Anotatornia 2 – an annotation tool geared towards historical corpora

Marcin Woliński, Witold Kieraś, Dorota Komosińska

Institute of Computer Science, Polish Academy of Sciences

Abstract

In the paper, we present a tool built to annotate historical corpora of inflected languages. Historical corpora pose a problem not seen in contemporary ones, namely the need to work with texts represented in two parallel forms: transliterated and transcribed. Besides a typical mode of operation where decisions of two annotators are confronted, the tool implements a mode where a single annotator is confronted with a tagger. Anotatornia 2 has been deployed for annotation of three rather different corpora of Polish: a contemporary one, a corpus of 19th century texts, and a corpus of Baroque texts (17th and 18th cent.).

1. Introduction

The direct incentive for building Anotatornia 2 was provided by the need to annotate inflected forms in a corpus containing 17th and 18th century Polish texts (Bronikowska et al., 2016). Compared to contemporary texts, processing historical texts is much more laborious, since the text can be in fact considered a foreign language with no native speakers available. Annotators can to much lesser extent rely on their linguistic competence. This emphasises the need for a comfortable working environment for annotation.

Moreover, historical texts exhibit much more orthographic variation. For example the word *komisja* ('commission') can appear in 19th century Polish texts in the following spellings: *komisja, kommisja, komissja, kommissja, komisya, kommisya, komissya, kommissya*. The general rule of historical corpora is to key in the text in a form as close to the original as possible (transliteration of the original). Then the variation can be coped with by means of (automatic) transcription (Bronikowska et al., 2016). The text is transcribed to a modernised spelling (so all the above words are represented as *komisja*). However, this means that processing tools need to operate on the text represented in this parallel form. Automatic morphological analysis is performed on the transcribed version (which simplifies the creation of a morphological analyser; Kieraś et al., 2017), but the results must remain coupled with the transliterated form, since this form has to be shown to human annotators. To the best of our knowledge, no readily available annotation tool works with data in such a form.

One more requirement typical of historical corpora is to track the number of page each token appears on in the printed original. This element of text structure was not needed in contemporary National Corpus of Polish but it was crucial for the Baroque corpus. To make things more difficult, page divisions cross all other levels of text structure (in particular division into sentences and paragraphs).

Morphological annotation of the National Corpus of Polish, NKJP (Przepiórkowski et al., 2012) was done with a Ruby on Rails application named Anotatornia (Hajnicz et al., 2008; Przepiórkowski and Murzynowski, 2011). An obvious decision would seem to be to improve that tool. Unfortunately, due to fast ageing of toolkits used to build web applications it turned out very difficult to deploy old Anotatornia in a new project. Also no other ready to use application satisfying requirements of historical corpora annotation was available. Thus we have decided to build a new tool for this task.

The work being reported was financed by a National Science Centre, Poland grant DEC-2014/15/B/HS2/03119.

2. Requirements for the tool

Manually validated morphological annotation of a corpus requires much human labour, which means that usually it is performed by a group of annotators working simultaneously. An obvious choice for this type of work is the use of a Web application.

In the present project, the annotation tool is required to facilitate the following tasks: tokenisation, sentence boundary determination, morphological analysis with disambiguation and validation. It is assumed that tokenisation and sentence detection are performed by automatic tools, which may make errors. The text is then passed to an automatic morphological analyser, which provides all possible interpretations of given words. Annotators will be required to validate tokenisation and sentence boundaries, and then to disambiguate inflectional tags and to provide interpretations of words unknown to the analyser.

It is assumed that the corpus consists of samples of a (more or less) fixed length. In corpora at hand the samples are contiguous pieces of text of about 200 words. The sample is a unit of work for an annotator.

In original Anotatornia tokenisation, sentence determination, and morphological disambiguation were three separate stages of processing (Przepiórkowski et al., 2012, §6.6), which meant that each sample was analysed by an annotator 3 times and problems spotted in a "wrong" phase could not be corrected immediately (for example it was not possible to correct an inflectional tag in the phase devoted to tokenisation). After a thorough discussion we have decided that the new tool will allow to perform all these actions in one processing phase.

Moreover, changes in tokenisation in Anotatornia required an intervention of the arbitrator (see section 3.), which resulted in lags caused by the need to pass the sample from annotator to arbitrator and back. Since we expected historical texts to have more problems with tokenisation, we have decided to avoid this lag and allow annotators to modify tokenisation directly.

