Manual annotation of the National Corpus of Polish with Anotatornia[*]

Adam Przepiórkowski and **Grzegorz Murzynowski**

Institute of Computer Science, Polish Academy of Sciences

Abstract

The aim of this paper is to present the procedure of the manual annotation of a 1-million-word subcorpus of the National Corpus of Polish, using a purpose-built tool, Anotatornia. The annotation is carried out at four levels: word-level segmentation, sentence-level segmentation, morphosyntax and word sense disambiguation. Some emphasis is put on the quality control of the resulting annotation: each information is introduced by two annotators, and the system implements a sophisticated procedure of resolving annotation conflicts.

Keywords: manual corpus annotation tool, segmentation, morphosyntactic annotation, word sense disambiguation, annotation quality control, National Corpus of Polish.

1. Introduction

Within the National Corpus of Polish (henceforth, NKJP, as in Pol. *Narodowy Korpus Języka Polskiego*; `http://nkjp.pl/`; Przepiórkowski et al. 2008, 2009), one of the subtasks concerns the manual annotation of a 1-million well-balanced subcorpus of NKJP at various linguistic levels.

The main tool used for this task is Anotatornia, whose early version (Hajnicz et al. 2008) was used in a project concerned with the automatic acquisition of argument structures. Since then, Anotatornia has been extended to new linguistic levels and otherwise adapted to the needs of NKJP, as described below.

The aim of this paper is to describe the process of the manual annotation of NKJP with Anotatornia. § 2 presents the preliminaries, § 3 discusses the levels of linguistic annotation, and § 4 describes the management of annotation conflicts within the system. § 5 concludes the paper.

[*]Research funded in 2007–2010 by a research and development grant from the Polish Ministry of Science and Higher Education.

2. Preliminaries

Each NKJP text contains three general types of annotation: metadata (mainly bibliographical information), structural markup (divisions within texts, e.g., into chapters and paragraphs) and linguistic annotation. The following levels of linguistic annotation are distinguished in the project: 1) segmentation into fine-grained word-level tokens, 2) segmentation into sentences, 3) morphosyntactic analysis, 4) coarse-grained syntactic words (e.g., analytical forms, constructions involving bound words, etc.), 5) named entities, 6) syntactic groups, 7) word senses (for a limited number of ambiguous lexemes).

The first three levels are crucial, as any further linguistic annotation will build on segmentation and morphosyntactic markup. For this reason, any information introduced at these levels must be double-checked. The best practice in such cases is the "2+1" annotation model, where two annotators introduce relevant information independently, and any conflicts are resolved by a referee (or a "superannotator").

Manual annotation is perhaps the most costly task in the development of a corpus, so a tool facilitating this task is needed. It should be ergonomic, it should facilitate work over the Internet, preferably with the use of any browser, and it should handle various linguistic levels. Moreover, it should implement the "2+1" annotation model and have built-in conflict management procedures. No off-the-shelf tool satisfying all project requirements was found at the start of the NKJP project, so it was decided to adapt Anotatornia, a tool whose prototype (Hajnicz et al. 2008) had been developed in an earlier project carried out at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS; the coordinator of NKJP).

The first task of the system is to assign appropriate paragraphs to annotators. This must be done in such a way that each paragraph is annotated by two different linguists. In fact, an annotator at one time fetches a 10-paragraph batch, where 5 of the paragraphs are assigned to another annotator, and another 5 — to a third annotator. The procedure is implemented in such a way, as to avoid "co-learning" of annotation errors; rather, annotators are forced to work with each other in different configurations (virtually all possible), thus increasing the overall consistency of the annotation.

