Interesting Linguistic Features in Coreference Annotation of an Inflectional Language*

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. This paper reports on linguistic features and decisions that we find vital in the process of annotation and resolution of coreference for highly inflectional languages. The presented results have been collected during preparation of a corpus of general direct nominal coreference of Polish. Starting from the notion of a mention, its borders and potential vs. actual referentiality, we discuss the problem of complete and near-identity, zero subjects and dominant expressions. We also present interesting linguistic cases influencing the coreference resolution such as the difference between semantic and syntactic heads or the phenomenon of coreference chains made of indefinite pronouns.

1 Introduction

For languages still lacking state-of-the-art coreference resolution tools, manual annotation of coreference over a substantially large dataset is traditionally the first step of the work: after the labor-intensive process is over, a supervised resolver can be trained on the hand-annotated documents. Since such resource was until recently unavailable for Polish, all coreference-related work concentrated on theoretical modelling, rule-based or projection-based approaches, and were evaluated on very small data samples.

All the issues above were highly motivating for creation of the first large-scale corpus of general direct nominal coreference of Polish (currently in last phases of construction). In this paper we present the decisions we made while selecting and adopting the annotation schema for this corpus and how they were influenced (and then verified against the real-world data) by recent works on the subject and our understanding of certain linguistic phenomena related to anaphora and coreference in highly inflectional languages.

Based on empirical data collected in the process of the corpus creation, we discuss how certain linguistic features of an inflectional language influence the annotation

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

schema and resolution tools. Statistics of the respective phenomena and inter-annotator agreement values are also presented.

2 Mentions and Coreference Clusters

Our annotation schema defines mentions as nominal groups (NGs) taking into account their *potential* referentiality, which is based on observation that certain stylistically marked cases allow of using traditionally non-referential expressions in referential contexts, as in (1)[1].

(1) *Nie wahał się włożyć kij w mrowisko.*
Mrowisko to, czyli cały senat uniwersytecki, pozostawało zwykle niewzruszone.
'He didn't hesitate to put a stick into an anthill (i.e. to provoke a disturbance).
This anthill, i.e. the whole university senate, usually didn't care.'

The following phrase types are treated as NGs:

1. nouns, nominal phrases, personal pronouns,
2. numeral groups (*trzy rowery* = '*three bicycles*'),
3. adjectival phrases with elided nouns ('*bukiet z czerwonych kwiatów i z tych niebieskich*' = '*a bouquet of the red flowers and these blue ones*'),
4. date/time expressions of various syntactic structures,
5. coordinated nominal phrases, including conjoining commas (*krzesło, stół i fotel* = '*a chair, a table, and an armchair*').

The boundaries of mentions are set to involve as broad contexts as possible to maximally disambiguate entities (to refer to '*the car which hit my wife*', not just '*the car*'). Elements allowed within mention contents are:

1. adjectives and adjectival participles in agreement (with respect to case, gender and number) with superior noun (*duży czerwony tramwaj* = '*big red tram*'),
2. subordinate nouns in the genitive case (*kolega brata* = '*my brother's colleague*'),
3. nouns in apposition (*malarz pejzażysta* = '*landscape painter*', pol. '*painter landscapist*'),
4. subordinate prepositional-nominal phrases (*koncert na skrzypce i fortepian* = '*a concerto for violin and piano*'),
5. relative clauses (*dziewczyna, o której rozmawiamy* = '*the girl that we talk about*').

The deep structure of NGs, i.e. all embedded phrases not containing finite verb forms having semantic heads other than those of the superior phrase (which reference different entities), are annotated, therefore the fragment *dyrektor departamentu firmy* '*manager of a company department*' contains 3 nominal phrases, referencing the manager of a company department ('*dyrektor departamentu firmy*'), the company department ('*departamentu firmy*') and the company ('*firmy*') alone.

This assumption is also valid for coordination — we annotate both the individual constituents and the resulting compound, because they can be both referred to:

[1] Henceforth, we will mark coreferent NGs with (possibly multiple) underlining, and non-coreferent NGs with dashed underlining.

