Annotation of Parallel Corpora (on the Example of the Bulgarian–Polish Parallel Corpus)[1]

L. Dimitrova[1], *V. Koseska–Toszewa*[2], *I. Derzhanski*[1], *R. Roszko*[2],
[1] IMI-BAS, *ludmila@cc.bas.bg*, [2]ISS-PAS

Abstract
In this paper we briefly describe a comparison of the morphosyntactic characteristics of the words of the first Bulgarian-Polish parallel corpus from the point of view of a prospective unification.

Keywords: *Bulgarian, Polish, parallel corpus, corpus annotation, morphosyntactic description, POS tagging*

1. Introduction

Corpus linguistics is a dynamic field which boasts many accomplishments in recent years. Among them are the MULTEXT corpus (Ide, Veronis, 1994), initially in seven West European languages (Dutch, English, French, German, Italian, Spanish and Swedish, with more in later editions, including Bambara, Catalan, Kikongo, Occitan and Swahili), and the MULTEXT-East annotated parallel corpus (Dimitrova et al., 1998), initially in six East European languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian, plus English as a "hub" language, in later editions including Croatian, Lithuanian, Resian[2], Russian and Serbian). MULTEXT-East is an extension of the language engineering project MULTEXT, one of the largest EU projects in the domain of language tools and resources.

The first Bulgarian–Polish corpus (currently under development in the framework of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI—BAS and ISS—PAS, coordinated by L. Dimitrova and V. Koseska) contains a total of approx. 3 million words and comprises two corpora: parallel and comparable (Dimitrova, Koseska, 2007, 2008). The first Bulgarian–Polish parallel corpus contains more than 1 million words, mostly fiction (a small part comprises official documents of the European Commission available through the Internet). The corpus is composed of two parts: original Bulgarian texts with Polish translations or vice versa and texts in other languages translated into both Bulgarian and Polish. The comparable corpus includes texts in Bulgarian and Polish, excerpts from newspapers, literary works, Internet textual documents, with the text sizes being comparable across the two languages. Some of the texts have been annotated at paragraph level. The bilingual Bulgarian–Polish corpus will be annotated according to the digital language resource annotation standards and will provide a sample of the vocabulary, which is to be included in an initial experimental version of the Bulgarian–Polish digital dictionary.

We endeavoured to perform a comparison of the morphosyntactic characteristics of the words of parallel texts in the two languages from the point of view of a prospective unification.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement 211938 MONDILEX

[2] Resian is a distinct dialect of Slovenian spoken in the valley Resia in Italy, close to the border with Slovenia. Resian and standard Slovenian are mutually unintelligible due to archaisms not preserved in modern Slovenian and significant Italian influence on Resian pronunciation and vocabulary, as well as Italian-induced innovations in Resian grammar (including prepositive definite and indefinite articles).

2. Corpus annotation

Corpus annotation is the process of adding linguistic information in an electronic form to a text corpus (Ide et al. 2000, Leech 2004, Monachini, Calzolari, 1996). Among the most common and important types of corpus annotation are **morphosyntactic annotation** (also called **grammatical tagging** or *part of speech* (*POS*) *tagging*), whereby a label or tag is associated with each word token in the text in order to indicate its grammatical classification, and *lemma annotation*, where the lemma of each word-token is indicated in the text. These two types may be regarded as mutually complementary.

POS tagging is the task of labelling each word in a sequence of words with its appropriate part-of-speech. Words are often ambiguous with respect to their POS; for example, in Bulgarian the neuter singular forms of most adjectives serve double duty as adverbs.

вероятно 'probable (neuter), probably'

вероятно → POS: adjective, Gender: neuter, Number: singular,

Definiteness: no

вероятно → POS: adverb, Type: adjectival

A tagset is a set of part-of-speech tags. The size and choice of the tagsets vary across languages. The classical system is based on a set of parts of speech including noun, verb, adjective, pronoun, adverb, numeral, preposition, conjunction, particle, interjection, and often (depending on the language) article, participle, etc. Morphologically rich languages need more detailed tagsets reflecting various inflexional categories.

The applications of POS tagging include lexicography, parsing, language models in speech recognition, disambiguation clues for ambiguous words (machine translation), information retrieval, spelling correction, etc.

