Coreference annotation schema for an inflectional language[⋆]

Maciej Ogrodniczuk[1], Magdalena Zawisławska[2],
Katarzyna Głowińska[3], and Agata Savary[4]

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Institute of Polish Language, Warsaw University
[3] Lingventa
[4] François Rabelais University Tours, Laboratoire d'informatique

Abstract. Creating a coreference corpus for an inflectional and free-word-order language is a challenging task due to specific syntactic features largely ignored by existing annotation guidelines, such as the absence of definite/indefinite articles (making quasi-anaphoricity very common), frequent use of zero subjects or discrepancies between syntactic and semantic heads. This paper comments on the experience gained in preparation of such a resource for an ongoing project (CORE), aiming at creating tools for coreference resolution.
Starting with a clarification of the relation between noun groups and mentions, through definition of the annotation scope and strategies, up to actual decisions for borderline cases, we present the process of building the first, to our best knowledge, corpus of general coreference of Polish.

1 Introduction

Although the notion of coreference is no longer a subject of much controversy and there are many more or less ready-to-use annotation guidelines available, in a case where a "new" language is being investigated — which has not yet received any formalized coreference description — they usually need to be supplemented with details specific to this language, and the task of creating a coreference corpus requires establishing detailed rules concerning annotation scope, strategies and typology of coreferential constructs.

This paper comments on the experience gained in the process of creating the first substantial Polish corpus of general coreference (500K words and 160K mentions are intended), which is currently being completed. We hope our analysis can provide a valuable source of information for creators of new coreference corpora for other inflectional and free-word-order languages. We believe that they could particularly benefit from studying our assumptions based on such specific properties as the absence of definite/indefinite articles (introducing quasi-anaphoricity), frequent use of zero subjects or discrepancies between

[⋆] 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).

syntactic and semantic heads. These phenomena are fundamental for building computational coreference resolvers.

Construction of a large high-quality corpus is of great importance in the context of further tasks in the ongoing CORE project, whose central aim is the creation of an efficient coreference resolver for Polish. We wish to surpass the previous early attempts, both rule-based [1] and statistical [2], which yielded tools trained and evaluated on a very limited amount of data. We believe that a more efficient tool can boost the development of higher-level Polish NLP applications, on which coreference resolution has a crucial impact [3]. Such applications include: 1) machine translation (when translating into Polish, coreferential relations are needed to deduce the proper gender of pronouns), 2) information extraction (coreference relations help with merging partial data about the same entities, entity relationships, and events described at different discourse positions), 3) text summarization, 4) cross-document summarization, and 5) question answering.

2 Reference, Anaphora and Coreference

In order to define the scope of coreference annotation we must bring back the underlying concept of *reference* to discourse-world objects, leading to an important limitation: only nominal groups (NGs), including pronouns, can be referencing expressions.

Recall that coreference annotation is usually performed (and evaluated) in two steps: (i) identifying *mentions* (or *markables*), i.e. phrases denoting entities in the discourse world, (ii) clustering mentions which denote the same referent. Consequently, the definition of a mention, and of the difference between a mention and a NG in particular, is of crucial importance to the whole process. We, unlike e.g. [4], consider this difference too controversial to be reliably decided in a general case.

For instance, multi-word expressions (MWEs) show opaque semantics, thus the NGs they include might be seen as non-referential. However, most MWEs do inherit some part of the semantics of their components, and might be coreferential in some stylistically marked cases, as in (1)[1]. Defining a clear-cut frontier between non-referential and referential NGs in these cases seems very hard.

(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.'

Thus, our annotation process consists in retaining – as mentions – all NGs (whether referential or not), and establishing coreference chains among them

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

wherever appropriate. In other words, we do not distinguish non-referential NGs from referential, but non-coreferential, NGs (e.g. singleton mentions). This decision obviously has a big influence on coreference resolution quality measures which take singleton mentions into account.

We also consider that the reference is context-dependent, not surface-form dependent, cf.

(2) *Spotkałam nową dyrektorkę. Osoba ta zrobiła na mnie dobre wrażenie.*
'I met the new manager. This person made a good impression on me.'