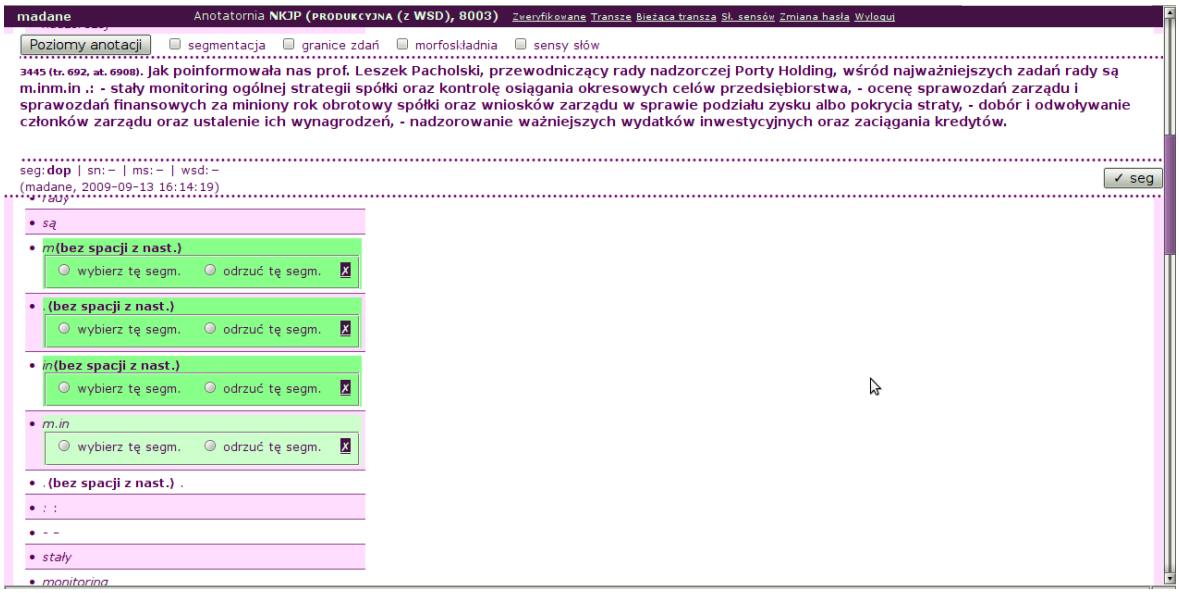

Figure 1: Word-Level Segmentation, annotator's view.

3. Levels of Annotation

Anotatornia handles both segmentation levels, the morphosyntactic level, as well as word sense disambiguation. Work at all these levels is currently, in September 2009, well under way in NKJP. For the annotation of syntactic words, named entities and syntactic groups, a tool will be selected from a number of general syntactic annotation tools available.[1]

The following subsections describe the four annotation tasks and their implementation in Anotatornia, in the order in which the tasks are performed by annotators.

3.1. Word-Level Segmentation

By default, word-level segments (henceforth, simply *segments*) are maximal strings of non-delimiter characters, where delimiter characters are white space characters and punctuation. That is, segments are normally "space-to-space" orthographic words, with typical punctuation characters treated as separate segments. Consequently, to a large extent word-level segmentation can be performed automatically.

One exception concerns the hyphen: sometimes words containing a hyphen are

[1] Originally, the intention was to implement all annotation levels in Anotatornia, but the complexity of the task and the costs of the system development turned out to be prohibitive.

best treated as single segments (e.g., *Mercedes-Benz*), and at other times — as sequences of segments (e.g., *biało-niebiesko-czerwony* 'white-blue-red'). More interestingly, so-called mobile inflections (agglutinates; e.g., *śmy* in *przyszliśmy* 'came-1PL'), the subjunctive particle *by* and the non-accented post-prepositional pronominal forms (e.g., *ń* in *nań* 'for him') are treated as separate segments. This may lead to segmentation ambiguities, as in *gdzieś*, which may be treated as a single adverb meaning 'somewhere' or as a concatenation of two segments: *gdzie* 'where' and *ś* '2SG', with the overall meaning roughly 'where have/did you...?'. Similarly, *długom* is either a dative plural form of the noun DŁUG 'debt', or the adverbial form *długo* 'long' with the agglutinate *m* '1SG' attached; etc.

