E} o_{e_1 e_2} \delta(e_1, e_2)}{\sum_{e_1, e_2 \in E} n_{e_1} n_{e_2} \delta(e_1, e_2)}$$

where $o_{e_1 e_2}$ is he number of mentions assigned by one of the annotators to e_1 cluster, and by the second one — to e_2 cluster, n_{e_1} is the number of all assignment of e_1 label and analogically, n_{e_2} is the number of assignment of e_2 label. In our example $o_{\{3\}\{2,4\}} = 1$, $o_{\{3\}\{2,5\}} = 0$, and $n_{\{3\}} = 2$, $n_{\{2,5\}} = 1$.

Passoneau in [16] defines $\delta(e_1, e_2)$ function in the following manner (with the result calculated with first matching rule counting from the top):

- $\delta(e_1, e_2) = 0$, when $e_1 = e_2$

– $\delta(e_1, e_2) = 0.33$, when $e_1 \subset e_2 \vee e_2 \subset e_1$,
– $\delta(e_1, e_2) = 0.67$, when $e_1 \cap e_2 \neq \emptyset$,
– $\delta(e_1, e_2) = 1$, when $e_1 \cap e_2 = \emptyset$.

In [17] she proposes another variant of this metric for the same task — MASI (Measuring Agreement on Set-valued Items). MASI is calculated as the product of the previous δ metrics and Jaccard coefficient [18]:

$$MASI(e_1, e_2) = \delta(e_1, e_2) * \frac{|e_1 \cap e_2|}{|e_1 \cup e_2|}$$

We calculated the agreement for the Polish Coreference Corpus following the procedure described in [16] (weighted Krippendorff α) to achieve 79.08% and 59.54% according to [17] (MASI).

7.4 Recasens Approach

In [9] Recasens describes the study of agreement between 8 annotators for 2 texts from the AnCora-CO-Es corpus (approx. 1,100 tokens in total). Assuming identity of mentions presented to annotators, the work were organised to test two aspects:

1. Annotator agreement concerning assignment of a mention to a cluster of a certain type. For each mention annotators could mark it as
   – non-coreference
   – discourse deixis
   – predicative
   – identity

   which made it a fairly standard classification task, investigated with weighted α (0.85 for the first text, 0.89 for the second one).
2. Annotator agreement concerning clustering of each mention from *predicative* or *identity* categories. Labels were cluster numbers, so it was a classification task again, investigated with κ (0.98 for text one, 1 for text two).

We have calculated the agreement for PCC by investigating, for each mention marked by both annotators, whether the mention is clustered or not. This binary decision is presented in Table 5 and examined with Cohen's κ (this time it can be calculated for the whole corpus at once, without averaging).

According to the data in Table 5 the observed agreement (p_{A_O}) is:

$$p_{A_O} = \frac{6238 + 9094}{6238 + 9094 + 975 + 1223} \approx 0.8746$$

while the predicted agreement (p_{A_E}) is:

$$p_{A_E} = \frac{6238 + 1223}{17530} * \frac{6238 + 975}{17530} + \frac{9094 + 1223}{17530} * \frac{9094 + 975}{17530} \approx 0.5132$$

which makes:

$$\kappa = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}} \approx 0.7424$$

		Annotation B	
		Clustered	**Singleton**
Annotation A	Clustered	6,238	975
	Singleton	1,223	9,094

Table 5. Inter-coder decision agreement on singleton/cluster element for all texts in PCC

7.5 BLANC-type agreement

Statistics of coreferential and non-coreferential links for mentions (as in BLANC metrics) marked by both individual annotations are listed in Table 6.

		Annotation B	
		Coreferential	**Non-coreferential**
Annotation A	**Coreferential**	16,638	3,448
	Non-coreferential	3,353	718,822

Table 6. Agreement of BLANC links in all texts

Cohen's kappa could be calculated for this data, but (again) it would not take into account that annotators cannot (even by chance) cluster mentions coming from two different texts. This means that κ should be calculated for each text separately and then averaged.