(3) *Nasza nowa dyrektorka to młoda kobieta.*
'Our new manager is a young woman.'

(4) *Nasza dyrektorka, młoda kobieta, przyszła na spotkanie.*
'Our manager, a young woman, came to the meeting.'

(5) *Młoda kobieta, która przejęła funkcję dyrektora, zrobiła na mnie dobre wrażenie.*
'The young woman who overtook the manager's duties made a good impression on me.'

In example (2) the NG *osoba ta* ('this person') has a defined referent, i.e. a concrete human being the speaker refers to. In (3)–(4), the nominal group *młoda kobieta* does not carry reference, but is used predicatively — assigns certain properties to the subject of the sentence. Our understanding of nominal coreference is therefore strictly limited to direct nominal constructs; expressions that do not denote the object directly are not included in coreference chains.

There is an additional, operational, criterion that we admit, contrary to many common coreference annotation and resolution approaches, e.g. [5]. If semantic identity relations between NGs are directly expressed by the syntax, we see no point in including them in coreferential chains. Typical cases here are predicates, as in (3), relative clauses, as in (5), and appositions, as in (4), where we see one, not two, mentions in the NG *Nasza dyrektorka, młoda kobieta* ('Our manager, a young woman').

Such definition of reference creates links between the text and discourse world and is of different nature than *anaphora* — an inter-textual reference to previously mentioned objects. Even if, in most cases, anaphora and coreference cooccur, it is not necessarily the case. In example (6), the underlined NGs are anaphoric but not coreferential, cf. [3]. Conversely, NGs in separate texts can be coreferential, but not anaphoric.

(6) *Człowiek, który dał piękne kwiaty swojej żonie, wydał mi się sympatyczniejszy niż człowiek, który odmówił kupienia ich swojej.*
'The man who gave beautiful flowers to his wife seemed nicer to me than the one who refused buying them for his (wife).'

3 Scope of Annotation

3.1 Mentions

As it was said in the previous section, all NGs (both referential and non-referential) are marked as mentions, while coreference chains can only concern referential NGs (mentions). In particular, some types of nominal pronouns, which seem non-referential by nature, are marked as mentions (since they are NGs) but never included in coreference chains: (i) indefinite pronouns (*ktoś* 'somebody'), (ii) negative pronouns (*nic* 'nothing'), (iii) interrogative pronouns (*kto* 'who')[2]. Note also that some Polish lexemes designated traditionally as pronouns behave morphosyntactically like other parts of speech. Namely, demonstrative pronouns introducing subordinates other than relative clauses (*o tym, że* 'of-this-that = of the fact that') are in fact parts of correlates. The reflexive pronoun (*się* 'oneself') is a particle. Finally, possessive pronouns (*mój* 'mine') behave like adjectives. Consequently, these three types of pronouns are never considered as NGs, i.e. they are never marked as mentions.

Finally, coreference relations between phrases other than nominal ones (e.g. *tam* 'there') are obviously never marked, since only NGs are considered as mentions.

3.2 Types of Relations

The major goal of coreference annotation is to determine the type of relation holding among discourse-world entities referred to by two or more mentions. We are essentially interested in identity relations. We also consider, experimentally, the notion of *near-identity* proposed by [6]. Due to the pioneering (wrt. Polish) nature of our project, all other types of relations (whether among entities or among mentions) have been explicitly ruled out, including non-identity, indirect anaphora, bound anaphora, ellipses (with the exception of zero anaphora), predicative relations, and identity of sense.

Identity Textual techniques used in Polish to signal the identity of referred entities are manifold:

- lexical and grammatical (personal and demonstrative pronouns),
- stylistic, such as *synonymy*,
- lexical and grammatical anaphora and cataphora between nominal groups,
- "quasi-anaphora" – when a group with syntactic-functional properties of anaphora introduces new information, e.g.

