KPWr: Towards a Free Corpus of Polish

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, Adam Wardyński

Institute of Informatics, Wrocław University of Technology,
{bartosz.broda, michal.marcinczuk, marek.maziarz, adam.radziszewski, adam.wardynski}@pwr.wroc.pl

Abstract

This paper presents our efforts aimed at collecting and annotating a free Polish corpus. The corpus will serve for us as training and testing material for experiments with Machine Learning algorithms. As others may also benefit from the resource, we are going to release it under a Creative Commons licence, which is hoped to remove unnecessary usage restrictions, but also to facilitate reproduction of our experimental results. The corpus is being annotated with various types of linguistic entities: chunks and named entities, selected syntactic and semantic relations, word senses and anaphora. We report on the current state of the project as well as our ultimate goals.

Keywords: corpus, Polish, free, manual annotation, creative commons

1. Introduction

Nowadays, corpora are fundamental resources for language studies and Natural Language Processing (NLP). Many corpora are available for various languages, with more being constructed. Approaches to building a corpus range from gathering the data from the Web with little manual effort, to very costly work on balancing and manual annotation.

In 2010, the SyNaT[1] and NEKST[2] projects were started. One of their goals was to develop Machine Learning (ML) algorithms for shallow syntactic and semantic processing of Polish. Thus, a corpus annotated with various types of linguistic entities was actually a prerequisite. The corpus, called *Korpus Języka Polskiego Politechniki Wrocławskiej* (Polish Corpus of Wrocław University of Technology, *KPWr*), is intended primarily as a training material for ML algorithms, although we want to make it publicly available to serve broader scientific community. We focus on: shallow parsing, word sense disambiguation, named entity recognition, anaphora resolution and automatic generation of metadata. As it takes a lot of effort to construct this kind of resource, we wanted the corpus to be useful for other activities involvinig manual or automatic linguistic analysis.

When we started, there were few freely available corpora for Polish: the IPI PAN Corpus[3] (IPIC) (Przepiórkowski, 2004) and the corpus of Frequency Dictionary of Contemporary Polish (FDCP) (Kurcz et al., 1990). Both were annotated only on the morpho-syntactic level. The National Corpus of Polish (NCP) (Przepiórkowski et al., 2010) was in the middle of construction, with unclear license and unknown release date.

2. Design of the Corpus

The first design decision to make was the source of texts. FDCP had been compiled in the 1960s — does not contain contemporary Polish, and the samples are insufficient for anaphora annotation (50 words on average). We could wait for the NCP, released under GNU GPL, but this is quite unfortunate a licence for a corpus[4]. Thus, we decided to gather texts from scratch. To avoid problems with interpretation of licences, we have settled on texts out of copyright or not subject to copyright laws, or using Creative Commons.

To be a good foundation for ML algorithms, the corpus should be large and contain text from diverse genres. The estimation of required sizes is difficult and still unsolved problem (Bishop, 2006). Thus, we wanted to create as large corpus as possible given the available resources. We have started with a small portion, ca. 50.000 words, and estimated we will be able to construct a corpus of maximum 500.000 words.

In an effort to make KPWr representative, we divided the texts into 14 categories (Tab. 1). The idea was to balance between different variants of Polish: written and spoken, contemporary and old use, general language and technical (scientific) one, official and slang usage (with few other oppositions left).

The problematic and time consuming phase of corpus construction was text acquisition, and especially its clean-up. Ideally, it should be already released on the Web, and we did find a large amount of text in that form, but in some cases we had to turn directly to the authors.

The gathering of text is an ongoing process and some genres are not yet present in the corpus.

To facilitate obtaining rights from authors, while still making anaphora annotation feasible, we settled on 300-word samples. Some of the numbers in Tab. 1, in *Progress* column are above 100% – this is because during the text clean-up parts or even whole documents had to be removed, thus we had to draw more samples than necessary.