3. Morphosyntactic descriptions for Bulgarian

For the purposes of morpho-lexical processing of corpora, the MULTEXT-East consortium developed language-specific word-form lexical lists covering at least the words appearing in this corpus. For each of the six MTE languages, a lexical list containing at least 15,000 lemmata were developed for use with the morphological analyser. Each lexicon entry includes information about the inflected-form, lemma, POS, and morphological specifications. A mapping from the morphosyntactic information contained in the lexicon to a set of corpus tags (used by the part-of-speech disambiguator) was also provided, according to the MULTEXT tagging model.

A lexicon entry has the following structure:

word-form ‹TAB› **lemma** ‹TAB› MSD ‹TAB› **comments**

where word-form represents an inflected form of the lemma, characterised by a combination of feature values encoded by *MSD*-code (**MSD**: **M**orpho**S**yntactic **D**escription); the fourth (optional) column, comments, is currently ignored and may contain either comments or information processable by other tools.

Here is an excerpt from the Bulgarian Lexicon:

Word-Form	*Lemma*	*MSD*
вещества	вещество	Ncnp-n
веществата	вещество	Ncnp-y
вещество	=	Ncns-n
веществото	вещество	Ncns-y

(*вещество* 'substance')

The **MSDs** are provided as strings, using a linear encoding; a relatively efficient and compact way to represent the flat attribute-value matrices. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way: the character at position 0 encodes part-of-speech; each character at position 1, 2, …, n, encodes the value of one attribute (person, gender, number, etc.), using the one-character code; if an attribute does not apply, the corresponding position in the string contains the special marker "-" (hyphen). By convention, trailing hyphens are not included in the lexical MSDs. Such specifications provide a simple and relatively compact encoding, and are similar to feature-structure encoding used in unification-based grammar formalisms. When the word form is the very lemma, then the equal sign is written in the lemma field of the entry ("=").

For Bulgarian morphosyntactic annotation was implemented in 1996–1997 for the purposes of the MULTEXT-East project. The morphosyntactic descriptions were designed on the basis of the traditional part-of-speech classification (Bulgarian Grammar 1993). Each word form is assigned a label encoding the major category (part of speech), type where applicable (e.g., proper versus common noun) and inflexional features. Punctuation is also included, as are abbreviations, numbers written in digits, and unidentified objects (residuals). A further non-standard category contains markers of degrees of comparison. Those are formed in Bulgarian with the particles *по* (comparative) and *най* (superlative), preposed to the adjective or adverb but separated from it by a hyphen (*лек* 'light', *по-лек* 'lighter', *най-лек* 'lightest'; *леко* 'lightly', *по-леко* 'more lightly', *най-леко* 'most lightly'). These particles are annotated as separate words:

по → POS: Particle, Type: comparative, Formation: simple,

най → POS: Particle, Type: superlative, Formation: simple.

4. The morphosyntactic descriptions for Polish

For Polish morphosyntactic analysis is performed by Marcin Woliński's Morfeusz (Woliński 2003, 2006). This analyser is based on an extended set of parts of speech (15 in total), so that groups of words traditionally counted under the same part of speech (and even parts of the same paradigm) are separated if they differ significantly in their inflexional categories or syntactic meaning. It constructs all possible analyses of each word and singles out one analysis as the most likely one (a suggestion that the user is free to endorse or decline).

The analyser has the shortcoming that it does not cover the bound clitic forms of the copula *-(e)m* 'I am', *-(e)ś* 'you are', etc., except when the host of the clitic is a past tense verb form (giving regular past tense forms inflected for person and number, though represented as two-word sequences here, which shows that the treatment of groups of words unseparated by blank space or punctuation is not a problem in principle). For example, the sentence *Coś zrobił?* is only analysed as 'Did he do something?' (*coś* 'something, anything' + *zrobił* '[he] did'), missing the alternative meaning 'What did you do?' (*co* 'what' + *-(e)ś* 'you' + *zrobił* 'did'), also expressible as *Co zrobiłeś?* (*co* 'what' + *zrobił* 'did' + *-(e)ś* 'you', analysed correctly by Morfeusz).

