Slavic Corpora Terminology Dictionary

LexMC Berlin results

Joanna Bilińska
j.bilinska@uw.edu.pl

7.12.2018, Berlin

The Dictionary

- corpora terminology
- digitally born dictionary
- languages:
  - Czech,
  - Slovene,
  - Bulgarian,
  - Polish,
  - Slovak,
  - (English) — only within the entries
- to be added:
  - Croatian,
  - Serbian,
  - Macedonian,
  - maybe Russian

Current state, CSS

korpus [N.; m.]
korpusy to elektroniczne zbiory autentycznych tekstów, utworzone według uprzednio określonych zasad i w określonym celu oraz wyposażone w narzędzia umożliwiające wielowarstwowe wyszukiwanie danych językowych. (Korpusi so elektronske zbirke avtentičnih besedil, nastale po vnaprej določenih merilih in z določenim ciljem ter opremljena z orodji, ki omogočajo večplastno iskanje jezikovnih podatkov.)
[źródło: Gigafida]
zbiór tekstów w formie elektronicznej z zestandaryzowaną formą zapisu i wybranymi danymi (zbirka besedil v elektronski obliki s standardizirano obliko zapisa in izbranimi podatki)
[źródło: SSKJ2]
besedilni korpus (sl) korpus (pl) corpus (en)
besedilni korpus [Adj. + N.; m.]
korpusy tekstowe (ew. w językoznawczym języku specjalistycznym także po prostu korpusy) to obszerne zbiory tekstów w języku naturalnym, zebrane w ustalonym okresie z

TEI file

```
<entry xml:lang="sl" xml:id="sl-1">
    <form type="lemma">
        <orth>korpus</orth>
    </form>
    <gramGrp>
        <pos>N.</pos>
        <gen value="masculine">m.</gen>
    </gramGrp>
    <sense>
        <def n="1">
            <cit type="translation" xml:lang="pl">
                <quote>korpusy to elektroniczne zbiory autentycznych tekstów, utworzone
                    według uprzednio określonych zasad i w określonym celu oraz
                    wyposażone w narzędzia umożliwiające wielowarstwowe wyszukiwanie
                    danych językowych. </quote>
            </cit>
            <cit type="original" xml:lang="sl">
                <quote>Korpusi so elektronske zbirke avtentičnih besedil, nastale po
                    vnaprej določenih merilih in z določenim ciljem ter opremljena z
                    orodji, ki omogočajo večplastno iskanje jezikovnih podatkov.</quote>
            </cit>
            <bibl>
                <ref target="http://gigafida.net/Support/About">Gigafida</ref>
```

Decisions

- we will leave **separate dictionaries**,
- no ontology and **no links** (no pointing) to the entries between the dictionaries,
- probably we will prepare **a list of Polish (English?) equivalents** that are to be incorporated into entries,
- there will be extracted and prepared Polish-other languages **glossary** with links at the end to facilitate comparing the terminology between languages,

Changes in TEI files

- `TEI Header` completed;
  - added licence information, publisher etc.
  - `respStmt` corrected (editor and translator)
- added `xml:id` to the entries
- `gramGrp` corrected: added `gen` with values
- forms corrected, made distinction between lemmas and phrases
- prepared sample entries for the editors of the other languages dictionaries so they can copy-paste them and fill in in most cases

XPath, searching

Example #1

```
//entry[.//gen="n."]//orth
```

searching for the headwords (orths) of the entries containing grammar information with gender value „n.” (neuter)

Description - 2 items	XPath location	Resource
korpusno jezikoslovje	/TEI[1]/text[1]/body[1]/entry[4]/form[1]/orth[1]	terminologa_kc
označevanje	/TEI[1]/text[1]/body[1]/entry[16]/form[1]/orth[1]	terminologa_kc

Example #2

```
//form[@type="phrase"]//orth
```

searching for the headwords (orths) containing phrases (not lemmas), the resuts are eg. *besedilni korpus*, and not *označevanje*

To discuss with the team

- what bibliographic information are we going to preserve within entries,
- whether to have definitions in original and translated form or only translation,
- whether to prepare the Polish (English?) list of terms to function as a kind of an ontlology,
- how to present data on a website

Lemmas and phrases