The annotation is performed using the Inforex collaborative web editor (Marcińczuk et al., 2012). Due to limited budget, we had to decide against using the standard 2+1 annotation model. After the annotation stage is finished, we plan to draw random samples and have them re-annotated by an

[1] `http://www.synat.pl`

[2] `http://www.ipipan.waw.pl/nekst/`

[3] The usage of the corpus is limited by its somewhat restrictive licence; most notably, it states that "conversion of the binary format of the IPI PAN Corpus or any part thereof to another format is prohibited".

[4] GPL is based on the notion of *source code*, which does not have a clear interpretation in the case of corpora. Numerous questions arise, which we could not answer; most importantly, does it force us to release the tools trained on those corpora under GPL?

Domain	Percentage	Progress
Blogs	10%	101%
Science	10%	0%*
Stenographic recordings	5%	107%
Dialogue	5%	32%
Contemporary prose	10%	0%
Past prose	5%	109%
Law	5%	110%
Long press articles	10%	100%
Short press articles	10%	101%
Popular science and textbooks	5%	110%
Wikipedia	10%	107%
Religion	5%	18%
Official texts	5%	105%
Technical texts	5%	28%

Table 1: Distribution of different genres in KPWr.
*Enough texts are gathered but are not yet sampled and cleaned-up.

additional group of linguists. This way we plan to measure the quality of annotation at lower cost. Should the quality turn out unacceptable, we will schedule additional work to fix as many problems as possible.

3. Annotation Layers

The annotation of KPWr involves the following layers:

1. tokenisation and morphological analysis (mostly automatic),
2. chunking and selected predicate-argument relations,
3. named entities and selected semantic relations between them,
4. anaphora (limited to identity-of-reference type),
5. word senses.

The above selection of employed annotation layers stems largely from the requirements and assumptions of the SyNaT and NEKST projects. For instance, we had to decide against inclusion of manual morphosyntactic annotation to allow allocation of linguistic workforce to critical annotation layers. The tokenisation and morphological analysis was performed using the MACA system (Radziszewski and Śniatowski, 2011) and the new version of Morfeusz analyser (Woliński, 2006). The division into sentences and tokens is not corrected manually except for the rare cases where misplaced sentence boundaries would interfere with manually placed annotations. The morphological information comes directly from Morfeusz SGJP and no disambiguation is performed[5] (i.e. the corpus is not morphosyntactically *tagged*).

In the rest of this section we discuss the remaining layers whose annotation is entirely manual.

3.1. Shallow Syntactic Annotation

We try to make a clear distinction between the following two issues: a declarative description of the structure to be annotated (annotation guidelines) and the procedure that will be used to produce a valid annotation (either manual or automatic). We do not base our definitions of syntactic entities on any rule-based grammar and do not favour any ML techniques. This way we hope to avoid unnecessary bias.

The planned stages of annotation include chunking (understood as recognition of phrase boundaries (Abney, 1991)), marking of chunks' syntactic heads, as well as annotating basic predicate-argument relations. Contrary to the approach presented in (Głowińska and Przepiórkowski, 2010), we prefer a small set of chunks. We focus on nominal and verbal chunks, sometimes joining what is traditionally distinguished as different syntactic groups under simple umbrella terms. This should simplify task formulation for ML and make the resulting parser easier to use. Also note that the syntactic groups (and syntactic words) as defined in the NCP (Głowińska and Przepiórkowski, 2010) are technically not chunks because two groups of the same name may overlap, rendering it hard to use standard chunking techniques.

We use the following set of chunks:

1. AgP, "Agreement Phrases" — simple nominal or adjectival phrases based on morphological agreement on number, gender and case, the building blocks for bigger nominal phrases. This seems to be the closest equivalent of Abney's original proposal, accounting for the difference between Polish and English and still practically useful. For simplicity, we also include phrase-initial prepositions if present, similar to (Grác et al., 2010) for Czech. We also allow adverbial modifiers if they clearly modify other parts of the phrase. E.g.: *bardzo ciekawa propozycja* (*very interesting proposal*), *bez popularnych dziś technologii* (*without the technologies popular nowadays*).