One further possibly questionable point is the treatment of gender. The category of animacy is unusually ramified in Polish, so that three varieties of the masculine gender are counted (human, animal and inanimate[1]). The analyser treats these as three separate genders (of a total of five in the language, according to Saloni's simplified version[2]), which gives rise to a proliferation of possible analyses due to the massive syncretism of gender in all parts of speech inflecting for this category (adjectives, numerals, pronouns, quasiparticiples and participles).

For example, the verb form (here called a quasiparticiple) *był* '(he) was' is assigned three genders (the three masculines); *były* '(they) were' is assigned four (masculine animal, masculine inanimate, feminine and neuter). Personal pronouns for the first and second persons are also assigned all possible genders, the most likely one being chosen on semantic grounds if possible. Thus the pronoun *ja* 'I' in *Ja przyszedłem* 'I came (m.)' is proclaimed most likely masculine human, although the other four analyses are also generated, because the verb form is in the masculine; in *Ja idę* 'I go' the same pronoun, now with a gender-neutral verb, is labelled as most likely feminine, perhaps because it happens to end in *-a*. This adds to the complexity of the analysis.

5. The experiment

We took two short stories for children ('The Gluttonous Little Bear' by Emilian Stanev and 'Soap Bubbles' by Svetoslav Minkov) in the original Bulgarian and in V. Koseska-Toszewa's Polish translation (just under 1000 words).

The translation is literary rather than literal. Some frequently recurring differences are due to the stylistic preferences characteristic of the two languages: Polish makes active use of constructions with participles and gerunds, which Bulgarian also possesses but employs significantly less, especially in informal speech and writing, preferring constructions with finite verb forms. Of course this is just a general tendency, and in individual sentences the correspondences may be of any complexity:

'"Great that I broke off from that vulgar straw!" said the first soap bubble, flushed with joy and floated above the bed of daisies.'

[1] Words denoting animals behave as human in the singular number and as inanimate in the plural.

[2] Different accounts distinguish between three and nine (or more) genders in Polish.

Bulgarian:

– Добре, че се откъснах от тая проста сламка — рече първият, като почервеня от радост и се понесе над лехата с маргаритките.

(*flushed and floated are coordinated and both are subordinated to said*)

Polish:

– Doskonale! oderwałam się od tej brzydkiej słomki! — odezwała się pierwsza bańka i zarumieniona z dumy i radości poleciała nad grządkę ze stokrotkami.

(*flushed is subordinated to floated, which is coordinated with said*)

We ran the tagger on the Polish and Bulgarian texts. Then we compared the tags.

The result of the automatic disambiguation of the first sentence of 'Soap Bubbles' by Svetoslav Minkov

*Имаше едно малко момиченце с червена панделка на коса*та.

Była sobie raz dziewczynka z piękną czerwoną wstążką we włosach.

'There was once a little girl with a red ribbon in her hair.'

can be found in the Appendix.

The table below shows the tags assigned to the words; where there are two or more possible analyses, the one which is actually chosen is shown first.

Bulgarian	Bulgarian MSDs	Bulgarian ctags	Polish	Polish ctags
имаше	Vmii3s Vmii2s	VMII3S VMII2S	*była*	adj:sg:nom:f:pos praet:sg:f:imperf
			sobie	siebie:dat siebie:loc
			raz	subst:sg:nom:m3 subst:sg:acc:m3
едно	Mcns-ln	MC		
малко	A--ns-n Ra Ncns-n	ANS RA NCNS-N		
момиченце	Ncns-n	NCNS-N	*dziewczynka*	subst:sg:nom:f
с	Sp	SP	*z*	prep:gen:nwok prep:inst:nwok qub
			piękną	adj:sg:acc:f:pos adj:sg:inst:f:pos
червена	A--fs-n Vmps-sfp-n	AFS VMPS-SF	*czerwoną*	adj:sg:acc:f:pos adj:sg:inst:f:pos
панделка	Ncfs-n	NCFS-N	*wstążką*	subst:sg:inst:f
на	Sp Qgs	SP QG	*we*	prep:loc:wok prep:acc:wok
*коса*та	Ncfs-y	NCFS-Y	*włosach*	subst:pl:loc:m3
.		PERIOD	.	interp

The Bulgarian tags stand for, as follows:

