Corpus of frequency dictionary of contemporary Polish (1963-67) – new annotation

Łukasz Szałkiewicz
Institute of Computer Science, Polish Academy of Sciences
lukasz.szalkiewicz@ipipan.waw.pl

1. About Corpus

The original purpose of the corpus was to create a general frequency dictionary of contemporary Polish. The work started in 1967. Partial results were published between 1972 and 1977, the completed dictionary in 1990. The corpus was later augmented in various respects, both by manual editing and automated procedures.

Corpus data contain 10,000 samples divided into 5 parts:

- essays,
- news,
- scientific texts,
- fiction
- plays.

Every sample is approximately 50 words long, they all come from texts published between 1963 and 1967 and contain bibliographic description of its source. Each word is tagged with its base form and some morphological properties. Sentence boundaries are also marked.

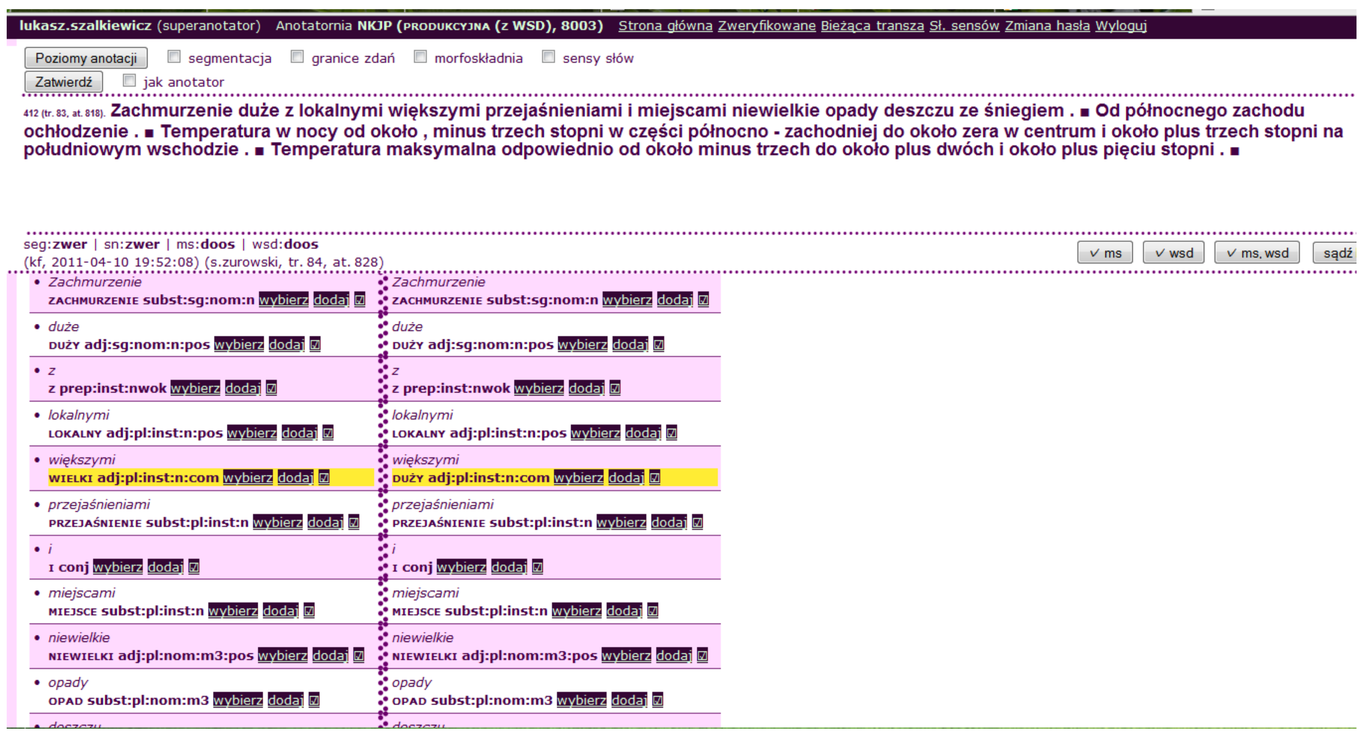
In 2001 corpus authors agreed to publish the data in the Internet under GNU licence.

2. New linguistic annotation

New annotation – according to rules established in annotation of National Corpus of Polish (NKJP).

1. All corpus data has been manually verified by two independent annotators (but see next paragraph), connecting with the server via a web interface.
2. Any conflicts have been resolved by a referee (so-called 'superannotator').
3. The linguistic annotation was carried out at three levels: word-level segmentation, sentence-level segmentation and morphosyntax.
4. There was used the same purpose-built tool, Anotatornia.
5. NKJP tagset and the same guidelines.

3. Screen from Anotatornia (tool for corpus annotation)



4. Using original annotation – tagset changes

Manual annotation is costly task in the development of a corpus. One way to reduce the cost was to use data from original annotation as outcome of work imaginary annotator, called 'KF' (in Polish *Korpus Frekwencyjny*). So **original annotation was treated as if it were made by the first annotator**. But these data were congruent with IPI PAN tagset not with NKJP tagset. Some automatical changes were needed to prevent repeated discrepancies between KF and second, real annotator.

There were created **tagsets overview of the differences** and it was transformed in a three-part list: what can be done automatically, semi-automatically and manually.

Examples:

- NKJP tagset differentiates between coordinate conjunctions (Pol. spójniki równorzędne; conj), e.g. *i, lub, oraz*, and subordinate conjunctions (Pol. spójniki podrzędne), sometimes called complementisers (comp), e.g. *że, aby, bowiem*.
- There where the class of particle-adverbs in IPI PAN tagset, separate from the class of adverbs (adv) which consists of only de-adjectival or gradable adverbs. The class of adverbs is larger in NKJP tagset than in IPI PAN, also includes traditional adverbs which are neither de-adjectival nor gradable.

NKJP tagset		IPI PAN tagset	
word	tag	tag	automatic change for all occurrences
że	comp (or qub, but much less)	conj	comp
stąd	adv	qub	adv
tu	adv	qub	adv
dlaczego	adv	qub	adv
we	prep:wok	prep	prep:wok
ponad	prep or qub	prep	prep, if next is num

5. Additional works

- there were no ready **sentence segmentation** in corpus data, so it was added automatically according to Marcin Milkowski's rules
- some minor changes in the texts of the corpus , such as the removal of double dots in the text of the drama
- methods of **changing the bases to begin with a capital letter** were developed, because in the original annotation all saved lemmas were always lower case, resulting in many discrepancies

6. Bibliography

You can see the most current data, corpus documentation and selected bibliography on http://clip.ipipan.waw.pl/PL196x

7. People

- Anna Andrzejczuk (annotator, tagsets conversion)
- Michał Lenart (computer engineer)
- Łukasz Szałkiewicz (superannotator, tagsets conversion)
- Sebastian Żurowski (main annotator)