2. NP — possibly complex Noun Phrases that may fill the role of verb arguments. Again, phrase-initial prepositions are included for simplicity. We limit the extent of NPs to clause boundaries, but do not require our phrases to be split on every preposition, which is otherwise a common practice in shallow parsing — sometimes PP attachment decisions must be made to annotate NPs correctly. We are aware that semantic knowledge may be necessary to make a PP attachment decision; nevertheless such decisions must be made to identify verbal arguments. Future experiments with ML will verify if this has been the right move.

3. AdjP, Adjective Phrases — top-level phrases similar to NPs but whose heads are adjectival.

4. VP, Verb Phrases. Since syntactic dependencies between the verb and NPs or PPs will be annotated as predicate-argument structures, our VPs do not include

[5] Again, this is due to legal considerations: the taggers available for Polish rely on manually annotated training corpora, while the only suitable corpus for tagger training in our scenario is the 1-million subcorpus of the NCP. Since it is licensed under GNU GPL, it is unclear if it is legal to release the results of disambiguation under a Creative Commons licence. If the answer turns out positive, we will happily include the tagging as well.

Category/Description	Types
people — names of people, nations, groups of people, etc.	6
toponyms — place names, i.e. names o continents, countries, cities, etc.,	16
urbanonyms — names of roads, districts, squares, etc.	5
hydronyms — names of geographical entities related with water, i.e. names of rivers, lakes, etc.	6
human-made, i.e. names of organizations, facilities, products, books, etc.	21
living — unique names of living objects, i.e. plants and animals	2
astral body — names of planets, stars, constellations, etc.	1

Table 2: Categories of named entities.

those arguments. A typical VP consists of a verbal predicate with possible subordinate verbs (e.g. infinitives) and adverbial adjuncts, e.g. *zaczęli pilnie się uczyć* (*(they) started to learn diligently*).

We define the following relations between chunks:

1. Verb subject, a relation between VP and NP.
2. Verb object, i.e. non-subject verb argument. Note that our NPs may in fact be PPs, such relations are also marked (e.g. *wyjechał do Anglii*, *he moved to England*).
3. *Copula relation* to link the verb with an NP or AdjP in predicative constructs, e.g. *on jest lekarzem* (*he's a doctor*), *to było mądre* (*it was wise*).

We have gathered the general principles of annotation as well as a number of practical guidelines on how to decide the correct annotation in troublesome cases. This document is now being used by the annotators; we plan to make it publicly available as a practical companion to the corpus.

3.2. Named Entities

The scope of named entity (NE) annotation is limited mainly to proper names and names which uniquely identify some categories of entities. We do not annotate numerical expressions (dates, times and other numbers), definite descriptions and noun phrases. All nested annotation are annotated. The annotation schema contains 57 types of named entities which were present in the first samples of KPWr. To organize the named entity types we have grouped them into 7 categories — description of the categories and number of assinged types is presented in Table 2.

3.3. Semantic Relations

On the basis of ACE English Relations Guideline v3 (LDC, 2005) and our named entity schema (see Section 3.2.) we have defined 8 types of relations. Because the level of NE is limited mainly to proper names, we focused on relations between proper names only. We have taken following assumptions: (1) NE connected by a relation must appear within one sentence; and (2) relation must be supported by context, i.e we do not annotate relations that are based only on common knwoledge, for example: (1) "*Poland and Germany are members of the EU*" and (2) "*A pair of towns along the Polish-German border ...*" — only the second sentence we will anotate with a neighbourhood relation, i.e. *neighbourhood(Poland,Germany)*. For every category we defined subtypes, which reflect types of NE that can be connected with given relation type. We have selected following relations: *affiliation*, *alias*, *composition*, *creator*, *location*, *nationality*, *neighbourhood* and *origin*.