AFS	Adjective feminine singular
ANS	Adjective neutral singular
MC	numeral cardinal
NCFS-N	noun common feminine singular indefinite
NCFS-Y	noun common feminine singular definite
NCNS-N	noun common neuter singular indefinite
PERIOD	Period
QG	particle general
RA	adverb adjectival
SP	Adposition prepositive
VMII2S	verb main indicative imperfect 2nd singular
VMII3S	verb main indicative imperfect 3rd singular
VMPS-SF	verb main participle past singular feminine

The Polish tags stand for, as follows:

adj:sg:acc:f:pos	adjective : singular : accusative : feminine : positive
adj:sg:inst:f:pos	adjective : singular : instrumental : feminine : positive
adj:sg:nom:f:pos	adjective : singular : nominative : feminine : positive
interp	punctuation
praet:sg:f:imperf	quasiparticiple : singular : feminine : imperfective
prep:acc:wok	preposition : accusative : vocalised
prep:gen:nwok	preposition : genitive : unvocalised
prep:inst:nwok	preposition : instrumental : unvocalised
prep:loc:wok	preposition : locative : vocalised
qub	qublik (particle-adverb)
siebie:dat	siebie : dative
siebie:loc	siebie : locative
subst:pl:loc:m3	noun : plural : locative : masculine (inanimate)
subst:sg:acc:m3	noun : singular : accusative : masculine (inanimate)
subst:sg:inst:f	noun : singular : instrumental : feminine
subst:sg:nom:f	noun : singular : nominative : feminine
subst:sg:nom:m3	noun : singular : nominative : masculine (inanimate)

Regarding the tagsets, the main differences between them (ignoring the mismatches in the names of matching tags, which can be amended easily) are due to the different morphological makeup of the two languages: Polish has morphological case pattern for all nominal parts of speech (seven cases) which Bulgarian has almost entirely lost (with the exception of a vestigial vocative in the noun and rudimentary declension of the personal pronoun); by contrast, Bulgarian has a definite article which was originally an enclitic but has merged with the noun, adjective or numeral into a single word form, giving an inflexional category of definiteness. For example:

Bulgarian

косата коса Ncfs-y

[wordform *косата* 'the hair', lemma *коса* 'hair'] POS: Noun, Type: common, Gender: feminine, Number: singular, Definiteness: yes;

Polish

włosach włos subst:pl:loc:m3

[wordform *włosach*, lemma *włos* '(strand of) hair'] POS: substantive (noun), Number: plural, Case: locative, Gender: masculine 3 (inanimate).

Bulgarian has also preserved more of the verb conjugation of Old Slavic, whereas in Polish verb conjugation is relatively simple, especially if the floating cliticised copula (a Polish innovation) is considered a separate word, as in this analyser (e.g., *przyszedłem* 'I came' is analysed as *przyszedł* 'came' + *-em* 'I').

The number of MSDs was reduced with respect to the c-tags in Bulgarian (from 324 MSDs, used in Bulgarian MTE corpus, to 117 c-tags to run the ISSCO tagger) due to software limitations 15 years ago. For example, instead of the three MSDs for masculine singular forms of adjectives (A--ms-f with full article, A--ms-s with short article, A--ms-n with none) the single c-tag AMS was used. The five MSDs for the demonstrative pronoun (Pd-------q, Pd--p----p, Pd-fs----p, Pd-ms----p, Pd-ns----p) were collapsed to the c-tag PD; for the 15 MSDs for relative pronoun (Pr-------q, Pr--p----a, Pr--p----p, Pr--p----s, Pr-fs----a, Pr-fs----p, Pr-fs----s, Pr-ms----a, Pr-ms----s, Pr-msa---p, Pr-msd---p, Pr-msn---p, Pr-ns----a, Pr-ns----p, Pr-ns----s) the c-tag PR is used. Due to increased computing power today we think that such a reduction is no longer necessary, so that the morphosyntactic descriptions are fully preserved at POS annotation.

We do not consider the different nomenclature of POS tags in Polish and Bulgarian to be a significant problem because a one-to-one correspondence could easily resolve it.