(2) *Jan z Marią przyszli na obiad. Oni są przemili, zwłaszcza Maria.*
'Jan and Maria have come to dinner. They are charming, especially Maria.'

Discontinuous phrases and compounds are also marked:

(3) *To był delikatny, że tak powiem, temat.*
'It was a touchy, so to speak, subject.'

Zero anaphora, very frequent in Polish due to rich inflection of verbs, is marked by including verbs (whose pronominal subjects are elided) into coreference clusters, as in (4). Zero anaphora is not considered for objects and complements.

(4) *Maria wróciła już z Francji. ØSpędziła tam miesiąc.*
'Maria came back from France. ØHad$_{singular:feminine}$ spent a month there.'

Coreference clusters group mentions referring to the same discourse-world entity. In our task we concentrate on identity of reference in its strict form (direct reference) with an extension of so called near-identity (see section 4).

3 Related Work

The annotation schema resumed in the previous section was presented in details in [1] It was also compared with several approaches to coreference annotation in languages that show coreference-relevant morphosyntactic similarities with Polish, i.e. Slavic languages [2,3] and Spanish [4,5] due to its frequent zero subject. A recent study dedicated to English [6] was also considered for obvious dominance reasons in NLP. In view of this contrastive study our annotation schema shows three major novel aspects which we deeply analyse in this paper:

- large-scale experiments with near-identity,
- introduction of dominant expressions,
- pointing at semantic rather than syntactic heads.

To a lesser extent, the fact of systematically taking zero subjects into account in our approach brings some new insights into the state of the art. In order to further verify these novelty issues, we present below some other bibliographic references reporting on coreference annotation schemas which were applied to corpora of about 200 thousand words or more — according to [7], p. 10.

The series of ACE (Automatic Content Extraction) program has been carried out from 1999 to 2008 for a varying number of languages, including Arabic, Chinese, English, and Spanish. It was meant to boost the development of automatic detection and characterization of meaning conveyed by human texts. The ACE-2007 annotation guidelines for Spanish [8] gives the rules of annotating and disambiguating entities. The entity typology is rather fine-grained: it consists of 7 main types (person, organization, geopolitical entity, etc.) and several dozens of subtypes (individual, group, governmental, commercial, etc.). Two coreference relations are considered: identity and apposition. The former is further subdivided into generic and non-generic. Mentions are

NGs (including attached prepositional phrases and relative clauses) and can be nested (*the president of Ford*). Each NG should have its (syntactic) head marked. Heads are marked but their definition is confusing (the syntactic head can be multi-word, and then its last token is marked). Semantic heads different from syntactic ones are not an issue. The problem of Spanish zero subject is not mentioned.

[9] describe the annotation of the 22-thousand-sentence Tübingen treebank of German newspaper texts (TüBa-D/Z) with a set of 7 coreference relations (coreferential, anaphoric, cataphoric, bound, split antecedent, instance, and expletive). These are no equivalence relations: they are non-symmetric and mostly non-transitive, thus they do not divide the set of referents into disjoint clusters. For instance, the split antecedent relation holds between a plural expression and a mention of a single member. E.g. in '*John and Mary were there... Both...*', *John* and *both*, as well as *Mary* and *both* are coreferential but *John* and *Mary* are not. Potential markables are definite NGs, personal pronouns, relative, reflexive, and reciprocal pronouns, demonstrative, indefinite and possessive pronouns. All of them correspond to nodes of the already existing parse trees resulting from prior syntactic annotation. Unlike in our approach, predicative nominal groups (NGs) seem to be considered coreferential with subjects. Zero subjects are not an issue in German. Neither dominant expressions nor semantic heads are mentioned.

[10] address the construction of a 182-thousand-word Italian Content Annotation Bank of newspaper texts (I-CAB). Mentions are NGs, possibly containing modifiers, prepositional complements or subordinate clause, representing entities (persons, organizations, etc.) or temporal expressions. ACE-2003 annotation guidelines for English are adopted and extended, to cover notably clitics contiguous with verbs (*vederlo*) and coordinated expressions (*John and Mary*). The paper announces future annotation of relations between entities but it is unclear where its results have been described.

