Computer Science Journal of Moldova, vol.**14**, no.1(40), 2006

Integrity and correctness checking of a lexical database*

S. Cojocaru A. Colesnicov L. Malahova

Abstract

A Romanian lexical database being the core of the Romanian Reusable Resources for Natural Language Technology should be thoroughly checked for integrity, correctness, and completeness before to be made widely available. This case study is presented.

1 Introduction

The core of Language Engineering applied to language understanding and generation resides in the acquisition of sufficient resources in the languages to be treated, which are used to provide morpho-syntactic, lexical and semantic information, as necessary for grammar development, statistical data for language models, etc.

Aiming this, the Romanian Reusable Resources for Natural Language Technology [1, 2] (the Resources) were developed. The Resources consist of a database with the linguistic information for Romanian at the word level, and a set of service programs. Extraction of linguistic information from the database can be done by formulating SQL queries.

Depending on user's command level in Romanian, it is possible to develop different applications based on the Resources for non-Romanian speaking users, ordinary Romanian speaking users, and expert users.

Applications dedicated to the non-Romanian speaking user may include a kind of e-learning system for Romanian morphology. The DB and its existing viewers may be used by international students and

*Supported by INTAS project Ref. Nr. 05-104-7633

language minorities in Moldova and Romania as directory in Romanian morphology.

Applications dedicated to the usual Romanian speaking user are Web interfaces intended to allow the use of morphological variations of Romanian words and synonyms in standard Web browsers. Web-service of spelling checking may be also developed.

Expert users of Romanian language can use the Resources to support dictionary development, including advanced lexicographic operations and support of complex browsing among dictionaries of different types. The DB with its programming possibilities seems to be a flexible and powerful tool for these operations.

Before to make our Resources widely available we should pass the stage of their correctness and integrity checking.

As the volume of the Resources is as big as hundreds thousand or even millions items, we should check their correctness and integrity in maximally automated mode. More precisely, we should use some programs to select suspicious information and to propose to the operator or the expert in philology to make the final decision.

The list of suspicious information items should be as short as possible for better reviewing by a specialist.

This article describes our approach to the checking of integrity and correctness of the lexical resources presented as a DB containing Romanian words, their morphological derivatives, synonyms, English and Russian translations, etc.

In Section 2, we shortly describe the DB structure. This section is mere technical and gives the presentation of information we work over.

Not only the information structure, but the methods the information for the DB was obtained influence the techniques of its checking. That's why we discuss methods of DB population in Section 3.

By applying automated methods, we can reveal only part of errors and suspicious items. The visual checking performed by an operator remains an important verification method. In Section 4, we describe the DB visualization tools (viewers) that were used for this.

In Section 5, we discuss some techniques we used to check the DB integrity and information correctness.

2 Database structure: main and auxiliary tables

The Resources DB has six main tables and a lot of auxiliary tables. Auxiliary tables contain different codes used in the main tables, e.g., codes of morphological characteristics or languages.

Six main tables are words, words_engl, words_rus, word_flexies, word_synonyms, word_translations. The former three tables map, correspondingly, Romanian, English, and Russian word to the numerical codes. These numerical codes are used in the latter three tables instead of textual word presentation. E.g., the word_synonyms table contains synonym pairs that consist of two numbers of the corresponding Romanian words from the words table.

There is other necessary information in these tables. Some examples follow.

The words table contains the part_code (part of speech) and field_code (domain of the word usage) fields.

Numerically encoded word/translation pairs in the word_translations table are marked by the language code to distinguish English and Russian translations.

The word_flexies table contains the flexy_word character field keeping derivatives of Romanian words. Each derivative is associated with its lemma in the words table through the integer prim_word_code field. The integer morpho_code field substantiates morphological information (tense, number, case, etc.).

As for auxiliary tables, the morpho_code field is substantiated using not one single table but ten auxiliary tables in correspondence to ten Romanian parts of speech: noun, adjective, verb, numeral, adverb, pronoun, preposition, conjunction, article, interjection. These are tables named noun_part_speech, adjective_part_speech, verb_part_speech, etc. The fields in these tables contain codes of Romanian morphologic categories corresponding to the part of speech.

You can see more detailed list of tables and their fields in Section 5, Tab. 1.

3 Database population

The DB population is one of the most important part of such project development. We took into account the request of highest quality of the DB population and decided therefore to populate the DB programmatically from textual information files.