It is interesting to compare the set of tags for Polish used by Morfeusz to MTE's set of tags for synthetic Slavic languages (Czech, Slovak, Slovene), whose grammatical categories are closer to those of Polish. Some of the differences are obvious (e.g., Polish has no dual number, whereas Czech has a vestigial dual and Slovene a full-fledged one). The following table summarises the less trivial differences between the approaches:

	MTE (Czech, Slovak, Slovene)	Morfeusz (Polish)
noun class	split into orthogonal categories of gender (m, f, n) and animacy (no, yes)	treated as a single multi-value category of gender
cliticised copula	treated as a feature of the host word	treated as a separate word
past tense	treated as a value of the category of tense (past)	treated as a compound of a quasi-participle and a cliticised copula
imperfective present vs perfective future	distinguished as different tenses	distinguished only by aspect
prepositional form of 3rd person pronoun	disregarded	labelled by means of an express category as postprepositional

Conclusion

Our scrutiny of the outcome of the tagging of the short texts leads us to believe that a unification of the morphosyntactic annotation for Bulgarian and Polish should be done within the perspective of the elaboration of a general tagset for Slavic languages.

References

1. Dimitrova, 1998: Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., and Tufis, D. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: Proceedings of COLING-ACL '98. Montréal, Québec, Canada, pp. 315-319.

2. Dimitrova, Koseska, 2007: Dimitrova, L., V. Koseska – Toszewa. Digital Dictionaries – Problems and Features. In: Proceedings of the Jubilee International Conference Mathematical and Computational Linguistics. 6 July 2007, Sofia, Bulgaria. 2007. Pages 25-34. ISBN 978-954-8986-28-1.

3. Dimitrova, Koseska, 2008: Dimitrova, L., V. Koseska–Toszewa. Some Problems in Multilingual Digital Dictionaries. In: International Journal *ÉTUDES COGNITIVES*. Vol. 8. SOW, Warsaw. 2008. Pages 237-254. ISSN 1641-9758.

4. Ide, Véronis, 1994: Ide, N., and Véronis, J.: Multext (multilingual tools and corpora). In *COLING'94*, pages 90-96, Kyoto, Japan, 1994.

5. Ide et al. 2000: Ide, N., Bonhomme, P., and Romary, L. XCES: An XMLbased Encoding Standard for Linguistic Corpora. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 2000. 825-830.

6. Leech 2004: Geoffrey Leech. Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. 2004.

7. http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm

8. Monachini, Calzolari, 1996: Monachini, M. and Calzolari, N. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG-CLWG-MORPHSYN/R. 1996. http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/

9. Piasecki, 2007: Piasecki M., Polish Tagger TaKIPI: Rule Based Construction and Optimisation. Task Quarterly. 2007, 11, p. 151-167.

10. Piasecki, Godlewski, 2006: Piasecki, M., Godlewski, G.. Reductionistic, Tree and Rule Based Tagger for Polish. In: Kłopotek, M. A., Wierzchoń, S. T., i Trojanowski, K., red. (2006). Intelligent Information Processing and Web Mining — Proceedings of the International IIS: IIPWM'06 Conference held in Zakopane, Poland, June, 2006. Advances in Soft Computing. Springer, Berlin.

11. Piasecki, Wardyński, 2006: Piasecki, M., Wardyński, A., Multiclassifier Approach to Tagging of Polish. In: Proceedings of 1st International Symposium Advances in Artificial Intelligence and Applications. 2006.

12. Woliński, 2006: Woliński, M., Analizator morfologiczny *Morfeusz SIAT*. (In Polish) http://nlp.ipipan.waw.pl/~wolinski/morfeusz/

13. Woliński, 2003: Woliński, M., System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica XII, 2003, p. 39-55. (In Polish)

14. Bulgarian Grammar, 1993: Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

ISSCO tagger: http://www.issco.unige.ch/staff/robert/tatoo/tagger.html#design

Appendix

The first sentence of 'Soap Bubbles' by Svetoslav Minkov is:

Имаше едно малко момиченце с червена панделка на косата.

'There was a little girl with a red ribbon in her hair.'

