Error Correction Environment for the Polish Parliamentary Corpus

Maciej Ogrodniczuk
Michał Rudolf
Beata Wójtowicz
Sonia Janicka

Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

ParlaCLARIN III Workshop, LREC 2022

Marseille, June 20, 2022

The Polish Parliamentary Corpus

In a nutshell:

- an 800M-token collection of linguistically annotated documents from the proceedings of Polish Parliament (Sejm and Senate)
- prepared in a series of subsequently running projects (CESAR, CLARIN-PL, MARCELL, ParlaMint, CLARIN-PL-Biz)
- gathering proceedings between 1919 and now
- three main document types: stenographic transcriptions of plenary sittings, committee sittings and parliamentary questions
- data linguistically analysed and saved in stand-off XML TEI National Corpus of Polish format
- primary link: `http://clip.ipipan.waw.pl/PPC`

Data cleanup still needed

Heterogeneous process of adding data to the corpus:

- from almost-direct inclusion of newest born-digital data already available in clean formats
- to tedious correction of automatically OCR-ed image-based PDF files containing older materials

Still many problems with the data:

- structural errors (such as unmarked speakers, enumerations, comments or retained unnecessary header information)
- typographical errors (punctuation errors, various misspellings)
- other errors (non-textual elements, HTML fragments etc.)

The solution

A new proofreading round:

- with pre-detected errors (how?)
  - with a language-based model?
  - with custom rules?
- in some (new?) error correction environment
  - easy to use by non-technical users (XML-based?)
  - how to consult the source?

Error candidate detection

Two experiments:

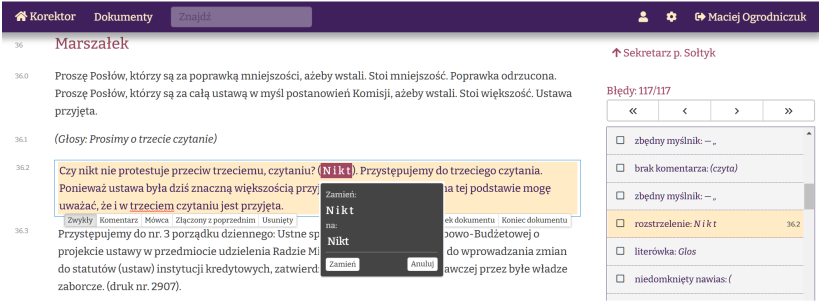
1. language model-based:
   - a sequence to sequence model using plT5 model for Polish
   - successful in discovering and correcting such cases as two words glued together, missing or excessive spaces and several types of grammatical errors
   - still, the number of false positives rendered its use impractical
2. rule-based:
   - very precise
   - composed of several modules corresponding to various error categories

Rule-based solution

Detected error types:

- **structural errors**: mostly merged enumerations or speaker names treated as normal text
- **comments and metadata** marked in original texts with simple brackets leading to many conversion errors
- **punctuation errors**, e.g. unmatched quotation marks or brackets, excessively hyphenated words etc.
- **broken or unfinished paragraphs** resulting from conversion errors or signalling missing content
- **misspellings** resulting in OOV words → use dictionary
- **common OCR errors or typos** resulting in highly improbable in-dictionary words → use frequency lists
- **other errors**, e.g. remains of non-textual elements such as tables or footnotes, characters outside the common character set or spaced-out words

A new Web-based correction environment



36 Marszałek

36.0 Proszę Posłów, którzy są za poprawką mniejszości, ażeby wstali. Stoi mniejszość. Poprawka odrzucona. Proszę Posłów, którzy są za całą ustawą w myśl postanowień Komisji, ażeby wstali. Stoi większość. Ustawa przyjęta.

36.1 *(Głosy: Prosimy o trzecie czytanie)*

36.2 Czy nikt nie protestuje przeciw trzeciemu, czytaniu? (N i k t). Przystępujemy do trzeciego czytania. Ponieważ ustawa była dziś znaczną większością przyj … na tej podstawie mogę uważać, że i w trzeciem czytaniu jest przyjęta.

Zamień:
N i k t
na:
Nikt

36.3 Przystępujemy do nr. 3 porządku dziennego: Ustne sp … owo-Budżetowej o projekcie ustawy w przedmiocie udzielenia Radzie Mi … do wprowadzania zmian do statutów (ustaw) instytucji kredytowych, zatwierd … awczej przez byłe władze zaborcze. (druk nr. 2907).

↑ Sekretarz p. Sołtyk

Błędy: 117/117

- □ zbędny myślnik: — „
- □ brak komentarza: *(czyta)*
- □ zbędny myślnik: — „
- □ rozstrzelenie: *N i k t* 36.2
- □ literówka: *Glos*
- □ niedomknięty nawias: *(*

https://korektor.rudolf.waw.pl

PDF page viewer add-on

An idea for a subproject:

- take a 'dirty OCR' of the original graphical source PDF
- compare it with the clean XML text of a transcript
- insert page boundary markers in the XML

Components:

- Tesseract OCR engine
- word on page boundaries compared with Levenshtein distance
- compensation mechanisms for special cases:
  - pages containing tables (previously removed from the corpus XML files)
  - hyphenated words at the end of the page

Detected errors

In the whole data set:

All detected errors	778 479
Punctuation errors	427 830
Broken or unfinished paragraphs	121 182
Misspellings	116 997
Structural errors	71 790
Comments and metadata	18 452
Other errors	40 680

Corrected errors

Until now:

All corrections	**606 506**	**100%**
Suggestion-based	344 929	57%
Newly introduced	261 577	43%
Structural (crossing paragraphs)	522 064	86%
Textual (inside a paragraph)	84 442	14%

Thank you!

And several funding institutions and helpful friends:

- **CLARIN-PL**: The 2014–2020 Smart Development Operational Programme, Priority IV: Increasing the scientific and research potential, Measure 4.2: Development of modern research infrastructure of the science sector, No. POIR.04.02.00-00C002/19, “CLARIN — Common Language Resources and Technology Infrastructure“
- **CLARIN-ERIC** for funding ParlaMint: Towards Comparable Parliamentary Corpora.
- **Krzysztof Wróbel** for his language model-based error candidate detection experiment.