TECHNOLOGIZATION OF ROMANIAN: LINGUISTIC RESOURCES, APPLICATIONS, TOOLS

Boian Elena, Ciubotaru Constantin, Cojocaru Svetlana, Colesnicov Alexandru,

Demidova Valentina, Malahova Ludmila

Institute of Mathematics and Computer Science of Academy of Sciences of Moldova, 5, Academiei str, MD2028, Chisinau, MOLDOVA, email: chebotar@math.md

ABSTRACT

In this paper we present the Reusable Resources for Natural Language Technology (RRNLT). The RRNLT contain: Tools for access (use) of Resources, Tools for maintenance (support) of Resources, A base Romanian lexicon of 70 – 100 thousands words, Morphological and base syntactic information for each entry in the lexicon, Synonyms and thesauri, Translations (in Russian and English, in this project), Model programs as examples of Resource use (spelling checker, thesaurus, translation dictionary and hyphenation support, prototype teaching and testing program for Romanian orthography and morphology)

Keywords:. Natural Language Technology, linguistic resources, morphological and syntactic information, Romanian lexicon, translation dictionary, thesauri, spelling checker.

In the present-day world, linguistic engineering is essential to promote linguistic and cultural diversity in the European Union. Technologization of a language, automatic processing and especially possibilities of automatic translation would assure the compatibility of the language used by a smaller national community with the languages widely circulated and the maintaining of the linguistic identity in the conditions of the globalization of the contemporary society.

Technologization of the Romanian language becomes obvious that any effort in this direction would finally contribute to the enhancement of communication between the Romanian language users (therefore belonging to countries as Romania, Republic of Moldova, Ukraine etc.) and the EU countries. In this moment there are very few linguistic resources of Romanian available in an electronic format but the corresponding works are under development in Romania and in the Republic of Moldova.

1. DATABASE STRUCTURE

Our core resource is a database. We can get from the database, using SQL queries, any information in any order and with any desired set of attributes and characteristics.

The database contains now 21 tables. The main three tables contain all base words for the Romanian, English, and Russian languages, correspondingly. They uniquely encode all words. For homonyms and other needs, words can be doubled. Encoded part of speech and usage area are also kept here. A special field word_sort contains words in some inner code for sorting. We do not use the standard sorting by the code page. 10 tables encode morphological_attributes codes of 97 grammatical attributes (1 – feminine gender, 2 – masculine gender, 10 – singular number, 11 – plural number, etc.).

Only base words are kept in the table words that is necessary for some support programs. All flexions are kept in the table word_flexies. The base form is also a flexion; therefore, it is included here.

Synonyms are kept in the table word_synonyms. This table contains information on the base word and its synonyms as pairs of references (codes) from the table words.

The table word_translations contains codes of the base Romanian word, language and the corresponding translation.

2. XML REPRESENTATION

Reusability of linguistic resources can be achieved only if the data and its annotations are describable using a common data model. The eXtensible Markup Language (XML) was selected as the basis for a standardized encoding format. Having XML enables use of powerful mechanisms from the XML framework, e.g., the XSLT Transformation Language. XSLT supports such kinds of document manipulation as: selection of elements or their portions; rearrangement or transformation of extracted information; addition of information in the target document. Therefore, an XML representation of our database can be manipulated for any application that uses part or all of its contents by researchers that are interested in Romanian lexicology and morphology, or in mapping Romanian words into English or Russian lexicon. XML representation of the database can be obtained and downloaded from the

3. WORD-LEVEL ROMANIAN RESOURCES – MORPHOLOGY

We developed a word inflection program that generates word-forms (flexions - 12 for nouns, 20 for adjectives, 35 for verbs) and shows all generated flexions on the screen. The user can correct them before writing them to the vocabulary database. We grouped registered affixes for noun declination in 32 groups and have based the declination algorithm on this grouping. In some cases the group for the given noun cannot be selected automatically, and we ask the user to select

the group giving him/her examples. Some irregularities in letter alterations also imply a dialog with the user. Adjectives were classified in 26 groups. They have irregularities also.