(1) in Bulgarian (ISSCO tagger)

```
<tok type=WORD>
      <orth>Имаше</orth>
      <disamb><base>имам</base><ctag>VMII3S</ctag></disamb>
      <lex><base>имам</base><msd>Vmii2s</msd><ctag>VMII2S</ctag></lex>
      <lex><base>имам</base><msd>Vmii3s</msd><ctag>VMII3S</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>едно</orth>
      <disamb><base>едно</base><ctag>MC</ctag></disamb>
      <lex><base>едно</base><msd>Mcns-ln</msd><ctag>MC</ctag></lex>
    </tok>
<tok type=WORD>
<orth>малко</orth>
   <disamb><base>малък</base><ctag>ANS</ctag></disamb>
   <lex><base>малък</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
   <lex><base>малко</base><msd>Ra</msd><ctag>RA</ctag></lex>
   <lex><base>малко</base><msd>Ncns-n</msd><ctag>NCNS-N</ctag></lex>
   </tok>
<tok type=WORD>
      <orth>момиченце</orth>
      <disamb><base>момиченце</base><ctag>NCNS-N</ctag></disamb>
      <lex><base>момиченце</base><msd>Ncns-n</msd><ctag>NCNS-N</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>с</orth>
      <disamb><base>с</base><ctag>SP</ctag></disamb>
      <lex><base>с</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
<tok type=WORD>
    <orth>червена</orth>
    <disamb><base>червен</base><ctag>AFS</ctag></disamb>
    <lex><base>червен</base><msd>A--fs-n</msd><ctag>AFS</ctag></lex>
    <lex><base>червя</base><msd>Vmps-sfp-n</msd><ctag>VMPS-SF</ctag></lex>
     </tok>
<tok type=WORD>
      <orth>панделка</orth>
      <disamb><base>панделка</base><ctag>NCFS-N</ctag></disamb>
      <lex><base>панделка</base><msd>Ncfs-n</msd><ctag>NCFS-N</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>на</orth>
      <disamb><base>на</base><ctag>SP</ctag></disamb>
      <lex><base>на</base><msd>Qgs</msd><ctag>QG</ctag></lex>
      <lex><base>на</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>косата</orth>
      <disamb><base>коса</base><ctag>NCFS-Y</ctag></disamb>
      <lex><base>коса</base><msd>Ncfs-y</msd><ctag>NCFS-Y</ctag></lex>
    </tok>
<tok type=PUNCT>
     <orth>.</orth>
     <ctag>PERIOD</ctag>
    </tok>
```

(2) in Polish (Piasecki, 2007)

Była sobie raz dziewczynka z piękną czerwoną wstążką we włosach.
<tok>
<orth>Była</orth>
<lex><base>były</base><ctag>adj:sg:nom:f:pos</ctag></lex>
<lex disamb="1"><base>być</base><ctag>praet:sg:f:imperf</ctag></lex>
</tok>
<tok>
<orth>sobie</orth>
<lex disamb="1"><base>się</base><ctag>siebie:dat</ctag></lex>
<lex><base>się</base><ctag>siebie:loc</ctag></lex>
</tok>
<tok>
<orth>raz</orth>
<lex><base>raz</base><ctag>subst:sg:nom:m3</ctag></lex>
<lex disamb="1"><base>raz</base><ctag>subst:sg:acc:m3</ctag></lex>
</tok>
<tok>
<orth>dziewczynka</orth>
<lex disamb="1"><base>dziewczynka</base><ctag>subst:sg:nom:f</ctag></lex>
</tok>
<tok>
<orth>z</orth>
<lex><base>z</base><ctag>prep:gen:nwok</ctag></lex>
<lex disamb="1"><base>z</base><ctag>prep:inst:nwok</ctag></lex>
<lex><base>z</base><ctag>qub</ctag></lex>
</tok>
<tok>
<orth>piękną</orth>
<lex><base>piękny</base><ctag>adj:sg:acc:f:pos</ctag></lex>
<lex disamb="1"><base>piękny</base><ctag>adj:sg:inst:f:pos</ctag></lex>
</tok>
<tok>
<orth>czerwoną</orth>
<lex><base>czerwony</base><ctag>adj:sg:acc:f:pos</ctag></lex>
<lex disamb="1"><base>czerwony</base><ctag>adj:sg:inst:f:pos</ctag></lex>
</tok>
<tok>
<orth>wstążką</orth>
<lex disamb="1"><base>wstążka</base><ctag>subst:sg:inst:f</ctag></lex>
</tok>
<tok>
<orth>we</orth>
<lex disamb="1"><base>w</base><ctag>prep:loc:wok</ctag></lex>
<lex><base>w</base><ctag>prep:acc:wok</ctag></lex>
</tok>
<tok>
<orth>włosach</orth>
<lex disamb="1"><base>włos</base><ctag>subst:pl:loc:m3</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>