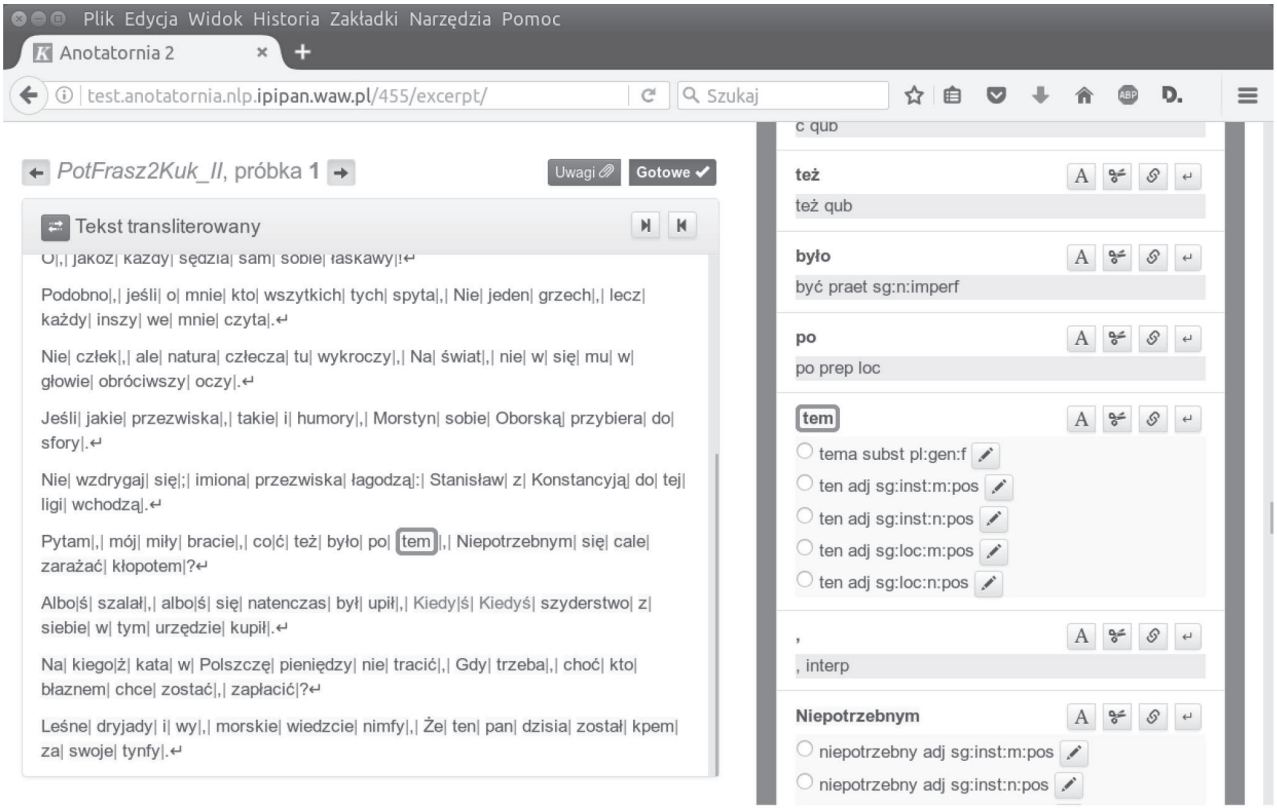

Figure 1: A sample being annotated as seen by an annotator

3. Annotation modes

The established best practice in manual corpus annotation is a procedure where each corpus sample is annotated independently by two annotators and conflicts solved by an arbitrator ("super-annotator"). This approach was in particular followed in the annotation of NKJP (Przepiórkowski et al., 2012, §6) and we also treat it as the basic mode for our tool (AA+A mode).

In more detail, the procedure includes an intermediate phase: when conflicts are detected, the sample is shown once again to both annotators. The conflicting tokens are highlighted (cf. Fig. 2), but only the user's own decisions are shown. This way annotators are encouraged to check their work for simple errors but are not tempted to switch to other annotator's version. When allotting samples to annotators, the system maximises the number of different pairs of annotators working in parallel. This is to minimise the biases introduced to annotation by particular annotators. The sample is passed to the arbitrator only if any conflicts remain after this additional stage.