Statistical estimations indicate that 88% of Romanian nouns and adjectives can be declined automatically and only 12% need a dialog.

The generation of verb word-forms is based on the infinitive of the verb. Generating the inflectional paradigm for the verb is the most complicated part of the program but we could automate this process fully. Verbs were divided into 5 types having their characteristic features. Each type is divided into some classes, characterized by affixes. Each class in its turn is divided into schemes, depending on specific letter alteration, sub-variants of affixes, word roots, etc. We founded in total 56 schemes for Romanian verb conjugation.

The described methods were used to create a Romanian word-form database that contains currently approx. 1,000,000 word-forms and is used in our Romanian spelling checker.

4. SYNONYMS

The main lexical source is a well-known Dictionary of Romanian Synonyms by Luiza and Mircea Seche of 44,240 words (electronic variants granted by Prof. Dan Cristea and the publishing house Litera International).

Operational unity is an entry in both synonym and translation dictionaries. Each entry has two components, the initial word and a list of synonyms. Lists of translations can be divided into descriptions and equivalents. Descriptions are defined completely but equivalents are in fact synonyms and are presented by a reference in the synonym dictionary. In the synonym dictionary, we use the fact that synonyms are also dictionary entries. We do not need therefore to include them several times and can use references instead.

A vector representation of the synonym dictionary consists of a matrix in which each line is an initial word and vector of references to synonyms. If some lists of synonyms contain common words in them, the corresponding reference vectors contain equal references.

5. TRANSLATIONS

Translations of Romanian words into Russian and English are provided. The user can get translations for all words that are like the given template.

6. SYNTAX

Working at the word level, it is possible to give only limited information on syntax indicating potential syntactical roles of word-forms that correspond their parts of speech and morphological attributes. E.g., the noun in the nominative case can be subject of the proposition, etc.

7. SUPPORT SYSTEM

A program complex was developed to populate the database. The database population programs provide five modes for this. The word-forms and the morphological information can be entered from files in the DOS and Windows encodings, and a module of entering just one base word with all its forms was also developed. The separate modules were developed to enter synonyms and translations. The programs analyzed the source text files, formatted the information as SQL queries for the database population and encoded the Romanian and Russian words into UTF-8. Then these queries were executed over the database. The database population programs were developed in Delphi using DLL and VBA for MS Word.

The interface for visualization of the database contents was developed using SQL queries to get information from the database, and PHP modules to produce HTML pages dynamically. Any Web browser can be used as visualization engine (MS IE and Mozilla Firefox were tested).

The main HTML page proposes selection of several demonstration modes: morphological information, word-forms, synonyms, English and Russian translations. The user can enter a word or a word template using wildcard characters '*' and '?'. The information is selected from the database form all words that satisfies the template. If the result uses several pages, the navigation through the pages in sequential or arbitrary order is provided.

8. EXAMPLE APPLICATIONS

The project includes creation of example applications using the developed database:

- analysis and spelling correction of a text (including all specific functions of a spelling checker: error detection; suggestions; dictionary completion; search for a word; search for similar words; lexicon packing; deletion of a word from the lexicon);
- synonym dictionary. The Romanian synonym dictionary was implemented as a set of HTML pages each containing twenty dictionary entries. There is on each page a page selector implemented as a drop-down list with the first and the last word on each page;
- bilingual dictionaries (Romanian-Russian, Romanian-English). They are implemented analogously as synonym dictionary;
- word derivation and morphological information visualization;
- prototype teaching and testing program for Romanian orthography and morphology. For its developing the free platform Claroline was used. The prototype contains two basic components: for teachers to assist them in lectures preparation and for students to guide and help them to study Romanian.

This work is a step to the entering of the Romanian language in the common computerized space of European languages.