  (7) *Duszą towarzystwa był zięć Kowalskich. Młody prawnik właśnie wrócił ze Stanów.*

  'Kowalski's son-in-law was the life and soul of the party. The young lawyer had just returned from the US.'
- zero-anaphora, very frequent in Polish – a personal pronoun may be omitted whenever the subject's person and gender are recognizable from the verb's

[2] Surprisingly enough, recent experiences show that such pronouns may be referential in stylistically marked cases such as: *Ktoś ukradł łopatę. Ten sam ktoś zniszczył ogrodzenie.* 'Someone stole the spade. The same someone broke the fence.'. We wish to review these cases in the final annotation stage.

agreement; therefore the annotation denoting the missing referential NG is most naturally attached to the verb, as in example (8).[3]

(8) *Maria wróciła już z Francji. ØSpędziła tam miesiąc.*

'Maria came back from France. ØHad$_{singular:feminine}$ *spent a month there.'*

Note that some approaches introduce a typology of coreference links which takes the above techniques into account. We, conversely, think that these types of linguistic data should be documented either at other annotation levels or in external linguistic resources. One – formal and practical – reason is that we see coreference chains as clusters, i.e. results of splitting the set of all mentions via a (unique and uniform) equivalence relation. If subtypes of this relation were to be used, clustering would no longer be possible and each pair of coreferent mentions would have to be marked explicitly. Such a methodology might not only have a prohibitive cost in some types of texts but would also be hard to evaluate by classical quality measures.

Near-identity [7] define the notion of *near-identity*, taking place in two contexts called refocusing and neutralization. Our understanding of these phenomena involves the following:

- Refocusing – two mentions refer to the *same* entity but the text suggests the opposite. This stylistic technique is often used to account for a temporal or spatial change of an object as in[4]:

  (9) *Warszawa przedwojenna i ta z początku XXI wieku*

  'Pre-war Warsaw and the one at the beginning of the 21st century'
- Neutralization – two mentions refer to *different* entities but the text suggests the opposite. This situation is typical for metonymy, as in example (10), where a container and its contents are merged, and unlike (11), which is a case of a classical identity:

  (10) *Wziął wino z lodówki i wypił je.*

  'He took the wine from the fridge and drank it.'

  (11) *Wziął wino z lodówki i włożył je do torby.*

  'He took the wine from the fridge and put it into the the bag'.

[6] put forward a detailed typology of near-identity relations. However, in the experimental annotation stage of our project, the annotators marked very few examples of near-identity, most of them concerning, in fact, more typical semantic relations, like homonymy, meronymy, metonymy, element of a set or — sometimes — hypernymy, e.g.:

(12) *Cała Warszawa była właściwie jednym wielkim cmentarzem. Ginęli ludzie, mnóstwo ludzi! Na podwórku, już tak po 15 sierpnia, praktycznie codziennie był pogrzeb przed kapliczką. Warszawa była bardzo pobożna...*

[3] Elliptical constructions concerning functions other than the subject, as in *Czytałeś książki Lema? Czytałem Ø.* 'Did you read Lem's books? I read Ø.' are not annotated in our model.

[4] Henceforth, near-identity-related mentions will be marked by a wavy underline.

> *’The whole Warsaw was in fact one big graveyard. People were dying, plenty of people! After the 15th of August there were funerals in the courtyard, in front of the chapel, almost every day. Warsaw was very pious...*[5]

That experience made us think that near-identity is either too infrequent to deserve a rich typology, or too hard to capture and classify reliably by annotators. That is why we mark near-identity links in our corpus, but we assign no type labels to them. Once the annotation has been completed, we plan to compare our examples of near-identity more thoroughly with the types proposed in [6].

3.3 Dominant Expressions

Despite the fact that all mentions within a cluster are (mathematically speaking) equivalent, we enrich each cluster with a pointer towards the *dominant expression*, i.e. the one that carries the richest semantics. For instance in the following chain the last element is dominant: *stworzenie* ’creature’ → *zwierzę* ’animal’ → *pies* ’dog’ → *jamnik* ’dachshund’.

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 linking mentions within different texts, and creating a semantics frame containing different descriptions of the same object.