Application of such procedure to all texts in PCC and grouping results for different text types is shown in Table 7. The data can be interpreted by taking into account several factors such as:

- discrepancy in the speaker and recipient's conceptual systems, resulting in difficulty in interpretation of academic books by a non-expert annotator,
- higher readability of fiction than academic or spoken texts, boosting the agreement value.

8 Conclusions

We have presented several approaches of calculating the inter-annotator agreement in coreference annotation of Polish and its results for four coreference-related tasks. We have investigated two typical tasks: mention detection and coreference resolution as well as two less common ones: semantic head annotation and near-identity markup.

The results of the analysis confirm the assumption that coreference is more of a semantic and conceptual phenomenon which cannot reach scores as high as

Text type	κ
Academic writing	0.699
Instructive writing and textbooks	0.727
Internet non-interactive (static pages, Wikipedia)	0.730
Dailies	0.740
Quasi-spoken (parliamentary transcripts)	0.746
Internet interactive (blogs, forums, usenet)	0.764
Spoken — conversational	0.765
Other periodical	0.772
Spoken from the media	0.785
Non-fiction	0.795
Unclassified written	0.807
Journalistic books	0.817
Misc. written (legal, ads, manuals, letters)	0.826
Fiction	0.871
Any	0.775

Table 7. κ values for individual text domains

those achieved in lower-level linguistic tasks such as segmentation or morphosyntactic annotation. The average coreference agreement result of 0.775 seems to show the upper limit of coreference resolution capabilities, currently being reached by the state-of-the art tools for Polish (cf. e.g. [6]). Results of near-identity annotation prove the difficulty of its reliable annotation in the current understanding of this phenomenon which should be verified in the further coreference annotation projects.

References

1. Recasens, M., Hovy, E., Martí, M.A.: A Typology of Near-Identity Relations for Coreference (NIDENT). In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 149–156
2. 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. Volume 8202 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg (2013) 97–108
3. 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)
4. Przepiórkowski, A., Buczyński, A.: Spejd: Shallow Parsing and Disambiguation Engine. In Vetulani, Z., ed.: Proceedings of the 3rd Language & Technology Conference, Poznań, Poland (2007) 340–344
5. 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) (2013) 103–122

6. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for Polish. In Vetulani, Z., ed.: Proceedings of the Fifth Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland (2011) 167–171
7. 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. Peter Lang, Frankfurt a.M., Germany (2006) 197–214
8. Ogrodniczuk, M., Zawisławska, M., Głowińska, K., Savary, A.: Coreference Annotation Schema for an Inflectional Language. In Gebulkh, A., ed.: Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics. Part I. Volume 7816 of Lecture Notes in Computer Science. Springer-Verlag, Heidelberg (2013) 394–407
9. Recasens, M.: Coreference: Theory, Annotation, Resolution and Evaluation. PhD thesis, University of Barcelona (2010)
10. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4) (2008) 555–596
11. Bennet, E.M., Alpert, R., Goldstein, A.C.: Communications through limited response questioning. Public Opinion Quarterly **18** (1954) 303–308
12. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1) (April 1960) 37–46
13. Passonneau, R.J.: Applying reliability metrics to co-reference annotation. CoRR **cmp-lg/9706011** (1997)
14. Krippendorff, K.H.: Content Analysis: An Introduction to Its Methodology. 2nd edn. Sage Publications, Inc (December 2003)
15. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th conference on Message understanding. MUC6 '95, Stroudsburg, PA, USA, Association for Computational Linguistics (1995) 45–52
16. Passonneau, R.J.: Computing reliability for coreference annotation. In: LREC, European Language Resources Association (2004)
17. Passonneau, R., Habash, N., Rambow, O.: Inter-annotator agreement on a multilingual semantic annotation task. In: In Proceedings of LREC. (2006)
18. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Sociète Vaudense des Sciences Naturelles **44** (1908) 223–270