The arbitrator can select from the variants provided by annotators or provide her own answer. She has also the right to intervene into description of any tokens in the sample, also those without conflicts.

This mode of operation is believed to provide high annotation quality thanks to the low(er) probability of two annotators making an error at the same spot. The obvious drawback is that the corpus has in fact to be annotated twice, which means more time and money.

In this paper we propose and test an annotation mode that alleviates this drawback. Since automatic taggers can provide results of quality similar to that of manual annotation, it is tempting to replace one annotator of the pair with an automatic tool. A natural question arises whether this results in more conflicts to be addressed by the arbitrator.

In this mode of Anotatornia 2 (AT+A mode) each sample is given to one annotator only and the results of an automatic tagger are stored in the system playing the role of the second annotator. As soon as the annotator finishes her work on a sample, resulting conflicts are shown, just as in the case of two annotators. Conflicts remaining after this phase are passed to the arbitrator.

These modes of operation are compared in section 6.

4. Anotatornia 2

Anotatornia 2 is an application built using Django toolkit. It allows annotators to interact with corpus samples in a Web browser, see Fig. 1.

The left panel of the program displays a corpus sample in the transliterated form (the view can also be changed to transcribed form). A sample is assumed to consist of a few paragraphs (or parts thereof), which consist of sentences, which in turn consist of tokens. Each token has its text provided in the transliterated and transcribed form. Interpretation of a token consists of a lemma and a morphosyntactic tag. These elements are visible and can be interacted with in the right panel.

The annotator is responsible for several tasks, which we describe below:

Tokenisation The main point in validating tokenisation is to resolve situations where tokenisation is ambiguous. Such cases are not very common in contemporary texts, main type is the ambiguity of words like *gdzieś*, which can be interpreted as a whole (meaning 'somewhere') or split into *gdzie* 'where' and *ś* being an auxiliary form of the verb *być* 'to be' (Przepiórkowski and Woliński, 2003).

In older texts tokenisation poses more of a problem since orthographic rules were much less fixed. In particular in Baroque Polish prepositions are often spelled together with a succeeding noun. Morphological analyser is able to handle some of such cases, but others have to be done by hand. Moreover, this mechanism can cause ambiguity in tokenisation (Kieraś et al., 2017).

The annotator can use two operations to change tokens. The first is splitting a given token into several new ones. The other operation applied to a token joins it with the following one. The annotator will have to add morphological annotation for new tokens produced in both ways.

Validation of transcription Anotatornia 2 in its left panel displays the text in the transliterated form, while the list on the right shows the transcribed version. The annotator should check whether the two forms correspond with each other. If not, the transcribed form can be changed by hand.

Sentence boundary determination Sentence boundaries are represented in Anotatornia 2 with a flag signalling that the given token is the last of a sentence (obviously usually this token is a period). The annotator can add or remove this flag from any tokens.

If a sample starts or ends in the middle of an incorrectly determined sentence, such part of a sentence is excluded from annotation.

Morphological disambiguation The most complex task of the annotator is to validate and complete inflectional interpretations. Anotatornia displays all interpretations generated by the morphological analyser allowing the annotator to choose one of them with a single click (see Fig 1).

If no interpretation is correct, the annotator can add a new one. This includes giving a lemma and a morphological tag. The tagset used can be configured for a given corpus. The system actively completes tags being entered by the annotator and checks if the resulting tag is correct with respect to the tagset.

Resolving conflicts All the tasks mentioned above can cause conflicts. A finished sample with conflicts is passed to the arbitrator, who is shown a list of differences between annotations (see Fig. 3). As seen in the picture, each conflicting token is shown three times on the list, first two corresponding to the choices of annotators, the third one providing the arbitrator with the possibility to select yet another answer.

The arbitrator can resolve a conflict either by selecting one of the provided solutions or building a new one exactly in the way an annotator does an analogous task. Moreover, the arbitrator has access to all elements of annotation of all tokens and can change any decision made by the annotators even for tokens without a conflict.