For morphological information, we used a set of log files produced under our precedent projects [5]. Information for translations and synonyms was taken from different lexicographical sources [3].

First of all, a uniform format for data input was developed, and existing data files were transformed to this format.

Data input files for DB population consist of line groups.

For the word_flexies table, each group contains one word-lemma with all its derivatives (word-forms). Encoded morphological information is included with each word-form. Part of speech and domain of usage is included with each word-lemma.

For synonyms, each group contains a main word and its synonyms; figurative synonyms are marked.

For translations, each group contains a main word and its translations into English or Russian. Information on the domain of usage is attached to the main word.

The DB population program produces log that shows if words were inserted, shows word codes, and the result for each operation. Errors are marked and can be easily found. We also see how many words were entered and which words were not entered because they double the existing in the DB ones.

Another tool for DB population with morphological information is a semi-automatic program that generates all word-forms for a given Romanian words. The program is wizard-like and the input should be done by an expert linguist.

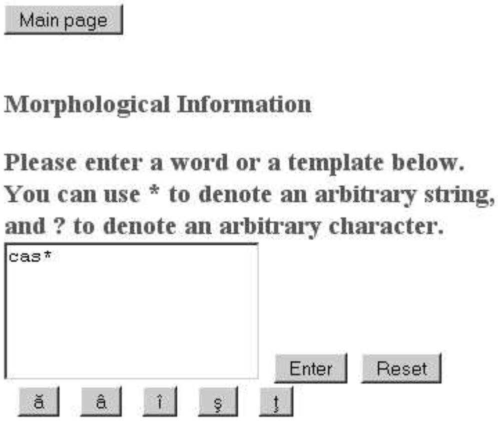
4 Viewers

Several viewers were programmed to check the DB visually, and to demonstrate possibilities and information contained in the DB.

We have five viewers: morphological characteristics viewer, wordforms (derivatives) viewer, synonyms viewer, English translations viewer, and Russian translations viewer.

All viewers have two common parts. The first common part is the input form. The user can ask not only for one word but he/she can use a regular expression with '?' meaning an arbitrary character and '*' meaning an arbitrary string (may be empty). The input form contains also five buttons that produce Romanian letters with diacritic for the case of absence of Romanian keyboard layout.

Fig. 1 shows the input form with the request to search morphological characteristics of a group of words.


Morphological Information

Please enter a word or a template below.
You can use * to denote an arbitrary string,
and ? to denote an arbitrary character.

cas*

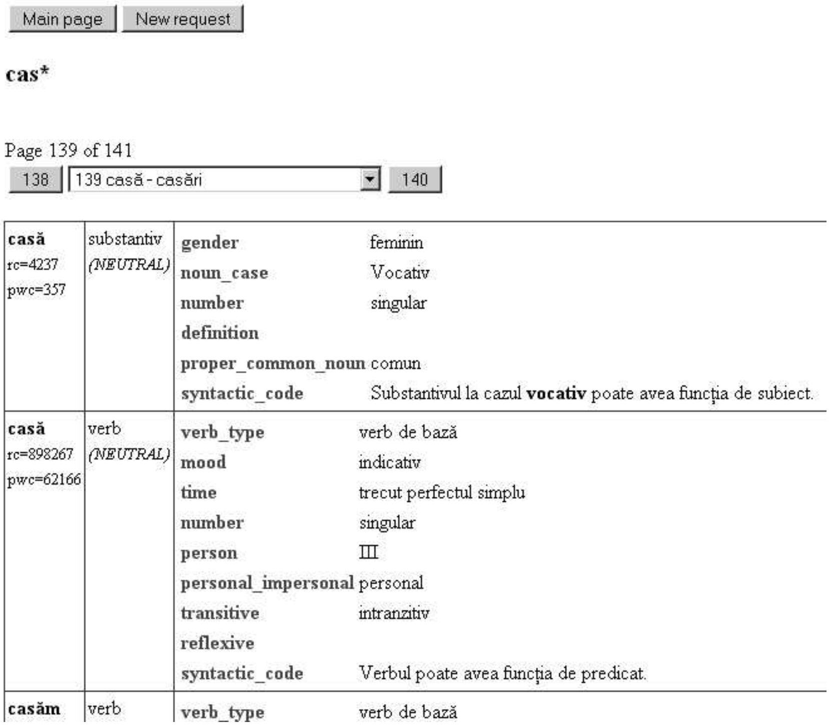
Figure 1. The input form for morphological characteristic search

The second common part of all viewers is called 'the pager' and controls the distribution of the DB data on pages (at 5 blocks per a page), page selection by its number, switching to the next or previous page, etc.

