The Polish Parliamentary Corpus

Maciej Ogrodniczuk

Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

CLARIN-PLUS Worskhop on Working with Parliamentary Records
Sofia, 27–29 March 2017

The Polish Parliamentary Corpus

In a nutshell:

- a large collection of linguistically annotated documents from the proceedings of Polish Parliament, Sejm and Senate
- based on the Polish Sejm Corpus prepared in October 2011 and recently updated by CLARIN-PL
- automatically created multi-layer annotation compatible with the National Corpus of Polish
- 42.4 GB compressed data, 300M segments
- available at: `http://clip.ipipan.waw.pl/PPC`

Corpus data

Data sources:

- 1993–: editable PDFs, other formats available (internally)
- 1989–93: scanned paper transcripts (OCR-ed and manually verified)

Corpus format and structure:

- stand-off TEI P5 annotation
- generated with in-house tools:
  - Morfeusz SGJP (text structure, utterance-level segmentation, tokenization, lemmatization)
  - Pantera (disambiguated morphosyntactic description)
  - Spejd (syntactic words and syntactic groups)
  - Nerf (named entities).

Source scans

(Początek posiedzenia o godz. 17 min. 05)

(Na salę wchodzą witane owacyjnymi oklaskami delegacje Związku Socjalistycznych Republik Radzieckich, Czechosłowackiej Republiki Socjalistycznej oraz Niemieckiej Republiki Demokratycznej)

(Przewodniczy na posiedzeniu Marszałek Sejmu Czesław Wycech)

Marszałek:

Otwieram posiedzenie.

Na sekretarzy powołuję posłów Reginę Pawlikowską i Zdzisława Kurowskiego.

Protokół i listę mówców prowadzi poseł Regina Pawlikowska.

Proszę Posłów Sekretarzy o zajęcie miejsc przy stole prezydialnym.

Prezydium Sejmu usprawiedliwia nieobecność na posiedzeniu posłów, których nazwiska będą umieszczone w załączniku do protokółu dzisiejszego posiedzenia.

Ustalony przez Prezydium Sejmu porządek dzienny obecnego posiedzenia obejmuje uroczyste uczczenie 25-lecia Polskiej Rzeczypospolitej Ludowej.

Czy w sprawie porządku dziennego są jakieś uwagi?

Nie słyszę.

Uważam, że Sejm porządek dzienny, przedstawiony przez Prezydium Sejmu, zatwierdził.

Obywatele Posłowie! Uroczyste nasze posiedzenie odbywa się w szczególnie podniosłym momencie — w przeddzień 25 rocznicy powstania Polski Ludowej. W posiedzeniu naszym biorą udział dostojni i drodzy nam Goście. Niech mi więc będzie wolno, otwierając posiedzenie, w imieniu Was wszystkich, Obywatele Posłowie, pod Ich adresem przede wszystkim skierować moje słowa.

Witam serdecznie przybyłą na nasze posiedzenie delegację Związku Socjalistycznych Republik Radzieckich w osobach Sekretarza Generalnego Komitetu Centralnego Komunistycznej Partii Związku Radzieckiego Towarzysza Leonida Breżniewa i Przewodniczącego Prezydium Rady Najwyższej ZSRR Towarzysza Nikołaja Podgornego. *(Owacyjne oklaski)*

Witam serdecznie delegację Czechosłowackiej Republiki Socjalistycznej w osobach Prezydenta Czechosłowackiej Republiki Socjalistycznej Towarzysza Ludvika Svobody i I Sekretarza Komitetu Centralnego Komunistycznej Partii Czechosłowacji Towarzysza Gustava Husaka. *(Owacyjne oklaski)*

Witam serdecznie delegację Niemieckiej Republiki Demokratycznej w osobach Prezesa Rady Ministrów i zastępcy Przewodniczącego Rady Państwa Niemieckiej Republiki Demokratycznej Towarzysza Willi Stopha oraz członka Biura Politycznego, Sekretarza Komitetu Centralnego Niemieckiej Socjalistycznej Partii Jedności Towarzysza Ericha Honeckera. *(Owacyjne oklaski)*

Witam wszystkich obecnych.

Przystępujemy do porządku dziennego.

Proszę o zabranie głosu I Sekretarza Komitetu Centralnego Polskiej Zjednoczonej Partii Robotniczej posła Władysława Gomułkę. *(Długotrwałe oklaski)*

I Sekr. KC PZPR
poseł Władysław Gomułka:

Obywatelu Marszałku, Wysoka Izbo! Przed 25 laty otwarła się nowa karta dziejów Polski.