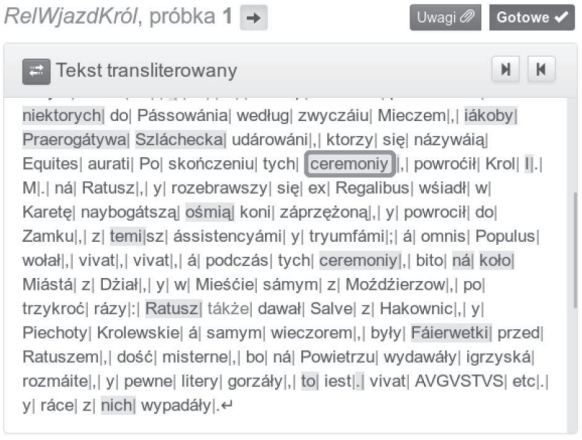

Figure 2: Conflicts in annotation as seen by an annotator (left panel of the interface)

5. Deployment

Anotatornia 2 has been already used for annotation of three corpora. The corpora represent three different periods of historical development of Polish and the annotating teams make use of different features of the application.

5.1. Korba

Chronologically the first task to be taken in Anotatornia 2 was the annotation of a 500,000 tokens large corpus of Baroque Polish. The process is still ongoing. The annotated corpus is extracted from a much larger collection of texts representing the Polish language of 17th & 18th century (until 1772).

The original texts were manually transliterated and automatically transcribed to modernized spelling using a rule-based method, automatically tokenized and morphologically analysed using a modified version of Morfeusz analyser (Woliński, 2014) with adjusted and enriched inflectional data. The full process of preparing text samples for manual annotation was described in detail in (Kieraś et al., 2017). The manually annotated corpus will serve as training data for a stochastic tagger with which then a larger (ca. 12 mln tokens) corpus will be annotated automatically and made publicly available.

The corpus is being annotated in the AA+A mode. The task is challenging for the annotators as Baroque Polish is considerably distant from contemporary language both lexically and grammatically. The ongoing project deals with so far the oldest Polish texts subject to systematic and extensive morphosyntactic annotation.

5.2. 1830-1918

Another historical corpus that is being annotated using Anotatornia is a 500,000 tokens large collection of samples excerpted from texts published between 1830 and 1918 (second half of the so called New Polish period). The original corpus (Bilińska et al., 2016) is twice as large and consists of 1000 samples, ca. 1000 words each. Every sample represents one of the five functional styles: science and popular science, short newspaper articles, essays, fiction and drama. The samples were distributed evenly be-