3.4. Anaphora

Anaphora is the linguistic phenomenon that occurs when one fragment of a discourse (*anaphor*) relates to another, previously mentioned fragment (*antecedent*).

We have decided to focus, at first, on direct, identity-of-reference relations, where an anaphor and its antecedent have the same referent (are *coreferential*), considering only the cases where either (or both) of them is a proper name (per our NE annotation, which is limited to proper names). This lends itself immediately to tasks of Information Extraction revolving around NEs. Detecting anaphoric relations can bring together various information contained in different fragments of text and link it all to the same entity. Inspirations came from GNOME/MATE project (Poesio, 2004), NP4E project (Hasler et al., 2006) and (Recasens et al., 2007), however we have narrowed the broader aspect of anaphora to focus on NEs. Also, in the lack of full parsing annotation, and for the ease of marking up anaphoras, *markables* (fragments identifying either anaphors or antecedents) are limited just to heads. For zero pronouns, the connected verb (which, in Polish, has gender and number) is treated as the markable. Actual ML approaches will likely require an initial step to consolidate all levels of annotation, including expansion of markables for anaphoras as close as possible to the NP level and dealing with zero pronouns.

In the end, we designated four types of direct identity-of-reference anaphora to annotate:

- NE (as mentioned, proper names only) to NE
- Personal pronoun to NE
- Zero pronoun (associated verb is marked) to NE
- AgP (its head) to NE

3.5. Word Senses

One of the difficult problems in semantic annotation is to assign sense labels to words. The task is conceptually simple: one has to assign a word sense for every occurrence of an ambiguous word in text. The selection of ambiguous words can be limited to some pre-defined dictionary called *sense inventory*. Some of the problems with word-sense annotation arise from difficulties in choosing among fine-grained sense distinctions. Thus, some researches encourage using coarse-grained senses (Hovy et al., 2006). On the other hand, merging of senses is easier than having to divide them later; some researchers even achieved high inter-annotator agreement using fine-grained senses.[6] We settled on using wordnet senses taken from Polish wordnet called plWordNet (Piasecki et al., 2009).

We did not have enough resources to disambiguate every ambiguous word in a corpus. Thus, we use a *lexical sample* approach. We started with nouns as at the time verbal part of plWordNet had not been finalised. The selection of nouns for annotation was data-driven. First, we collected a frequency dictionary from available Polish corpora and the Internet. Then, using plWordNet we divided the nouns according to their polysemy into 12 bins (2, 3...12 senses and more then 13 senses). We used the categories for devising more precise categories: homonyms (H_x), polysemous words (P_y) and mixed ($M_{x,y}$), where x is number of homonymous senses and y is number of polysemous senses. Some of the categories were underrepresented, as not all of the (x, y) combinations could be found among the most frequent nouns. Nevertheless, we achieved good coverage of different polysemy-related phenomena. Later, we used similar methodology for selecting verbs, but the frequency list was taken from the KPWr.

The final list of nouns contains 54 nouns and 30 verbs. We have made a preliminary experiment with annotation of word senses and so far we didn't stumble upon any hard problem. We suspect that during the validation phase more problems will require our attention.

As plWordNet does not contain glosses (yet), we had to prepare definitions of meanings. The definitions are accompanied by the examples taken from corpora. During the work we found some missing senses in plWordNet, which had to be added. This was very laborious part of the work, taking about 100 hours of lexicographer work.

4. State of the Corpus

So far, we have gathered 1 458 documents, making up 402 849 words, i.e. 80.6% of the planned corpus size.

Still ca. 20% of the general language corpus is to be collected. The main problem is copyright law. In order to achieve our goals the texts must be released on the Creative Commons ShareAlike licence. For scientists and modern writers the copyright law poses a difficult problem because of contracts signed with the publishers. We collect mainly unpublished PhD theses and try to gather unpublished texts of young writers. Several catholic priests have decided to support our corpus with newly written sermons.