W Chełmie, w pierwszym uwolnionym od okupanta mieście polskim powołany został z woli Krajowej Rady Narodowej — podziemnego parlamentu polskiej demokracji — Polski Komitet Wyzwolenia Narodowego. Jego siedzibą i zarazem pierwszą stołeczną kolebką Polski Ludowej było miasto Lublin.

Manifest Polskiego Komitetu Wyzwolenia Narodowego obwieszczający odrodzenie Polski stał się zapowiedzią nowej epoki w historii naszego narodu.

I Sekr. KC PZPR poseł Gomułka

Obóz demokracji polskiej, któremu przewodziła Polska Partia Robotnicza włożył w słowa Manifestu Lipcowego całe swe historyczne doświadczenie, wyraził w nich wiekowe dążenia i pragnienia ludu pracującego. Program Manifestu stał się fundamentem, na którym krok po kroku, piętro za piętrem, od 25 lat wznosimy gmach nowej Polski — Polski socjalistycznej.

W tych pierwszych dniach wolności, gdy lud pracujący rozpoczynał budowę swojej państwowości, nasz ówczesny Rząd — Polski Komitet Wyzwolenia Narodowego — tyle miał ziemi pod stopami, ile jej co dzień zdołały wywalczyć korpusy radzieckie i dywizje polskie w krwawych bojach z hitlerowskim okupantem. Tyle miał siły, ile mu udzieliły świeżo wyzwolone wsie i miasteczka, ile mu jej dali z siebie ludzie pracy, skupiający się wokół swej władzy.

Wówczas w Lublinie, kiedy zakładaliśmy społeczne, polityczne i prawne podstawy pod nową Ludową Rzeczpospolitą, rzeczą najważniejszą była wojna z najeźdźcą hitlerowskim, wojna, którą naród polski toczył nieustannie od 1 września 1939 r. i która daleka jeszcze była od zakończenia. Gdy PKWN schodził ze sceny historycznej, ustępując miejsca Rządowi Tymczasowemu, front stał na Wiśle, Narwi i Wisłoce, 75% ziemi polskiej znajdowało się jeszcze pod władzą okupanta, 75% ludności polskiej narażone było wciąż na postępującą eksterminację. Dorobek materialny Polaków ulegał dalszemu niszczeniu przez okupanta. Toteż sprawę wojny, sprawę przyśpieszenia wyzwolenia i „wzmożenia udziału narodu polskiego w walce o zmiażdżenie Niemiec hitlerowskich", traktowaliśmy jako naczelne zadanie nowej władzy, najwyższy nakaz dany nam przez naród.

W dniu swych narodzin ludowe państwo polskie miało 65 tysięcy żołnierzy na froncie — I Armię Wojska Polskiego stworzoną przez lewicę polską na ziemi radzieckiej, przy jakże doniosłej radzieckiej pomocy, a także 40 tysięcy młodego żołnierza w powstających dopiero formacjach, nie licząc partyzantów Armii Ludowej, oraz ludowego, antyhitlerowskiego podziemia. Po pięciu miesiącach działalności PKWN mieliśmy już 286 tysięcy nowocześnie uzbrojonych żołnierzy w szeregach zjednoczonego ludowego Wojska Polskiego. Pod koniec wojny, wiosną 1945 roku, mieliśmy pod bronią prawie 400 tysięcy ludzi, z czego 185 tysięcy wzięło udział w wielkich wiosennych bitwach, które przyniosły wolność całej Polsce i zadecydowały o ostatecznym zwycięstwie nad hitlerowską tyranią.

Udział ludowego Wojska Polskiego w ofensywie styczniowej, a zwłaszcza w walkach na Pomorzu w lutym i marcu 1945 roku, ważył na przebiegu operacji wojskowych na froncie wschodnim i przyczynił się w sposób istotny do skrócenia okupacji ziem środkowej i północnej Polski. Armie polskie, uczestniczące w ofensywie kwietniowej, stanowiły 8 do 10% ogółu wojsk walczących w gigantycznej bitwie między Sudetami a Bałtykiem, która przesądziła o ostatecznej klęsce Hitlera. Toteż fakt, iż właśnie nasz żołnierz, jedyny z sojuszników stanął obok żołnierza radzieckiego w Berlinie, stał się symbolem dziejowej sprawiedliwości.

Ten udział Polaków w ostatecznym zwycięstwie nad hitleryzmem odegrał wielką rolę w historycznych rozstrzygnięciach, mających na celu zabezpieczenie przyszłości Polski i Europy przed zaborczością niemiecką, rozstrzygnięciach, które przywróciły Polsce jej ziemie piastowskie po Odrę i Nysę Łużycką i wybrzeże Bałtyku po Szczecin.