[11] present NAIST, a 38-thousand sentence Japanese corpus annotated for coreference and predicate-argument relations (including nominal predicates relating to events). They consider identity-of-reference relations for the former, and both identity-of-reference and identity-of-sense relations for the latter. They pay a special attention to zero anaphora, whose role — not only as a subject but as an object or a complement as well — is particularly visible when coreference and predicates' arguments are annotated jointly. Namely, the frames for elided arguments have to be filled out with antecedents appearing in other sentences than the predicate itself. The reported inter-annotator agreement for coreference annotation is 0.893 for recall and 0.831 for precision. No mention of dominant expressions or semantic heads is made.

[12] describe OntoNotes, a system of multi-layered annotated corpora in English, Chinese and Arabic. It is supposed to make up for the drawbacks of previous annotation schemata, mainly MUC and ACE in that the coreference annotation is not restricted to NGs and a larger set of entity types is considered. The English corpus consists of 300-thousand-word newspaper texts, later completed by broadcast conversation data [13]. All data have been previously annotated for syntax, verbal arguments and word senses (by linking words to nodes of an external ontology). Thus, mention candidates correspond to nodes of pre-existing syntax trees. As in ACE, two coreference relations are considered: identity and apposition. The main mention candidates are specific NGs,

pronouns (*they*, *their*) and single-word verbs coreferent with noun phrases (e.g. *the sales rose by 18%* ← *the strong growth*). Expletive (*it rains*), pleonastic (*there are*) and generic (*you need*) pronouns are not marked. Generic, underspecified or abstract entities are only partly considered by identity: those cannot be interlinked among themselves, even if they can be linked with referring pronouns (*parents* ← *they*). Nested structures are generally marked but exceptions occur in dates (e.g. no subphrase of *Nov. 2, 1999* is coreferent with *November*). Only intra-document coreference is annotated, thus dominant expressions (motivated in our approach notably by future inter-document coreference annotation) are not an issue. Zero subjects are addressed with respect to pro-drop Arabic and Chinese pronouns [13]. Since such pronouns are materialized in parse trees as separate nodes, their inclusion in coreference chains is straightforward. The existence of semantic heads different from syntactic ones is not mentioned.

[14] describe a 200-thousand-word coreference-annotated corpus of Dutch newspaper texts, transcribed speech and medical encyclopedia entries. Its annotation schema is largely based on the MUC-7 annotation guidelines for English[2]. Annotation focuses mainly on identity relations between NGs but other non-equivalence relations are also introduced: bound relations (*everybody* ← *they*), bridging relations (e.g. superset–subset or group–member), and predicative relations (e.g. John is a painter). Syntactic heads are pointed at but semantic heads different from syntactic ones do not seem to be an issue (e.g. *tons* is the head of *200,000 tons of sugar*). The ideas of dominant expressions and zero subjects are not present. Predicative NGs and appositions are considered as mentions coreferent with their subjects. Discontinuous NGs are taken into account. The inter-annotator agreement measured as the MUC-like F-score, is 0.76 for identity relations, 0.33 for bridging relations and 0.56 for predicative relations.

The above bibliographic study confirms the novelty of annotating at least three coreference-related aspects:

- large-scale annotation of near-identity introduced by [4], cf. Section 4, which obviously could not be performed by approaches published before 2010, and seems not to have been applied since then except in our work;
- dominant expressions (cf. Section 5), whose idea appears in none of the studied approaches, despite its utility e.g. for cross-document coreference annotation;
- semantic heads (cf. Section 6), whose difference from syntactic heads does not seem to be an issue in other languages than Polish, for which coreference annotation has been performed.

As for taking zero subjects into account, the five major approaches concerned are [2] (for Bulgarian), [4] (for Spanish), [11] (for Japanese), [13] (for Arabic and Chinese) and ours (for Polish). In Section 7 we revisit this notion in order to report on its nature and frequency in our corpus.

[2] See `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`.