Such segmentation ambiguities should be resolved at the word-level segmentation. In some cases, segmentation ambiguities are recognised by the morphological analyser used to preprocess texts (cf. § 3.3 below), and the action to be performed by annotators boils down to the selection of an appropriate segmentation variant. Annotators may also propose to split one segment into multiple segments or to join multiple segments into one segment. This is a sensitive operation, as it amounts to postulating the existence of words which, at the same time, are unknown to the morphological analyser and do not follow the default segmentation rules, so it must always be confirmed by a referee (cf. § 4).

3.2. Sentence-Level Segmentation

Text preprocessing does not involve any sentence-level segmentation, as automatic tools may make mistakes where the full stop apparently ending a sentence is possibly a part of an abbreviation, as in the fragment *na ul. Stefana*, which may be a part of *Mieszkał na ul. Stefana Batorego.* 'He lived at the Stefana Batorego street', where *ul.* stands for *ulicy* 'street', or a part of *Zbierał na ul. Stefana to jednak nie obchodziło.* 'He collected (money) for a beehive. But Stefan didn't care.', where *ul* means 'a beehive'.

Even though such cases are rare, there are also texts in NKJP which do not follow standard punctuation (transcripts of speech being the extreme case), and on the other hand the task of sentence-level segmentation is inexpensive compared to other annotation tasks, so it is performed fully manually in the project. Technically, a paragraph is presented to an annotator in such a way that each segment is a separate box, and clicking on a given box

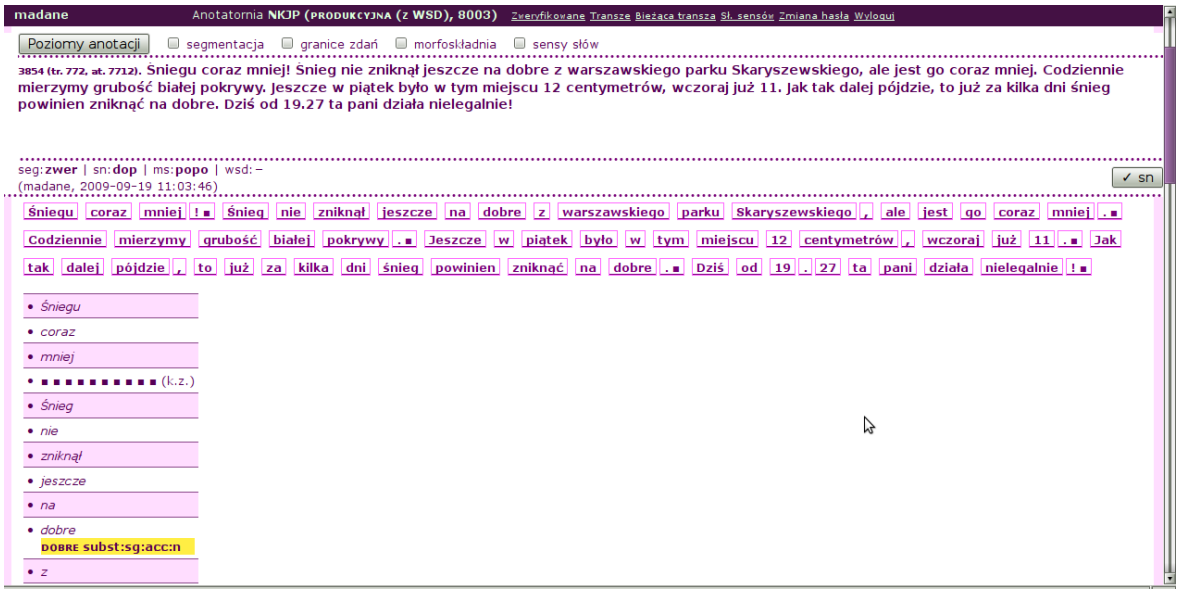

Figure 2: Morphosyntactic annotation, annotator's view. A discrepancy is marked in yellow.

declares the segment as the last segment in a sentence.