The morphological characteristics viewer queries the word_flexies table for the given word(s) and shows morphological attributes for each matching word.

The word-forms viewer queries the words table for the given word-lemma(s), then selects all derivatives of each word from the word_flexies table and shows them with their morphological attributes. The synonyms, English translations, and Russian translations viewers work analogously over the corresponding tables.

Fig. 2 shows a page produced by the morphological characteristics viewer. We asked for the regular expression 'cas*' that matches 702 derivatives in the word_flexies table.



cas*

Page 139 of 141

casă
rc=4237
pwc=357
	substantiv
	(NEUTRAL)
		gender	feminin

		noun_case	Vocativ
		number	singular
		definition	
		proper_common_noun	comun
		syntactic_code	Substantivul la cazul **vocativ** poate avea funcţia de subiect.

casă
rc=898267
pwc=62166
	verb
	(NEUTRAL)
		verb_type	verb de bază

		mood	indicativ
		time	trecut perfectul simplu
		number	singular
		person	III
		personal_impersonal	personal
		transitive	intranzitiv
		reflexive	
		syntactic_code	Verbul poate avea funcţia de predicat.
casăm	verb	**verb_type**	verb de bază

Figure 2. A page of found morphological characteristics

5 DB integrity

The building of a lexical resource is a difficult process. We tried to automate it maximally using specially developed programs. Meanwhile, at least three sources of errors remain that can influence the final result, namely:

- errors in the used lexicographical sources;
- errors in programs processing lexicographical information;
- operator's errors.

We can therefore suppose that each field of our DB can be potentially erroneous. We can not solve our task of information verifying using only software tools. E.g., it is impossible to decide if one word is a real synonym or translation of another word without consulting with an expert in philology. Meanwhile, it is possible to develop a set of techniques that solve this problem partially. The techniques developed for our case are described below.

5.1 Formal DB validity

First of all, we can apply formal methods to check validity of the DB structure. These methods can be formulated using the semantics and interdependencies of the DB fields and tables. All DB fields are divided for it in four categories:

1. fields containing textual representation of words (in our case, Romanian, English, and Russian);
2. fields containing references that connect different tables, e.g., numbers of Romanian words that replace words themselves in the word_synonyms table;
3. fields containing morphological and other attributes;
4. fields containing textual representation (deciphering) of attributes; these fields exist only in the auxiliary tables.

Fields of our main tables are listed in Tab. 1. For each field, its category and the method of its formal checking are shown.

Depending of the used DB engine, some formal relationships can be supported automatically.

5.2 Checking of words

Non-formal checking may be executed by variety of techniques depending on the field category. E.g., category 1 fields can be checked by usual spell checkers. For Romanian, we have our own spell checker RomSP [5]. The corresponding list of Romanian words was carefully tested and updated by developers and many users of the product, and we can take it as being quite reliable. We used also Romanian, English, and Russian spell checkers from MS Office. For Romanian, we marked words that were rejected by both spell checkers as highly suspicious. The analysis show that most of them were erroneous.

We can use other word lists, e.g., those coming with free spell checkers like ISpell [6].

A different method of word checking supposes the selection of n-grams (word fragments of n letters, $n > 2$) from the given set of words, and calculation of their frequencies. Less frequent n-grams are considered to be suspicious. Words that contain such n-grams should be checked by experts.

5.3 Checking of attributes

We saw that category 3 fields can be formally checked as containing in one of additional tables as the record number. The correspondence between fields of categories 3 and 4 can be checked informally using interval of values for different attributes but this is partial checking only. In any case, additional tables are short and can be checked visually. We can also search for unused codes in them. The correspondence of codes in the morpho_categories table and tables for each part of speech was checked by issuing requests that show in parallel decoded values of each code.