4 Near-Identity

Near-identity is a novel coreference relation defined in [15]. Our understanding of this concept, as discussed in [1] includes two phenomena: (i) two mentions refer to the *same* entity but the text suggests the opposite (refocusing, e.g. *pre-war Warsaw* vs. *today's Warsaw*), (ii) two mentions refer to *different* entities but the text suggests the opposite (neutralization, e.g. *wine* as a bottle vs. its contents). [16] state that the binary distinction between coreference (identity) and non-coreference (non-identity) is too limited, since a continuum of values exists between these two extreme cases. Near-identity should help in bridging this gap. The same paper puts forward a fine-grained typology of near-identity with 4 types (name metonymy, meronymy, class, spatio-temporal function) and 15 subtypes (role, location, organization, etc.).

While the idea of near-identity itself was inspiring, the applicability of its typology seemed uncertain. Thus, before including it in our annotation schema we first studied the reliability of detecting near-identity alone, i.e. regardless of its type. As discussed in Section 9, the value obtained for inter-annotator agreement in untyped near-identity links, in terms of Cohen's κ, was only 0.222. Some annotators never even used the near-identity links, which proves that the concept is hard to capture.

These results bring strong doubts not only about the utility of the mentioned typology but also of the near-identity as such. The concept of near-identity might be, in our opinion, a result of mixing two different levels of language: the meaning of a word and its reference. The former is independent of the context while the latter is a function of a word used in a given context. Words very often have common elements of meaning – that is why it was possible to create semantic networks like WordNet (web of words which are linked with each other without any context). Reference though is related more to pragmatics than to semantics. Very often phrases formed with words sharing no semantic elements can refer to the same referent (e.g.: *football players of Polonia* and *"Black shirts"*), and the link between such phrases can be established only due to the external knowledge (Polonia football players wear black shirts). On the other hand the same words can refer to different referents, e.g.:

(5) *Te tipsy bardzo niszczą paznokcie [. . .] ostatnio właśnie już mi całe paznokcie odrosły już nawet już nie mam takiej strasznie zniszczonej płytki po tych paznokciach.*

'These artificial nails damage nails a lot [. . .] lately my nails have just grown back I don't have so awful haggard nail any more after those nails.'

In this text the three occurrences of *nails* have different referents (although still the same basic semantics): generic nails, nails of the speaker, and artificial nails.

Our experience with the corpus annotation shows that people usually have no problem with distinguishing these two linguistic levels: the word meaning and the word reference. Conversely, 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.

5 Dominant Expressions

In every cluster we indicate the *dominant expression*, i.e. the expression that carries the richest semantics or describes the referent the most precisely. The best candidates for dominant expressions are named entities, as well as periphrastic phrases that denote a particular object in the discourse world, e.g:

(6) Cluster: *David Beckham, rozgrywający Realu Madryt* *'David Beckham, Real Madrid playmaker'* Dominant expr.: *David Beckham*

In many cases, pointing at the dominant expression helps the annotators sort out a large set of pronouns denoting various persons (e.g. in fragments of plays or novels). We think that it might also facilitate cross-document annotation or the creation of a semantics frame containing different descriptions of the same object.

In 62% of all cases, the dominant expression was selected from among NGs contained in the cluster. 77% of them were taken without any changes (which means that there was the base form of the NG in the cluster) as in (7), while 23% of them were transformed into their base forms.

(7) Cluster: *tamtejszy dziennikarz, dziennikarz, Ja, napisał, pismak*
'local journalist, journalist, I, wrote, hack.'
Dominant expr.: *tamtejszy dziennikarz*
'local journalist'.

For 38% of the clusters, the dominant expression was not present in the text but given by the annotator instead (e.g., *Halloween* for the cluster containing a repeated phrase: *tej okazji* 'this occasion'). This was necessary in particular when the cluster consisted of verb forms only, e.g.:

(8) Cluster: *stwierdzili, powiedzieli* *'stated, said'*
Dominant expr.: *lekarze w Polsce* *'doctors in Poland'*

As mentioned in Section 9, dominant expressions can be annotated with a much higher reliability (66.78%) than near-identity. A detailed study of disagreement cases shows that many of them are superficial rather than essential. They are due e.g. to different letter case or spelling errors in the dominant expressions, or to the fact that some annotators produce the base form of the dominant expressions while others cite the (inflected) forms occurring in the corpus. Such cases may be corrected mostly automatically, which will enhance the inter-annotator agreement indicator.