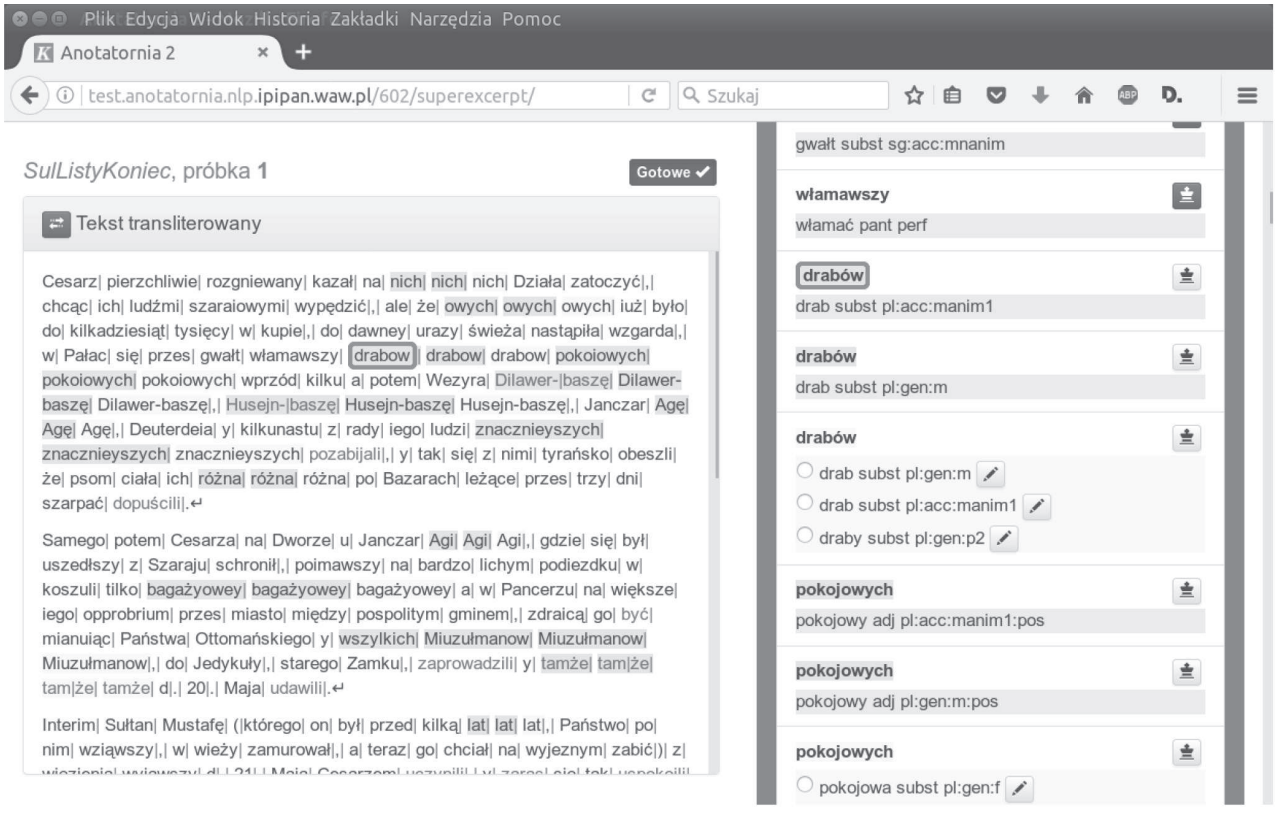

Figure 3: A sample with conflicts as seen by an arbitrator

tween the styles. All the corpus samples were extracted from scans of original printed editions and transliterated into plain text digital files. For the purpose of manual annotation a set of over 3,100 shorter samples (ca. 160 words each) was excerpted from the original corpus. The samples were automatically transcribed into modernized spelling using the same tools as in the case of BCP and morphologically analysed with a variant of Morfeusz in which the dictionary was modified to adhere to 19th inflection.

The corpus is being annotated in the AT+A mode. The tagger used in the task is Concraft (Waszczuk, 2012). Samples are added to Anotatornia in relatively small batches, so the tools supporting the annotation process could be updated constantly. In the beginning a version of manually annotated subcorpus of NKJP was used as tagger's training data. The subcorpus was automatically converted (to the possible extent) to adhere to the task's tagset and annotation guidelines. During the annotation process the tagger's model is periodically, incrementally retrained with newly annotated data to improve its performance in the further process of annotation. The same applies to the analyser which is being updated constantly based on annotators' feedback.

5.3. Poleval

Finally, a small corpus of text samples was manually annotated in Anotatornia for the purpose of evaluating tagging systems competing in the Poleval contest.[1] The corpus is ca. 55,000 tokens large and consists of text samples extracted from the corpus of coreferences (Ogrodniczuk et al., 2015).

Each Poleval corpus sample was annotated by human annotator and Concraft tagger in AT+A mode. The tagger was trained on the manually annotated subcorpus of the NKJP. For the purpose of automated tagging a version of Morfeusz 2 was built to adjust its dictionary to NKJP tagset.

6. Evaluation

Table 1 shows the number of tokens in parts of respective corpora that have already been annotated (in case of Poleval it is the whole corpus). The percentage of words absent from respective morphological dictionaries used for each corpus is highest in the corpus of oldest texts, which was expected.

The last three rows of Table 1 show the number of times an annotator had to add a piece of information (as opposed to selecting from automatically provided options). Tags

	Korba	**19c**	**Poleval**
tokens	198,601	126,894	55,448
unknown to Morfeusz	6.44%	1.72%	1.54%
	added interpretations:		
transcription	2.14%	0.33%	0.10%
tokenisation	1.31%	0.42%	0.25%
tags	6.92%	2.80%	3.88%

Table 1: Corpus sizes and interpretations added by annotators

[1] see `http://poleval.pl/`

	Korba	19c	Poleval
annotation mode	AA+A	AT+A	AT+A
conflicts	9.21%	14.51%	11.34%
	who was right:		
one of annotators	85.90%	89.61%	86.78%
tagger	—	4.90%	9.42%
arbitrator's own answer	14.10%	5.50%	3.80%
arbitrator without conflict	3.96%	2.66%	0.75%

Table 2: Conflicts in annotation

had to be added for all tokens unknown for the analyser, and indeed all numbers in the last row are slightly larger than the numbers in the second row. It is a bit startling that the difference is largest for the Poleval corpus – it seems that its annotators were most eager to intervene in the tags. The number of manually delimited tokens is largest for Korba, which is expected due to orthographic features of Baroque texts. The number of changed token transcriptions in Poleval is non-zero, because this was used as a way to correct evident typos in the text (in this corpus only).