Table 1. Formal checks in the main DB tables

Table	Field	Cat.	Check method
word_flexies	prim_word_code	2	This field should be a record number in the words table.
	flexy_word	1	This field should be non-blank; only Romanian letters are permitted.
	morpho_code	3	The check is made through the chain words.prim_word_code → part_of_speech → record_number in the corresponding table.
word_synonyms	prim_word_code	2	This field should be a record number in the words table.
	synonym_code	2	This field should be a record number in the words table.
	figurat_code	3	0 or 1; shows if the synonym has figurative meaning.
word_translations	lang_code	3	Record number in the languages_table; ≠Romanian.
	prim_word_code	2	Record number in the words table.
	translation_code	2	Record number in the word_engl or word_rus depending of language.
	figurat_code	3	0 or 1.
words	prim_word_code	2	This code should be used in the word_synonyms, word_translations, or word_flexies tables.
	part_code	3	Record number in the part_code table.
	word_ii	1	Non-blank field; only Romanian letters permitted.
	word_aa	1	Non-blank field; only Romanian letters permitted.
	field_code	3	Record number in the field table.
words_engl	record_code	2	Used in translations with lang_code=English only.
	word_engl	1	Non-blank field; only English letters, spaces, and punctuation (e.g., apostrophe) permitted.
words_rus	record_code	2	Used in translations with lang_code=Russian only.
	word_rus	1	Non-blank field; only Russian letters, spaces, and punctuation permitted.

5.4 Checking of references

The next category of checks is search for duplicates. Our DB population programs query for existence of the information before its insertion into any of tables, therefore, absence of duplicates can be supposed. Meanwhile, search for duplicates can expose some errors in the prepared data for population of the DB, or in DB population programs themselves.

In the words table the unique field is prim_word_code. The corresponding information consists of the Romanian word in its textual form, its part of speech and field of usage. These data are checked for uniqueness during DB population. Non-unique combination found means something wrong with these programs, and we can check their logs visually for this combination.

We do not enter specific field of usage for a word where we enter its morphological derivatives. In this case, the corresponding field is always set to 1 ("general"). Therefore, we can check for uniqueness of the combination of a word's textual form and part of speech and analyze the corresponding fields of usage and tables where are used "non-general" words. We created the list of uninflected words that coincide with some inflected pairs of text and part of speech, and the list of "truly" uninflected words. We discovered several cases when the information for synonyms and translations was erroneous.

Moreover, we checked the words table for uniqueness of word's textual form ignoring even its part of speech. In Romanian, adjective can coincide with adverb and noun can coincide with adjective, but such cases are relatively rare. It differs from English where the same word can be verb or noun as a rule. This check permitted to detect several errors also.

Uniqueness of records in the word_flexies, word_synonyms, word_translations, word_engl, and word_rus tables is also checked during DB population. The corresponding check can be performed after population to test the DB population programs.

5.5 Statistics of derivatives

We performed also the following informal semantic checks.

Normally, Romanian words have some standard number of inflective derivatives depending of the part of speech, e.g., 35, 39, or 40 for verbs, etc. We queried for the actual number of derivatives for words from the words table. Fig. 3 shows the result of the first such test in the aggregated form. part_of_speech= 1 means 'verb', etc. You see in the graph, e.g., one verb with 160 derivatives.

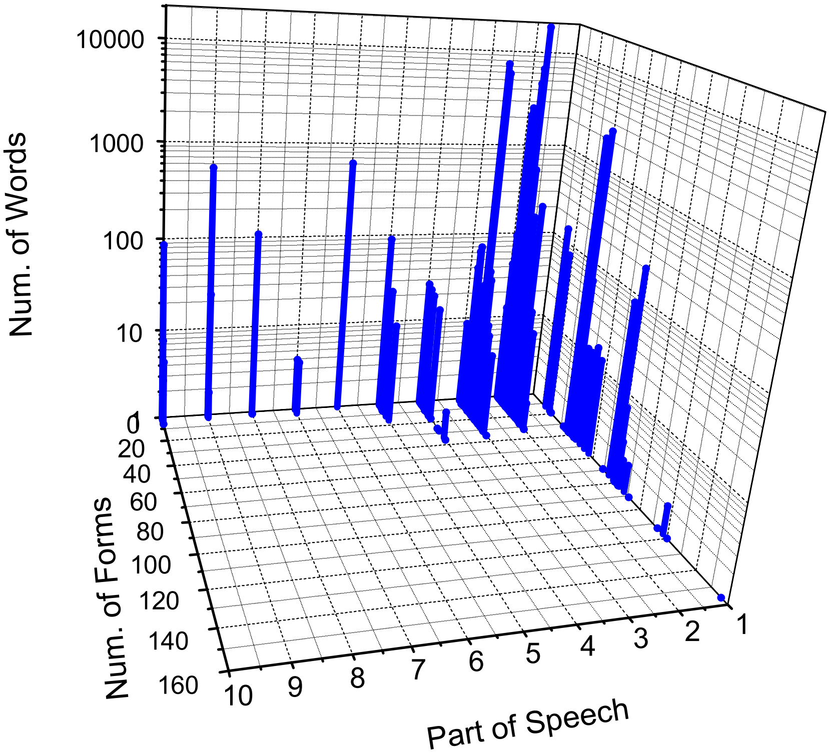


Figure 3. Statistics of derivative number

The unsuspected number of derivatives for some words permitted us to correct some errors. E.g., it was found analyzing the case of verbs with more derivatives than necessary that some details of Romanian grammar were misunderstood during the design stage.