4 Annotation Strategies

4.1 Mention Boundaries

In order to encompass the wide range of mentions, we set the boundaries of nominal groups as broadly as possible. Therefore, an extended set of elements is allowed within NG contents, i.e., 1) adjectives as well as adjectival participles in agreement (with respect to case, gender and number) with superior noun, 2) subordinate noun in the genitive case, 3) nouns in case and number agreement with superior nouns (i.e. nouns in apposition); but also 4) prepositional-nominal phrase that is a subordinate element of a noun (e.g. *koncert na skrzypce i fortepian* ‘a concerto for violin and piano’)[6]; 5) relative clause (e.g., *dziewczyna, o której rozmawiamy* ’the girl that we talk about’). Moreover, the following phrases are treated as nominal groups: 1) numeral groups (e.g., *trzy rowery* ‘three bicycles’), 2) adjectival phrases with elided nouns (e.g., *Zrób bukiet z tych czerwonych kwiatów i z tych niebieskich.* ‘Make a bouquet of these red flowers and these blue ones.’), 3) date/time expressions of various syntactic structures,

[5] *The whole Warsaw* refers to the place, while *Warsaw* is a metonymy and refers to people who lived in the city.

[6] Such cases should be distinguished from situations where a prepositional-nominal phrase is a subordinate element of a verb, e.g. *Kupił mieszkanie z garażem.* ‘He bought a flat with a garage.’

4) coordinated nominal phrases, including conjoining commas (*krzesło, stół i fotel* 'a chair, a table, and an armchair').

For each phrase, the semantic head is selected, being the most relevant word of the group in terms of meaning. The semantic head of a nominal group is usually the same element as the syntactic head, but there are some exceptions, e.g., in numeral groups, the numeral is the syntactic head, and the noun is the semantic head.

4.2 Mention Structure

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

(13) *Asia i Basia mnie lubią. One są naprawdę ładne, szczególnie Asia.*
'Asia and Basia like me. They are really pretty, particularly Asia.'

Discontinuous phrases and compounds are also marked:

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

5 Task Organization

Texts for annotation were randomly selected from the National Corpus of Polish [8]. Similarly to this resource, we aimed at creating a 500-thousand-word balanced subcorpus. It was divided into over 1700 samples between 250 and 350 segments each. These samples were automatically pre-processed with a shallow parser detecting nominal groups and their semantic heads[7], and a baseline coreference resolution tool marking potential mentions and identity clusters.

The manual revision of this automatically performed pre-annotation is being carried out in the MMAX2 tool [12] adapted to our needs. In particular, the

[7] More precisely, nominal groups were precomputed from parse trees produced by a shallow parser Spejd [9] supported with the Pantera tagger [10] and a named entity recognizer Nerf [11]. The scope of each NG was heuristically determined in that the longest NG was retained among all potential NGs sharing the same head, e.g. *dyrektor departamentu firmy* 'the director of the department of the company' was retained rather than *dyrektor departamentu.* Nested NGs were then marked within each retained maximal NG, e.g. *[dyrektor [departamentu [firmy]]].*

annotators can correct the pre-annotation results (i.e., remove or change NG marking, change the semantic heads, and modify the content of identity clusters). They can also add mentions and clusters that were not detected by the tool.

Each fragment of the corpus is prepared by one annotator (entitled to change the pre-annotation in all respects) and then checked by the supervising annotator. Note that although the best practice in other annotation tasks [8] is parallel annotation, in which two independent annotators work on each text, and an adjudicator reviews cases of disagreement, we find this practice hard to apply in coreference annotation.

Nevertheless, a part of the corpus, namely 210 texts, has been annotated independently by two people, and then adjudicated by the supervising annotator, in order to check the inter-annotator agreement. Statistics were calculated for each level of annotation separately, i.e., 1) NG scope: $F1 = 85,55\%$, 2) semantic heads: 97%, 3) identity clusters: see below, 4) near-identity links: 22,20%, and 5) dominant expressions: 63,04%.