Sprawiedliwe dla Polski rozstrzygnięcie w Poczdamie, mające u swych podstaw decyzje konferencji trzech mocarstw w Jałcie, naród polski opłacił straszliwą daniną krwi, cierpień i strat w czasie II wojny światowej. Rachunek zbrodni hitlerowskich dokonanych na narodzie polskim jest przerażający. 6 milionów poległych i pomordowanych — 22% biologicznej substancji narodu — to liczby, których wstrząsająca wymowa jest nie do zatarcia. A przecież nie wyczerpują one wcale ogromu strat Polski. Z każdego tysiąca obywateli polskich, zamieszkałych w Polsce w 1939 r., 220 zostało zamordowanych, 37 — okaleczonych do ciężkiego inwalidztwa; na gruźlicę zapadło około 100, do obozów pracy i na przymusowe roboty wywieziono przeszło 100, pozbawiono całego dobytku i wysiedlono — 90, pozostawiono bez dachu nad głową 150.

Całkowitemu zniszczeniu uległo prawie 40% majątku narodowego, wartość równa dorobkowi pracy dwóch poprzednich pokoleń Polaków.

Ten straszliwy bilans wyrażał jedynie częściową realizację zbrodniczych planów okupanta, planów całkowitej likwidacji narodu polskiego jako narodu zdolnego do życia. Zwycięzcy — jak cynicznie oświadczał Hitler w przededniu napaści na Polskę — nikt nie odważy się sądzić.

Dziś raz jeszcze przypomnieć wypada, że jeśli te obłędne plany pełnej eksterminacji narodu polskiego nie zostały urzeczywistnione, zawdzięczamy to w stopniu decydującym Związkowi Radzieckiemu, jego bohaterskiej armii.

Przyśpieszenie wyzwolenia o każdy tydzień oznaczało dla nas uratowanie życia 20 tysiącom ludzi — tylu bowiem przeciętnie mordowali hitlerowcy w ciągu każdego tygodnia okupacji w naszym kraju.

Główną rolę w zwycięstwie nad faszyzmem odegrała Armia Radziecka. Ona to — jak świadczą tajne wykazy strat armii hitlerowskiej — zadała prawie 80% wszystkich strat poniesionych przez hitlerowskie wojska lądowe w okresie od 1 września 1939 r. do 29 kwietnia 1945 r.

Linguistic annotation format

Text structure:

```
<teiCorpus>
  <xi:include href="corpus_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <body>
        <!-- ... -->
        <div xml:id="txt_7-div">
          <u xml:id="txt_7.1-u" who="#The_Speaker">
            <!-- ... -->
          </u>
          <u xml:id="txt_7.2-u" who="#MP_Jan_Nowak">
            <!-- ... -->
```

Linguistic annotation format

Sentence- and token-level segmentation:

```
<p xml:id="segm_txt_7.1-u"
   corresp="text_structure.xml#txt_7.1-u">
  <s xml:id="segm_txt_7.1-u.1-s">
    <!-- ... -->
    <!-- Proszę -->
    <seg xml:id="segm_txt_7.1-u.4-seg"
         corresp="text_structure.xml
         #string-range(txt_7.1-u,0,6)"/>
    <!-- państwa -->
    <seg xml:id="segm_txt_7.1-u.5-seg"
         corresp="text_structure.xml
         #string-range(txt_7.1-u,7,7)"/>
    <!-- ... -->
```

Linguistic annotation format

Morphosyntactic annotation:

```
<seg xml:id="morph_txt_7.1-u.45-seg"
     corresp="ann_segmentation.xml
     #segm_txt_7.1-u.45-seg">
  <fs type="morph">
    <f name="orth">
     <string>państwa</string>
    </f>
    <!-- państwa [7,7] -->
    <f name="interps">
      <fs type="lex" xml:id="...">
       <f name="base">
         <string>państwo</string></f>
       <f name="ctag">
         <symbol value="subst"/></f>
       <!-- ... -->
```

What's next?

CLARIN-PL activities:

- processing data with newest analytic tools
- live corpus
- more annotation layers
- better search and presentation
- rich(er) metadata
- consultation of other versions
- audio/video linking
- more data!

More data

Which data?

- high-quality scans prepared by the Sejm library
- manually cleaned
- 2000 documents, 90K pages from 1918–1989
- two political periods most interesting for the humanities

How to process them?

- human-assisted OCR (area and typo corrections)
- scripts for structure and error detection