

Developing a Lexicographic and Publishing Platform for the Serbian Language

The experience of foreign language teaching has shown that a significant number of students never manage to increase their spoken and written fluency beyond the intermediate level, despite taking intermediate and advanced courses.¹ Reading authentic, unabridged, and “difficult” literary texts can be extremely frustrating if the student lacks the vocabulary or sufficient knowledge of complex syntactic structures and cultural references. Reading simplified, abridged versions of literary works, on the other hand, is often an insult to the student’s intelligence.² Annotated editions of authentic literary texts written in the Serbian language, which are easily available on the Internet, would greatly improve and enrich the learning process of Serbian as a foreign language, while possibly attracting a larger number of students to the study of South Slavic literatures and cultures.

The author describes a scalable, web-based, digital platform for publishing annotated, fully-glossed study editions of literary works in the Serbian language through the use of non-proprietary formats and interoperable standards such as XML³, TEI⁴ and Unicode⁵, together with an integrated, collaborative, WordNet-based⁶ bilingualized Serbian-English dictionary. The integrated approach to textual and lexicographic production provides a set of tools for intermediate or advanced students, which facilitate reading, comprehension and discussion of authentic literary texts while encouraging students to become producers of knowledge in a global learning network.

The publishing platform allows editors to work with XML documents and establish links between individual words in a text and corresponding dictionary entries. At the same time, the editing process is rooted in the lexico-semantic approach to language learning by stressing the importance of units larger than the single word (polywords, fixed and free collocations, institutionalized expressions etc.). When published (through the use of XSL and CSS stylesheets) as valid HTML documents viewable in standards-compliant web browsers, texts contain two types of glosses: direct dictionary glosses, which are stored centrally in the dictionary database, and annotations, which are text-specific. Both are easily available to the user with a familiar interface method of clicking on individual words⁷ and alt-clicking on highlighted phrases⁸.

Although based on the English WordNet, the proposed lexicographic platform extends the Serbian-English semantic ontology by providing methods of encoding a host of important lexicographic properties which are aimed primarily at foreign students: detailed grammatical information, full accented declensions and conjugations, usage labels (dialectological, temporal, functional), and, when appropriate, stylistic distinction among members of a single synset. At the same time, however, the dictionary – like any WordNet-based semantic database – serves as a platform for exploring semantic relations between words and their senses (synonyms, antonyms, hyponyms, hypernyms etc.) The modular project architecture will allow further extensions such as introduction of etymologies, morphogenetic representations of lexical meaning (i.e. morphosemantic links between dictionary entries based on their constitutive roots), valency frames, visualizations of semantic hierarchies and so forth.

¹ Felicity O'Dell, Teaching advanced learners, <http://uk.cambridge.org/elt/teachers/youngadult/articles.htm>

² See also Gillian Lazar, Literature and Language Teaching, Cambridge 1993; Widdowson, H. Stylistics and the Teaching of Literature, Longman 1975.

³ <http://www.w3.org/XML/>

⁴ <http://www.tei-c.org/Guidelines/P5/>

⁵ <http://unicode.org/>

⁶ <http://wordnet.princeton.edu/>

⁷ A single-click on a word shows the dictionary gloss:

и белином ишчезло је и мало гробље
јом **вире** из дубоког снега. Једино ту се
зц **вирити**
кра делимично се видети, помаљати се, штрчати
џе **to protrude**
жа рана у општој белини која се
риметно у сивој пустињи неба још увек

⁸ Annotated phrases are highlighted when the mouse goes over them:

е слегла, и запута, али одмах настави да кара м
ек : **нисити Растислав, него Распислав!** Ни име
луги. Лок су се фратри звали фпа-Марко, фпа-
е слегла, и запута, али одмах настави да кара м
ек : **нисити Растислав, него Распислав!** Ни име
луги. Лок су се фратри звали фпа-Марко, фпа-
е слегла, и запута, али одмах настави да кара м
ек : **нисити Растислав, него Распислав!** Ни име
луги. Лок су се фратри звали фпа-Марко, фпа-

Alt-clicking an annotated phrase displays the annotation:

слегла, и запута, али одмах настави да кара младића.
к : **нисити Растислав, него Распислав!** Ни име ти, болан,
УТ **нисити Растислав, него Распислав**
је The root of the name Растислав stems from **расти** → **to grow**
'рсе and **раст** → **growth**. Распислав is not a real name, but a word
play with the prefix **паз-**, which usually denotes division into
two or more parts or dispersion and scattering, i.e. the opposite
of growth. Cf. **расипник** → **squanderer**; **распикућа** →
spendthrift; **распикућство** → **wastefulness**.