The agreement in identity clusters annotation was calculated using κ coefficient (taking agreement by chance into account) for the decision about each mention, whether it is a singleton or not. This method is similar to the agreement computation in the An-Cora corpus [4]. We achieved κ of 0.7424. Note also the particularly low agreement in near-identity links which indicates the hardness of this task.

6 Difficult cases

Recall that the annotators' main task in the project is to indicate identity of reference, i.e. that two or more linguistic elements point to the same extralinguistic referent in the text. The task does not sound very difficult, but in practice, things turn out to be different. There are relatively many cases when the recipient cannot decide if the NGs are coreferential or not. The mistakes can occur on the three main levels: lexical, grammatical and conceptual.

6.1 Lexical Level

Frequent occurrences of annotator's "false friends" are due to polysemy and homonymy. In the first sentence of example (15), the noun *misja* 'mission' means "a responsible task someone is entrusted with", while in the second it is "a representation of a country/organization with special assignment". Such cases of graphically identical but non-coreferential NGs may be hard to detect.

(15) *Misja rozpoczęła się 8 kwietnia. W skład misji weszły 24 osoby.*
'The mission started on April 8th. 24 people were members of the mission.'

In some extreme cases, two NGs may be both graphically and semantically identical, and still remain non-coreferential as in example (16). The speaker clearly assigns here different characteristics to both expressions, i.e. he means that there are many different types of mothers, e.g. good ones and bad ones. Detecting this particular type of repetition might be useful in a future automatic coreference resolver.

(16) *Są matki i matki.* *'There are mothers and (then there are) mothers.'*

Another, perhaps the most problematic, lexical issue involves the so-called *co-extension*. It occurs when two or more NGs refer to objects which belong to the same conceptual field. The referents of these NGs can be linked by various semantic relations, e.g. hypo-/hypernymy, meronymy, antonymy, etc. Such relations very often make it difficult to decide if the NGs are coreferential or not. In example (17), the annotator made an excessive cluster in which s/he placed phrases *mity* 'myths' and *mitologia* 'mythology', while a myth is a meronym of a mythology, and a mythology is a holonym of a myth.

(17) *[...] mity są niezastąpionym narzędziem dla psychologa, usiłującego prześledzić wzorce ludzkich zachowań. Wysiłki archeologów, religioznawców, antropologów doprowadziły z jednej strony do porzucenia eurocentrycznego spojrzenia na mitologię...*

'[...] myths are irreplaceable tools for a psychologist, who is trying to follow through the standards of human behaviour. Efforts of archaeologists, specialists in religious studies and anthropologists resulted, on the one hand, in giving up the Eurocentric point of view on the mythology...

Similarly, in example (18), the annotator had a problem with deciding on an identity connection between *okupacja* 'occupation' and *wojna* 'war'. Obviously, those two words have something in common (occupation is a result of war, therefore the WordNet *entailment* relation would be relevant here), but they are not coreferential.

(18) *Od czasu okupacji (...) — Ale tu je masz z powrotem, w metryce — powiedział dyrektor. — Kiedy to jest stara metryka, którą mi odtworzono zaraz po wojnie.*

'After occupation (..) — But here you have it back, in your birth certificate — said the headmaster. — But it is an old birth certificate, which was reconstructed after the war.'

6.2 Grammar Level

We omit the most obvious cases, e.g. a speaker's grammar mistakes, and concentrate on the less typical examples. There are no articles in Polish, therefore the most difficult task for the annotators was to distinguish definite and indefinite objects. In example (19), an annotator wrongly created one cluster in which s/he placed all forms of the word *asystent* ('*assistant*'), e.g.:

(19) *Każdy szanujący się poseł ma asystenta. Asystentami są z reguły ludzie młodzi, ale nie brakuje również szczerze zaangażowanych emerytów. Pracują jako wolontariusze tak jak Marek Hajbos, asystent Zyty Gilowskiej.*

'Every decent Member of Parliament has an assistant. Assistants are usually young people, but there are also genuinely involved senior citizens. They work as volunteers like Marek Hajbos, the assistant of Zyta Gilowska.