6 Semantic Heads

For each mention, its semantic head is selected, being the most relevant word of the group in terms of meaning. The semantic head of typical nominal group is the same element as the syntactic head but in numeral groups the numeral is the syntactic head, and the noun is the semantic head. Numeral groups are regarded as nominal groups in

our project (e.g., *dużo pieniędzy* 'a lot of money', *trzech z was* 'three of you'). They can be also embedded in other nominal groups (e.g., *sąsiad dwóch kobiet* 'neighbour of two women'). In these examples, words *dużo, trzech, dwóch* 'a lot, three, two' are syntactic heads and *pieniędzy, was, kobiety* 'money, you, women' are semantic heads.

The reason why we are interested in semantic rather than in syntactic heads is the same as when we admit a very broad definition of nominal groups (including numeral phrases, some adjectival phrases, etc., cf. Section 2). Namely, coreference is a phenomenon on the level of semantics and discourse more than syntax. Thus, understanding the semantically central elements should help establish discourse links, notably in future automatic coreference resolvers. In particular, it seems promising to examine agreement in case, gender, number, synset, etc. between semantic heads in potentially co-referring mentions.

As shown in Section 9, the reliability of annotating the semantic heads was very high (97.00%). Disagreement resulted mainly from inattention in distinguishing syntactic from semantic heads, e.g., a) adjectives or numerals were selected instead of nouns, b) the head of a subordinate phrase was selected instead of the head of the main phrase (e.g., *metropolii* 'metropolis' was marked as the semantic head in: *niedobrą dzielnicę jakiejś wieloetnicznej metropolii* 'bad quarter of a multi-ethnic metropolis').

7 Zero subjects

Similarly to most other Slavic languages, Polish grammar permits independent clauses to lack explicit subjects. The form of the null referent is then partially indicated by the morphology of the verb. We annotate such cases with identity links, as in (9) while other types of elliptic expressions are not linked, unlike in [11], cf. (10)-(11).

(9) *Maria wróciła już z Francji. ØSpędziła tam miesiąc.*
'Mary returned from France. She spent a month there.'

(10) *Janek kupił duże pudełko czekoladek, ale niewiele Ø już zostało.*
'John bought a huge box of chocolates, but there were just a few Ø left.'

(11) *Czytałeś książki Lema? Czytałem Ø.*
'Have you read Lem's books? I have Ø.'

Even with such limitation, the phenomenon seems frequent — there are 4678 coreference clusters containing at least one zero subject (26.89% of the total number of non-singleton clusters).

8 Pronominal Coreference and Other Issues

Originally, we had excluded some types of pronouns (indefinite, negative, reflexive, interrogative) from the annotation on the assumption of their non-referentiality. Surprisingly, the analysis of the corpus showed that they can frequently form coreferential chains. Very often an indefinite pronoun is a subject of a verb sequence (in which the second verb is a zero subject verb), e.g.:

(12) *Jak ktoś jest zazdrosny, znaczy, że naprawdę Ø kocha.*
'If someone is jealous, it means, that (he/she) really loves.'

Sometimes pronouns make a typical anaphoric link: a demonstrative (pronoun) refers to indefinite pronoun used in the first clause, cf.:

(13) *Jeśli coś przestanie być potrzebne, można to usunąć z dysku, zwalniając miejsce na inne zasoby.*
'If something is no longer needed, one can remove it from the disk to save on storage for other resources.'