Although the text-collecting is in progress, we have started to annotate the existing files on almost every level (WSD, anaphora, chunks, named entities and semantic relations). The KPWr corpus is still under intensive development. The detailed statistics are presented in Table 3. The most advanced work has been performed on NE annotation and syntactic chunks. The work on the other levels started later, hence more is still to be done.

Level	**Ann. instances**	**Documents**
Chunks	22 054	155
Named Entities	16 316	732
Word Senses (WSD)	5 911	1129
Level	**Rel. instances**	**Documents**
Chunk relations	4 892	154
Semantic relations	3 092	725
Anaphora	6 101	624

Table 3: KPWr statistics: the number of annotation and relation instances, as well as the documents fully annotated per annotation level.

5. Plans

Text gathering and annotation are in progress. We plan to enrich documents with typical metadata, such as title and publication date, and also general semantic classification, e.g., text genre and keywords. This information will be used for ML experiments, e.g., with automatic topic identification. The metadata will be assigned manually, except for the parts that might be extracted automatically in an unambiguous manner, e.g. source URL.

As already mentioned, the corpus needs evaluation. The planned procedure is to draw random samples, have them re-annotated and estimate the discrepancy at various annotation levels. If the quality of annotation turns out not to meet our standards, we will designate an extra group of linguists to perform the necessary corrections, perhaps using automatic means of detecting potentially erroneous annotations. Last, but not least we will release the corpus on Creative Commons license.

6. Acknowledgements

This work was co-financed by Innovative Economy Programme project POIG.01.01.02-14-013/09 (`http://www.ipipan.waw.pl/nekst/`) and the National Centre for Research and Development (NCBiR) project SP/I/1/77065/10 (`http://www.synat.pl`).

7. References

Steven Abney. 1991. Parsing by chunks. In *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers.

Enko Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

C.M. Bishop. 2006. *Pattern recognition and machine learning*.

Katarzyna Głowińska and Adam Przepiórkowski. 2010. The design of syntactic annotation levels in the National Corpus of Polish. In *LREC 2010 Proceedings*.

Marek Grác, Miloš Jakubíček, and Vojtěch Kovář. 2010. Through low-cost annotation to reliable parsing evaluation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 555–562, Tokio. Waseda University.

[6] For example 85.5% for nouns in Senseval-2 (Agirre and Edmonds, 2006).

Laura Hasler, Constantin Orăsan, and Karin Naumann. 2006. NPs for Events: Experiments in Coreference Annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006)*, pages 1167 – 1172, Genoa, Italy, May, 24 – 26.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60. Association for Computational Linguistics.

I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, and J. Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej*. Wydawnictwo InstytutuJęzyka Polskiego PAN, Cracow.

LDC. 2005. Ace (automatic content extraction) english annotation guidelines for relations. Technical report, Linguistic Data Consortium.

M. Marcińczuk, J. Kocoń, and B. Broda. 2012. Inforex – a web-based tool for text corpus management and semantic annotation. In *The eighth international conference on Language Resources and Evaluation (LREC)*. ELRA.

M. Piasecki, S. Szpakowicz, and B. Broda. 2009. *A wordnet from the ground up*. Oficyna wydawnicza Politechniki Wroclawskiej.

Massimo Poesio. 2004. The mate/gnome proposals for anaphoric annotation, revisited. In *Michael Strube and Candy Sidner (editors), Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science PAS.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. 2010. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Adam Radziszewski and Tomasz Śniatowski. 2011. Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

M. Recasens, M. A. Martí, and M. Taulé. 2007. Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria.

Marcin Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In *Intelligent Information Processing and Web Mining – Proceedings of the International IIS: IIPWM '06 Conference held in Wisła, Poland, June, 2006*, pages 511–520.