6.3 Conceptual Level

A crucial problem in establishing the identity relations between NGs is the lack of annotator's competence in some fields. In example (20) the annotator was unaware that the players of the Silesian football team Ruch Chorzów wear blue shirts.

(20) *W trzecim kwartale 2010 roku Ruch Chorzów zarobił na czysto aż 5.5 mln zł. Wiadomość o zysku Niebieskich na pewno ucieszy jego kibiców.*
'In the third quarter of 2010 year Ruch Chorzów earned 5.5 million zloty net. The news about the profit of the Blues will please their supporters for sure.'

7 Related Work

In this section we review some of the coreference annotation schemes admitted in previous efforts for several languages. While an exhaustive state-of-the-art contrastive study is beyond the scope of our paper, we are particularly interested in languages that show coreference-relevant morphosyntactic similarities with Polish. Slavic languages are obviously of highest importance, but Spanish is also relevant, in particular due to its frequent zero subjects. Finally, for obvious dominance reasons in NLP, we also address one of the most recent studies dedicated to English.

[13] describes BulTreeBank, a syntactically annotated corpus of Bulgarian based on an HPSG model. Coreferential chains link nodes in HPSG trees. Each noun phrase is linked to an (extra-linguistic) index representing, roughly, the discourse-word entity. Coreference is expressed by linking several phrases to the same index. In principle, only coreferential relations which cannot be inferred from the syntactic structure are annotated explicitly, however, some inferable ones are annotated too (it is unclear which ones). Zero subjects and other elliptical elements (e.g. headwords missing due to coordination) are represented whenever they belong to coreference chains. Syntactic trees may help represent split mentions but it is uncertain if they do. Possessive pronouns are considered as mentions. Three relations are encoded: identity, member-of, and subset-of. Discourse deixis is probably taken into account. It seems that the annotation concerns coreference occurring within one sentence only. No inter-annotator agreement results are given.

[14] presents annotation efforts for a 94,000-word English corpus. Special attention is paid to two difficult phenomena: discourse-inherent coreference ambiguity and discourse deixis. The former yields an annotation scheme in which coreference is not an equivalence relation (one mention can appear in several chains). All nominal groups are considered mentions, but some are later marked as non-coreferential. A limited set of bridging relations is taken into account. Problems related to zero subjects, nested, split and attributive NGs, as well as semantic heads, are not discussed and are probably not addressed in the annotation scheme.

[15] extends the coreference annotation in the Prague Dependency Treebank of Czech, a language rather close to Polish. It builds on previously constructed annotation layers including the so-called tectogrammatical layer, which provides ready mention candidates and (probably) their semantic heads. Mentions include nominal phrases and coreferential clauses (discourse deixis). Nested groups are delimited except in named entities (where only embedded groups which are NEs themselves are marked). Attributive phrases are not considered uniformly: appositions are marked as mentions even if they are never included in coreference chains, while predicate nominals are not considered at all. The notable contribution of this approach is addressing a wide range of bridging relations between nominals. The relatively low scores of the inter-annotator agreement might be an evidence that coreference annotation is particularly difficult in Slavic languages.

[16] describes coreference annotation in AnCora-CO, a 400K-word corpus of Spanish and Catalan, for which, like in [15] and [13], other annotation layers had previously been provided, including syntax. Thus, possible candidates for mentions had already been delimited. The annotation schema is rather complete. Three types of relations are considered: identity, predicative link and discourse deixis. Zero subjects are marked, clitic pronouns which get attached to the verb are delimited, embedded and discontinuous phrases are taken into account, and referential NGs are distinguished from attributive ones. Bridging references are not considered.

[17] addresses the anaphoric relations in a parallel, 5-language Copenhagen Dependency Treebank, in which unified annotation of morphology, syntax, discourse and anaphora is being performed. It consists of a 100,000-word Danish corpus with its translations into English, German, Italian and Spanish. Possible specificities of mention detection are not addressed, however, relation typology is extensively discussed. Both coreference and bridging relations (called *associative anaphora*) are considered. The former are split into 6 categories, according to linguistic techniques used to express the coreference, including discourse deixis. The latter count as many as 12 types. The inter-annotator agreement (expressed in percentages, i.e. not accounting for agreement by chance) varies highly among relation types.