Indefinite pronouns can also implicitly refer to a specific person. The speaker in the example below does not want to speak openly about the former director and uses the indefinite pronoun *kogoś* 'someone' instead:

(14) *Po rezygnacji z pracy w szpitalu były dyrektor zniknął z życia publicznego. ØWrócił dopiero, gdy starosta Andrzej Barański zaproponował mu współpracę. Posunięcie starosty, wywołało ostrą reakcję kilku radnych. W trakcie ostatniej sesji kilkakrotnie pytano, czy nowy pracownik ma odpowiednie kwalifikacje, by zdobywać dla powiatu unijne środki pomocowe. – Uważam, że powołanie kogoś, kto nie sprawdził się w szpitalu i jako szef spółdzielni mieszkaniowej, może budzić wątpliwości — mówi Wojciech Wenecki.*
'After giving up the job in the hospital, the former director had disappeared from the public life. He came back only when starost Andrzej Barański offered him cooperation. The move of the starost provoked sharp reaction of several councillors. During the last session they repeatedly asked if the new employee has suitable qualifications to obtain for the local government administration UE aid resources. - I think that appointing someone who haven't performed well in the hospital and as chief of the housing cooperative may raise doubts — says Wojciech Wenecki.'

The examples of coreferential chains containing indefinite pronouns show that these pronouns should have been allowed in coreference chains. We wish to reconsider this phenomenon at the end of the annotation process. It makes us think that the problem of coreference in the text might be somehow different than the one of the reference to the world. Maybe it should be examined as a separate phenomenon.

9 Inter-Annotator Agreement

For the purpose of measuring the inter-annotator agreement, a sample of the corpus (henceforth called the IAA-sample) was annotated independently by two annotators. More precisely, several annotators participated in the experiment but each text was annotated by exactly two of them. The IAA-sample consisted of 210 texts selected so as to uniformly represent the 14 existing text genres (prose, press, dialogues, etc.). It contained 60674 tokens in total, i.e. about 12% of the whole corpus.

The establishment of the inter-annotator agreement with respect to mention detection remains a challenge [6]. Thus, we based our estimation on the F-measure (which does not take agreement by chance into account) — it amounted to 85.55% in the IAA-sample. Namely, the two annotators produced 20420 and 20560 mentions, respectively, and 17530 of them had identical borders in both sets (including the internal borders in

Table 1. Detailed inter-annotator agreement

Mentions	**Near-identity**	**Dominant expressions**	**Semantic heads**	**Identity clusters**
$F_1 = 85.55\%$	$\kappa = 0.222$	66.78%	$S = 97.00\%$	$\alpha = 79.08\%$

case of discontinuous mentions). Further calculations, resumed in Table 1, take only those 17530 mentions into account whose borders were marked identically by both annotators.

Chance-corrected agreement of near-identity links with Cohen's κ [17] was calculated for each text separately (because cross-document links are not allowed) and then averaged. For a single text the agreement was measured for the decisions on every pair of mentions (marked by both annotators) in the text – whether they were marked as near-identical or not. The probability of the annotator marking two mentions as near-identical was estimated for each text and annotator separately. If there were no near-identity links in a text its agreements was equal to 1. If however one annotator marked some near-identity links in a text and the other annotator marked no such link in the same text, the agreement for this text was 0 (because the expected agreement was the same as the observed agreement). The final result is very low – only 0.222. may be due to the methodology of measuring chance-corrected agreement. Namely, if one annotator marked a single near-identity relation in a given text and the other annotator marked no such relation in the same text, the κ for this text was equal to 0. Frequently (128 times) a link was marked as near-identity by one annotator and as identity by the other. These figures prove the difficulty of reliably annotating near-identity links.

The agreement in annotating dominant expressions was calculated only on those non-singleton clusters which were identically designated by both annotators (among the identically delimited mentions). There were 6162 such clusters, out of which 4115 (about 66,78%) had the same dominant expression. If we count such proportion for only one mention from each cluster (as each cluster member has the same dominant expression), for 1818 such cluster "representatives" 1146 mentions (63,04%) have the same dominant expression in both annotations. No chance-corrected analysis was conducted because the dominant expressions could have been entered by the annotators as free text, which jeopardizes the probability model of agreement by chance.