3.3. Morphosyntactic Annotation

As mentioned above, texts are preprocessed by a so-called morphological analyser (in fact, little more than an appropriately compressed morphological dictionary), Morfeusz (Woliński 2006). This tool tokenises texts into word-level segments and assigns morphosyntactic interpretations to each segment. Morfeusz does not attempt to disambiguate these interpretations according to context; instead, it marks segments with all interpretations present in its dictionary. The dictionary of the version of the tool used in NKJP, Morfeusz SGJP, is essentially that of Saloni et al. 2007, with subsequent improvements, and the interpretations follow the NKJP Tagset (Przepiórkowski 2009a), a conservative modification of the IPI PAN Tagset (Przepiórkowski & Woliński 2003a,b).

The basic task of annotators is to select the appropriate morphosyntactic interpretation from among those known to Morfeusz. Technically, such interpretations are presented in a column, as radioboxes, and morphosyntactic disambiguation boils down to scrolling the page down and clicking appropriate boxes.

It may happen that the morphological analyser does not know a word or does not

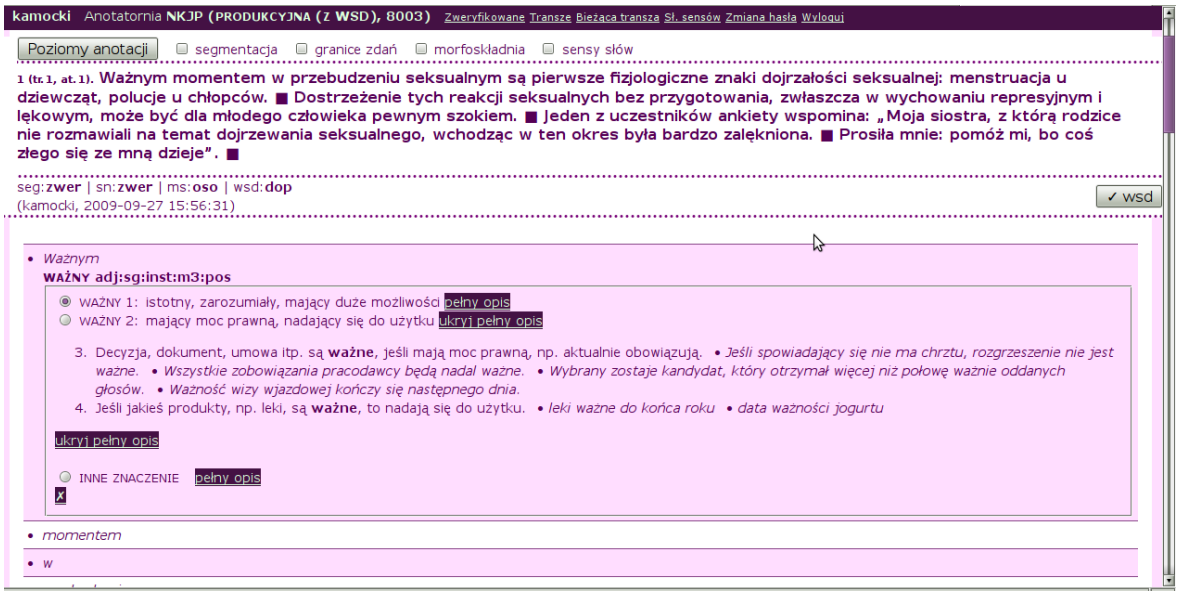

Figure 3: Word Sense Disambiguation, annotator's view, brief description of meaning 1 and full description of meanig 2 are displayed.

contain information about some of its interpretations. Anotatornia makes it possible to add a new interpretation to a segment. On the basis of the formal specification of the NKJP Tagset, Anotatornia suggests completions as the annotator types in a new interpretation, thus facilitating and speeding up the process.

3.4. Word Sense Disambiguation

In NKJP, word sense disambiguation (WSD) is performed only for a little over 100 lexemes, selected from the most frequent clearly ambiguous (strictly homographic, rather than simply polysemous) Polish lexemes.