Table 1 shows a contrastive study of some coreference annotation schemata and of our approach. In view of this analysis, our approach shows some novelty. It seems to be the first one to experiment with near-identity (introduced by [7]) on a large scale. It is the only one to focus on pointing at dominant expressions and semantic heads. Along with [13], it belongs to two out of three approaches dedicated to Slavic languages which introduce zero subjects in coreference chains.

8 Conclusions and Further Work

We believe that the reported notes on coreference annotation could prove valuable for other coreference corpora creators and the underlying resource is in itself an important step towards general-purpose coreference resolution for Polish.

Table 1. Contrastive analysis of coreference annotation schemes and tools

Reference	**Language**	**Mention scope**					**Semantic Head Markup**	**Relations**	**Dominating Expression Markup**	**Inter-Annotator Agreement**
		Zero Subject	**Nested NGs**	**Split NGs**	**Attributive NGs**	**Discourse Deixis**				
[13]	Bulgarian	✓	✓	?	?	✓		identity, part-of, subset-of		unknown
[14]	English				✓	✓		identity, predicative link, discourse deixis bridging relations		$\alpha = 0.6 - 0.7$
[15]	Czech		partly	✓	partly	✓	?	identity, predicative link, discourse deixis with several sentences, bridging relations		F_1-measure= $0.39 - 0.80$
[16]	Spanish, Catalan	✓	✓	✓	delimited, predicative- linked	✓		identity, predicative link, discourse deixis		$\alpha = 0.85 - 0.89$
[17]	Danish, English, German, Italian, Spanish					✓		identity, discourse deixis bridging relations		$25 - 100\%$
Our approach	Polish	✓	✓	✓	delimited but never linked		✓	identity, near-identity	✓	$\kappa = 0.7424$

We also hope that the current work could help harmonize efforts aimed at creating similar corpora for a group of related languages and, in turn, testing new cross-lingual concepts such as coreference projection (see, e.g., [18]), which require stable and consistent annotation model for all languages involved.

References

1. 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
2. Kopeć, M., Ogrodniczuk, M.: Creating a Coreference Resolution System for Polish. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, ELRA (2012) 192–195
3. Mitkov, R.: Anaphora Resolution. In Mitkov, R., ed.: The Oxford Handbook of Computational Linguistics. Oxford University Press (2003)
4. Recasens, M.: Coreference: Theory, Annotation, Resolution and Evaluation. PhD thesis, Department of Linguistics, University of Barcelona, Barcelona, Spain (2010)
5. Haghighi, A., Klein, D.: Simple coreference resolution with rich syntactic and semantic features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 1152–1161
6. 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
7. Recasens, M., Hovy, E., Martí, M.A.: Identity, non-identity, and near-identity: Addressing the complexity of coreference. Lingua **121**(6) (2011)
8. 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)
9. 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
10. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In Loftsson, H., Rögnvaldsson, E., Helgadóttir, S., eds.: Advances in Natural Language Processing. Volume 6233 of Lecture Notes in Computer Science., Springer (2010) 3–14
11. 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 (to appear)
12. 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
13. 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)
14. Poesio, M., Artstein, R.: Anaphoric Annotation in the ARRAU Corpus. In: Proceedings of the International Conference on Language Resources and Evaluation

(LREC 2008), Marrakech, Morocco, European Language Resources Association (2008)

15. 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
16. Recasens, M., Martí, M.A.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation **44**(4) (2010) 315–345
17. Korzen, I., Buch-Kromann, M.: Anaphoric relations in the Copenhagen Dependency Treebanks. In: Proceedings of DGfS Workshop, Göttingen, Germany (2011) 83–98
18. Rahman, A., Ng, V.: Translation-Based Projection for Multilingual Coreference Resolution. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada, Association for Computational Linguistics (June 2012) 720–730