The agreement in annotating semantic heads was evaluated in terms of the chance-corrected agreement (for identically delimited mentions). Uniform probability distribution among all possible head choices was assumed and the S measure [18] was used. The observed agreement and the expected agreement were close to 99.05% and 68.32%, respectively. The former was quite high mainly because of the large number of mentions consisting of only one token (therefore only one possible head). These two values yield the final S result of approximately 97.00%.

For completeness, we also present the chance-corrected inter-annotator agreement of the identity clustering task, calculated according to the version of weighted Krippendorf's α [19] proposed by Passoneau in [20].

10 Conclusions and Perspectives

We have presented a detailed study of several particularly interesting phenomena related to coreference annotation in an inflectionally rich language such as Polish. We have applied the novel concept of near-identity [15] on a large scale and we came to the conclusion that it cannot be reliably applied in coreference annotation. We also argue that semantic heads are more relevant to coreference than syntactic heads. We have introduced a novel idea of dominant expressions which represent the expressions carrying the biggest semantic load with respect to a reference cluster. These may be useful e.g. in cross-document coreference annotation. Finally, we have reviewed our previous conviction that indefinite, negative and other particular types of pronouns can never appear in coreference chains. We hope that all these observations may contribute to a high-quality methodology and usefulness of future coreference annotation projects, particularly in highly inflected languages.

References

1. Ogrodniczuk, M., Zawisławska, M., Savary, A., Głowińska, K.: 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., Heidelberg, Springer-Verlag (2013) 394–407
2. Osenova, P., Simov, K.: BTB-TR05: BulTreeBank Stylebook. BulTreeBank Version 1.0. Technical Report BTB-TR05, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia, Bulgaria (2004)
3. 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), Goa, India, AU-KBC Research Centre, Anna University, Chennai, AU-KBC Research Centre, Anna University, Chennai (2009) 1–16
4. Recasens, M., Martí, M.A.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation **44**(4) (2010) 315–345
5. Korzen, I., Buch-Kromann, M.: Anaphoric relations in the Copenhagen Dependency Treebanks. In: Proceedings of DGfS Workshop, Göttingen, Germany (2011) 83–98
6. Poesio, M., Artstein, R.: Anaphoric Annotation in the ARRAU Corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, European Language Resources Association (2008) 1170–1174
7. Recasens, M.: Coreference: Theory, Annotation, Resolution and Evaluation. PhD thesis, Department of Linguistics, University of Barcelona, Barcelona, Spain (2010)
8. Consortium, L.D.: ACE (Automatic Content Extraction) Spanish Annotation Guidelines for Entities (2006) Available at `http://projects.ldc.upenn.edu/ace/docs/Spanish-Entities-Guidelines_v1.6.pdf` (accessed on Feb. 18, 2013).
9. 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, Ann Arbor, Michigan, USA (2005) 13–20
10. Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V.B., Sprugnoli, R.: I-CAB: the Italian Content Annotation Bank. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, European Language Resources Association (2006) 963–968

11. 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), Stroudsburg, PA, USA, Association for Computational Linguistics (2007) 132–139
12. 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), Washington, DC, USA, IEEE Computer Society (2007) 446–453
13. Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M.: OntoNotes Release 4.0 (2010) Available at `http://www.bbn.com/NLP/OntoNotes` (accessed on Feb. 18, 2013).
14. 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), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) 144–149
15. Recasens, M., Hovy, E., Martí, M.A.: Identity, non-identity, and near-identity: Addressing the complexity of coreference. Lingua **121**(6) (2011)
16. Recasens, M., Hovy, E., Marti, M.A.: A Typology of Near-Identity Relations for Coreference (NIDENT). In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, European Language Resources Association (2010) 149–156
17. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1) (1960) 37–46
18. Bennet, E.M., Alpert, R., Goldstein, A.C.: Communications through limited response questioning. Public Opinion Quarterly **18** (1954) 303–308
19. Krippendorff, K.H.: Content Analysis: An Introduction to Its Methodology. 2nd edn. Sage Publications, Inc (December 2003)
20. Passonneau, R.J.: Computing reliability for coreference annotation. In: LREC, European Language Resources Association (2004)