IB047

From Corpus to Dictionary

Pavel Rychlý

pary@fi.muni.cz

April 1, 2022

Formats for storing dictionary entries

- text files vs. database
  - database contains entries
  - each entry in text format
- XML, JSON, NVH
- tree structure
- schema

XML

- TEI Guidelines
- TEI Lex-0
  - A baseline encoding for lexicographic data
  - Europian projects: ENeL, ELEXIS, DARIAH
  - https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

```
<entry type="mainEntry" xml:lang="cz" xml:id="en000008">
  <form type="lemma" xml:id="en000008.hw1">
     <orth>abeceda</orth>
  </form>
  <pc>,</pc>
  <form type="inflected">
     <gramGrp>
        <gram type="case" value="genitiv"/>
        <gram type="number" value="singular"/>
        <gram type="gender" value="feminine"/>
     </gramGrp>
     <orth extent="suffix" expand="abecedy">-y</orth>
  </form>
  <!--...-->
</entry>
```


JSON

- JavaScript Object Notation
- data interchange format
- attribute–value pairs and arrays

JSON – example

```
{ "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```


NVH: Name-Value Hierarchy

- less verbose alternative to XML
- name, value, list of child elements

```
headword: bank
  partOfSpeech: noun
  definition: an institution where you store or borrow money
    translation: банка
    example: I got a large loan from the bank.
      translation: Я получил крупный кредит в банке.
  definition: a stretch of land along a river
    translation: беper
    example: The house is on the north bank of the river.
      translation: Дом находится на северном берегу реки.
```

NVH: Compare to XML

```
<entry>
  <headword>bank</headword>
  <partOfSpeech>noun</partOfSpeech>
  <sense>
    <definition>an institution where you store or borrow money</definition>
    <translation>банка</translation>
    <exampleContainer>
      <example>I got a large loan from the bank.</example>
      <translation>Я получил крупный кредит в банке.</translation>
    </exampleContainer>
  </sense>
  <sense>
    <definition>a stretch of land along a river</definition>
    <translation>берег</translation>
    <exampleContainer>
      <example>The house is on the north bank of the river.</example>
      <translation>Дом находится на северном берегу реки.</translation>
    </exampleContainer>
  </sense>
</entry>
```


Informations from Corpora

- multi-word expressions (MWE)
- collocatons
- thesaurus
- domains
- examples
- translations

Examples

A good example must be:

- typical, exhibiting frequent and well-dispersed patterns of usage
- informative, helping to elucidate the definition
- readability
  - intelligible to learners,
  - avoiding gratuitously difficult lexis and structures, puzzling or distracting names, anaphoric references,
  - can be understood without access to the wider context.

GDEX – Good Dictionary EXamples

- sentence length: 10 – 25 words, longer/shorter penalized
- word frequencies: non common words (top 17,000) penalized
- pronouns and anaphors penalized
- target collocation in the main clause preferred
- whole sentence: beginning with a capital letter and ending with .!?

Domains/usage

- *usually spoken*, *business*
- domain annotation in corpus
- *only in plurals*
- condition specificaton + threshold
  - values of a structure attribute
  - subcorpus
  - query (tags)

```
=plurals
HR plural
Q1 [lempos="%s" & tag="NN2"]
Q2 [lempos="%s" & tag="NN1"]
RE -n$
```