If a form of one of these ambiguous lexemes is (automatically) detected in a paragraph, this paragraph is marked as due to be annotated at the WSD level. The presentation of the WSD task is fully analogous to that of morphosyntactic annotation: senses are presented in a column, next to vertically aligned radioboxes, so again annotation boils down to scrolling and clicking.

More specifically, brief (usually, 3–7 words) definitions differentiating senses of a lexeme are presented next to radioboxes, with a button next to each such brief definition

for displaying the full dictionary definition of the sense, together with examples. This way annotators may work efficiently relying on the brief definitions, but they may also easily access full definitions in more problematic cases.

Although definitions are based on a relatively comprehensive Polish dictionary (Bańko 2000), situations have arisen where a word is used in a sense not included in the dictionary. For this reason, there is an additional "sense" for each lexeme, for out-of-dictionary uses. However, annotators are asked to make a very sparing use of this artificial "sense".

WSD is dependent on the morphosyntactic level in a subtle way. The word sense dictionary contains verbal, nominal and adjectival lexemes. There are cases where a given word form is morphosyntactically ambiguous between different grammatical classes, and it should be semantically disambiguated only if it is judged to belong to one grammatical class, but not when it belongs to the other. For example, the form *spraw* may be analysed as a form of the ambiguous noun SPRAWA 'matter, case; trial, hearing', or as a form of the verb SPRAWIĆ 'cause, make'. Only the former lexeme is present in the dictionary of ambiguous lexemes, so the form should be subjected to sense disambiguation only if it was assigned a nominal interpretation at the morphosyntactic level. Anotatornia fully implements this feature.

4. Annotation Management

NKJP sets a careful procedure for controlling the quality of annotation. The two annotators working on the same paragraph do not know each other's identity, and any questions regarding the annotation should be asked on a mailing list to which all annotators subscribe. This way the results of a discussion are known to all annotators, whether or not the discussion gives rise to a modification of annotation guidelines. Even when the two annotators can guess each other's identity, e.g., as a result of such a mailing list discussion, they are forbidden to communicate in private.

Any discrepancy in the annotation of two linguists is registered by Anotatornia. In such a case, the annotators are informed about the fact and place of the disagreement, but not about the other linguist's decision. This way, rather than perhaps blindly accepting the annotation of the other linguist, both annotators are forced to think again about the