5.6 Checking through parallel dictionaries

Parallel dictionaries are very useful and widely used in computer linguistics (see, e.g., [4]). Our DB contains translations of many Romanian words into English and Russian. We could not get sufficient results from the English translations. The Russian translations permitted us to formulate several useful criteria because Russian is a highly inflective language like Romanian.

We used endings of Russian translations, that are more or less standard depending of part of speech, for:

- Check for words that are not verbs but Russian translations have "verbal" endings -ти -тись -ть -ться -чь -чься. We found 4119 of them, being mostly OK, but several errors were found.
- Check for words that are not adjectives but Russian translations have endings -ая -ев -ий -ин -ов -ое -ые -ый -ье -ья. No such words were found.
- Check for words that are not adverbs but Russian translations have endings -е -о -у -ем -ём -мя -ой -ом -ски. This check was not so successful (18974 words) but we shortened the result by deleting all verbs, adjectives, and nouns, and found several errors more.

5.7 DB completeness checking

The following test can be proposed to check the DB completeness. Having a list of Romanian words from any source, it is possible to sort it and to compare with the sorted list of words from the words or word_flexies tables. The word_flexies table should be used if word derivatives are permitted. This technique finds words that do not exist in our DB, and we can add them.

5.8 Correction of errors

As errors were found, they were corrected in the source data files. At a small quantity of corrections, erroneous records were deleted taking into account all interdependencies, and the corresponding part of the data file was entered anew. Having a lot of corrections, we populated anew the whole DB (six main tables) that takes quite acceptable time.

6 Conclusions

We selected the DB as linguistic information stock because of possibility of quick parallel and distant access, flexibility of possible queries, wide use and availability of the corresponding programming techniques. Other forms of information presentation like, e.g., word lists, can be easily obtained from the DB. Applications can be developed using our DB directly or indirectly.

The information containing in the DB should be thoroughly checked using different techniques. We proposed a set of methods that were found useful in our case. The discussed techniques can be applied at checking of lexical information in other cases.

Acknowledgements

The work was supported by the grant MM2-3042 of CRDF/MRDA.

References

[1] E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, L. Malahova. *Lexical resources for Romanian.* In: Scientific memoirs of the Romanian Academy, ser.IV, **vol. XXVI,** Bucharest, Romania, 2005, pp. 267–278.

[2] E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, L. Malahova. *Lexical Resources for Romanian – a project overview.*

In: Proceedings of Symposium on Intelligent Systems and Application, September 19–20, 2003, Iaşi, România, 12 pp. – ISBN 973–97737–2–9.

[3] E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, L. Malahova, T. Verlan. *Elaboration of tools to support an electronic dictionary of synonyms and transaltions.* In: International Conference "Trends in the Development of the Information and Communication Technologies in Education and Management", March 20–21, 2003, Academy of Economic Studies, Chişinău, Republic of Moldova, pp. 175–177. – In Romanian.

[4] D. Tufiş and A.M. Barbu. *Pevealing Translator's Knowledge: Statistical Methods in Constructing Practical Translation Lexicons for Language and Speech Processing.* International Journal of Speech Technology **5**, 2002, pp. 199–209.

[5] L. Malahova, A. Colesnicov. *Implementation of the Romanian Spelling Pack for Windows.* In: The International Conference on Technical Informatics CONTI'96. Proceedings. Computer Science and Engineering, **vol. 1**, Timişoara, România, 1996, pp. 23–28.

[6] `http://www.gnu.org/software/ispell/ispell.html`

S. Cojocaru, A. Colesnicov, L. Malahova,

Received March 10, 2006

Institute of Mathematics and Computer Science,
5 Academiei str.
Chişinău, MD−2028, Moldova.
E–mail: *sveta@math.md*, *kae@math.md*, *mal@math.md*