Tagger WCRFT2 Polish

Description

In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context — i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Schools commonly teach that there are 9 parts of speech in English: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. However, there are clearly many more categories and sub-categories. For nouns, the plural, possessive, and singular forms can be distinguished. In many languages words are also marked for their "case" (role as subject, object, etc.), grammatical gender, and so on; while verbs are marked for tense, aspect, and other things. Linguists distinguish parts of speech to various fine degrees, reflecting a chosen "tagging system".

In case of Polish WCRFT2 Tagger for the example input sentence:

```
PL:     Mamy   bardzo miłego psa.
EN: We have a very   nice   dog.
```

the output is:


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE chunkList SYSTEM "ccl.dtd">
<chunkList>
 <chunk id="ch1" type="p">
  <sentence id="s1">
   <tok>
    <orth>Mamy</orth>
    <lex disamb="1"><base>mieć</base><ctag>fin:pl:pri:imperf</ctag></lex>
   </tok>
   <tok>
    <orth>bardzo</orth>
    <lex disamb="1"><base>bardzo</base><ctag>adv:pos</ctag></lex>
   </tok>
   <tok>
    <orth>miłego</orth>
    <lex disamb="1"><base>miły</base><ctag>adj:sg:acc:m2:pos</ctag></lex>
   </tok>
   <tok>
    <orth>psa</orth>
    <lex disamb="1"><base>pies</base><ctag>subst:sg:acc:m2</ctag></lex>
   </tok>
   <ns/>
   <tok>
    <orth>.</orth>
```

```
    <lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
   </tok>
  </sentence>
 </chunk>
</chunkList>
```

The input sentence is tokenized and each token is represented by <tok></tok> section, with the information about base form of a word (lemma), part of speech and other attributes:

```
   <tok>
    <orth>miłego</orth>                  //text word
    <lex disamb="1">
      <base>miły</base>                  //lemma
      <ctag>adj:sg:acc:m2:pos</ctag>     //tags
    </lex>
   </tok>
```

For the given example *tags* are:

- adj - adjective (part of speech)
- sg - singular (number)
- acc - accusative (case)
- m2 - animate masculine (gender)
- pos - positive (degree)

The full list of tags for Polish is described here.

Input

Plain text file (UTF-8) in Polish.

Output

File in CCL format. Morphological tags are presented in NKJP format.