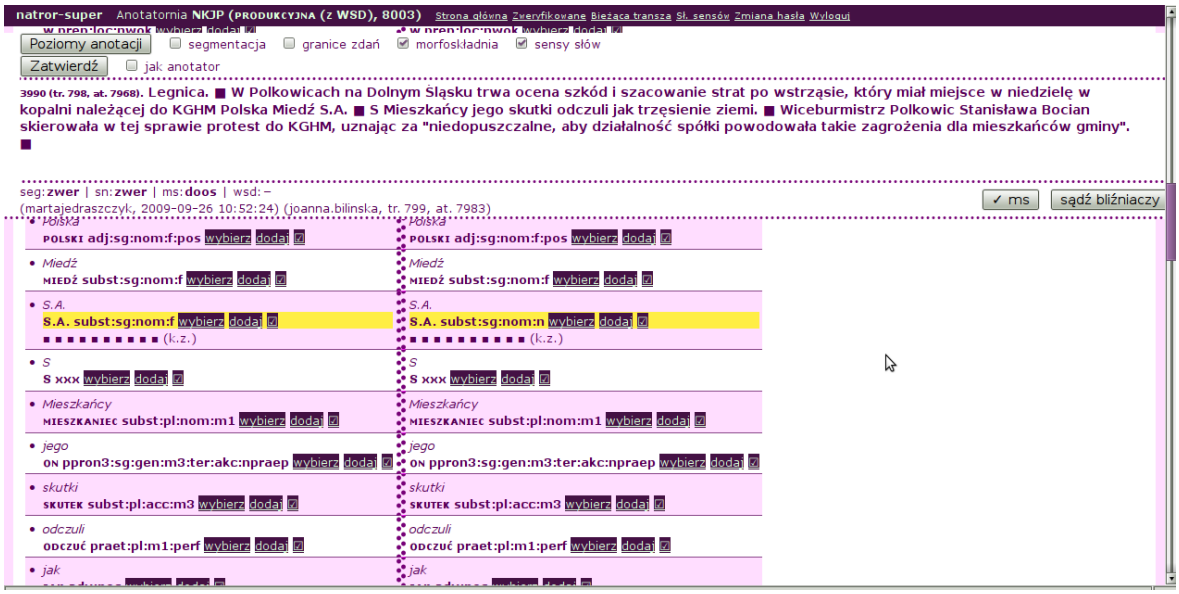

Figure 4: Morphosyntactic annotation, referee's view. A discrepancy is marked in yellow.

annotation adduced earlier, and either explicitly confirm it or change it.

In many cases such differences result from simple inattention and are easy to correct. However, when — after the step described in the previous paragraph — a discrepancy still exists, the paragraph is presented to a referee, who makes the final decision. Annotators have read-only access to previously annotated paragraphs, and they can see the results of this refereeing process. Moreover, if the disagreement reflected a genuine weakness of the annotation guidelines, and if a general and precise solution may be formulated, it is added to the guidelines.

5. Conclusion

Anotatornia is perhaps unique among the tools for the manual annotation of corpora in that it implements careful procedures of assigning texts to annotators, controlling the consistency of their annotations, and resolving any conflicts. It is currently employed and fine-tuned in the annotation of the National Corpus of Polish. Once the code is relatively stable, it will be published as open source at `http://nlp.ipipan.waw.pl/Anotatornia/`.

Bibliography

Bańko, M. (Ed.) (2000) *Inny słownik języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Hajnicz, E., Murzynowski, G. & M. Woliński. (2008) 'ANOTATORNIA – lingwistyczna baza danych'. In: *Materiały V konferencji naukowej InfoBazy 2008, systemy * aplikacje * usługi*. Gdańsk: Centrum Informatyczne TASK, Politechnika Gdańska, 168–173.

Przepiórkowski, A. (2009a) 'A comparison of two morphosyntactic tagsets of Polish'. In: Koseska-Toszewa, V., Dimitrova, L. & R. Roszko (Eds.), *Representing semantics in digital lexicography: Proceedings of MONDILEX fourth open workshop*. Warsaw, 138–144.

———. (2009b) 'Zasady znakowania morfosyntaktycznego w NKJP'. Unpublished manuscript, ICS PAS. Version 1.23 of 27 September 2009.

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B. & M. Łaziński. (2008) 'Towards the National Corpus of Polish'. In: *Proceedings of the sixth international Conference on Language Resources and Evaluation, LREC 2008*. Marrakech.

Przepiórkowski, A., Górski, R. L., Łaziński, M. & P. Pęzik. (2009) 'Recent developments in the National Corpus of Polish'. In: Levická, J. & R. Garabík (Eds.), *Proceedings of Slovko 2009: Fifth international conference on NLP, corpus linguistics, corpus based grammar research, 25–27 November 2009, Smolenice/Bratislava, Slovakia*. Brno: Tribun.

Przepiórkowski, A. & M. Woliński. (2003a) 'A flexemic tagset for Polish'. In: *Proceedings of Morphological processing of Slavic languages, EACL 2003*. Budapest, 33–40.

———. (2003b) 'The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish'. In: *Proceedings of the 4th international workshop on linguistically interpreted corpora (LINC-03), EACL 2003*, 109–116.

Saloni, Z., Gruszczyński, W., Woliński, M. & R. Wołosz. (2007) *Słownik gramatyczny języka polskiego*. Warsaw: Wiedza Powszechna.

Woliński, M. (2006) 'Morfeusz — a practical tool for the morphological analysis of Polish'. In: Kłopotek, M. A., Wierzchoń, S. T. & K. Trojanowski (Eds.), *Intelligent information processing and web mining*. Berlin: Springer-Verlag, 511–520.