Table 2 shows data about conflicts in the annotation. The first row reports the number of conflicts as percentage of the whole corpus. As can be seen, the number of conflicts in corpora where the AT+A mode is used is higher than in the AA+A corpus. This is probably due to the tagger making more errors than a human annotator. However, even in the worst case of 19th century text tagged with a tagger trained on contemporary data 14.51% of conflicts is manageable – the workload for the arbitrator is larger but in reasonable boundaries and it clearly outweighs the cost of second manual annotation.

The next rows show how the conflicts were resolved by the arbitrator (in percentages of conflicts). The solution provided by one of the annotators was accepted by the arbitrator in a similar number of cases in Korba and Poleval. This number is higher for 19c probably because of the tagger making more simple errors in this corpus. However, the next row shows that even in this corpus it happened that the version provided by the tagger was selected over the version of the annotator. This proves that a confrontation even with a weak tagger is a good means of catching human errors.

The last row shows the number of times the arbitrator has changed the decision of annotators (or an annotator and the tagger) even though they provided the same answer. Differences in this row seem to be caused by Korba text being difficult and they support the claim that the AT+A mode does not impair the annotation process.

7. Conclusions

A new tool for morphological annotation of corpora has been presented. The tool caters for the specific needs of processing historical texts such as the dual form of text (transliteration/transcription).

The innovative "annotator against tagger" mode provides results of similar quality with about half of workload as the typical mode of two independent annotators.

The tool uses TEI XML modelled on NKJP as its input and the output, so it can be easily used for other corpora. No element of the tool is language dependent.

We believe that the tool provides an ergonomic working environment, with clean visualisation of morphological interpretations. Although implemented as a web application, the tool is very reactive. Simple decisions are preformed with a single click and a tag validator is activated when an interpretation has to be added by hand.

Anotatornia 2 is publicly available under the terms of an open source license and can be found at its website: `http://zil.ipipan.waw.pl/Anotatornia2`.

8. References

Bilińska, Joanna, Magdalena Derwojedowa, Witold Kieraś, and Monika Kwiecień, 2016. Mikrokorpus polszczyzny 1830-1918. *Komunikacja specjalistyczna*, 11:149–161.

Bronikowska, Renata, Włodzimierz Gruszczyński, Maciej Ogrodniczuk, and Marcin Woliński, 2016. The Use of Electronic Historical Dictionary Data in Corpus Design. *Studies in Polish Linguistics*, 11(2):47–56.

Hajnicz, Elżbieta, Grzegorz Murzynowski, and Marcin Woliński, 2008. ANOTATORNIA – lingwistyczna baza danych. In *Materiały V konferencji naukowej InfoBazy 2008, Systemy * Aplikacje * Usługi*. Sopot: Centrum Informatyczne TASK, Politechnika Gdańska.

Kieraś, Witold, Dorota Komosińska, Emanuel Modrzejewski, and Marcin Woliński, 2017. Morphosyntactic annotation of historical texts. the making of the baroque corpus of polish. In *International Conference on Text, Speech, and Dialogue*. Springer, Cham.

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska, 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (eds.), 2012. *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Przepiórkowski, Adam and Grzegorz Murzynowski, 2011. Manual annotation of the National Corpus of Polish with Anotatornia. In Stanisław Goźdź-Roszkowski (ed.), *Explorations across Languages and Corpora: PALC 2009*. Frankfurt am Main: Peter Lang.

Przepiórkowski, Adam and Marcin Woliński, 2003. 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*.

Waszczuk, Jakub, 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India.

Woliński, Marcin, 2014. Morfeusz reloaded. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*. Reykjavík, Iceland: ELRA.