Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries

Ana-Maria Barbu
Institute of Linguistics, Romanian Academy
Calea 13 Septembrie nr. 13, 050711, Bucharest, Romania
E-mail: anabarbu@unibuc.eu

Abstract
This paper presents two lexical data bases for Romanian: RoMorphoDict, a dictionary of inflected forms and RoSyllabiDict, a dictionary of syllabified inflected forms. Each data basis is available in two Unicode formats: text and XML. An entry of RoMorphoDict, in text format, contains information on inflected form, its lemma, its morpho-syntactic description and the marking of the stressed vowel in pronunciation, while in XML format, an entry, representing the whole paradigm of a word, contains further informations about roots and paradigm class. An entry of RoSyllabiDict, in both formats, contains information about unsyllabified word, its syllabified correspondent, grammatical information and/or type of syllabification, if it is the case. The stressed vowel is also marked on the syllabified form. Each lexical data base includes the corresponding inflected forms of about 65,000 lemmas, that is, over 700,000 entries in RoMorphoDict, and over 500,000 entries in RoSyllabiDict. Both resources are available for free. The paper discribes in detail the content of these data bases and the procedure of building them.

1. Introduction
This paper presents two data bases, one of complete paradigms of Romanian words (a morphological dictionary called RoMorphoDict) and the other of syllabified (inflected) words (a syllable dictionary called RoSyllabiDict). Each data basis is available in two Unicode formats: text and XML. The main information of an entry in the morphological dictionary consists of an inflected word, its lemma and its morphological description. If a word has no inflected form, this position is occupied by the lemma form. An entry of the syllable dictionary contains an inflected form, its syllabified form and an observation field. The procedure of building them and the presentation of the results make up the content of this paper.

2. The utility of such dictionaries in NLP
A morphological dictionary can be used wherever a lemmatizer is needed. Explaining what a lemmatizer is good for would be a waste of time, because it is a real basic tool in NLP.

For Romanian, there are concerns to build rather morphological analyzers or generators, than such dictionaries of huge dimensions. We mention, in this sense, (Bîrlădeanu & Burciu, 2006) and (Dumitriu, 2006a, 2006b). The latter work uses the tool Unitex described in (Paumier 2006). Another work is a complex tool named RoLingva, which includes inflected lemmas, stress information and a morphological analyzer, but this is a commercial self-contained tool and cannot be used in NLP applications. A previous step in building a morphological dictionary is represented in (Ionescu, 2003). We think lemmatizers (analyzer or/and generator) for Romanian have two major challenges. They have to face, on the one hand, with a rich system of phonetic alternations and irregular forms, and, on the other hand, with the high degree of ambiguity given by the rich inflectional morphology of Romanian. Furthermore, they are strongly time-consuming. However, they have the advantage of treating unknown word. A morphological dictionary, instead, presents a high-level accuracy and it is much faster to use. Its weak points are, indeed, the unknown words.

With respect to syllables dictionary, it has an uncontroversial utility in speech research. For previous work in Romanian syllabification, see (Dinu, 2006).

3. The morphological dictionary: RoMorphoDict

3.1 Building procedure
The Romanian morphological dictionary RoMorphoDict is based on the printed dictionary that prescribes the correct writing, pronunciation and inflection of the Romanian words, known with the abbreviation DOOM (1989). It contains about 65,000 entries of words in contemporary Romanian lexicon, covering all parts of speech. It also provides combinations of words which induce writing difficulties, but these were ignored in our task.

Actually, for automatically building RoMorphoDict, we had at our disposal an electronic copy of DOOM and an explicit inventory of Romanian paradigms for nouns and verbs. We consider our paradigm inventory explicit because we have considered two paradigms to be different if they differ by at least one form. For each paradigm all the corresponding endings are mentioned.

An entry in DOOM, Fig.1, has the following basic structure, where POS means part-of-speech, MSD – morphosyntactic description and INF – inflectional form.

| lemma | POS | MSD1 | INF1 | MSD2 | INF2 |
|-------|-----|------|------|------|------|
| abandoná vb., ind. prez. 1 sg. abandonéz, 3 sg. şi pl. abandoneáz |

Figure 1: DOOM entry

This entry describes the verb (vb.) abandoná (the
Table 1: The Paradigm-DOOM Forms Correspondence

| Paradigm # | Indicative present |
|------------|-------------------|
|            | 1 sg   | 2 sg   | 3 sg   | 1 pl   | 2 pl   | 3 pl   |
| Paradigm endings | -éz | -ézi | -eáz | -ám | -áti | -eáz |
| DOOM forms | abandon-éz | abandonéáz | Abandonéáz |
Verbal moods:

| ind | conj | ger | imper | inf | part | part-adj |
|-----|------|-----|-------|-----|------|---------|
| indicative | conjunctive | gerundive | imperative | infinitive | participle | adjectival participle |

Verbal tenses:

| imperf | prez | mmperf | perf |
|--------|------|--------|------|
| imperfect | present | plusqueparfait | simple perfect |

Persons & Numbers:

| 1 | 2 | 3 | sg-pl |
|---|---|---|------|
| first person | second person | third person | singular and plural |

Combinations: 1sg, 1pl, 2sg, 2pl, 3sg, 3pl

Cases:

| nv  | a  | g  | vc  | d  | g-d  | n-a |
|-----|----|----|-----|----|------|-----|
| nominative | accusative | genitive | vocative | dative | genitive and dative | nominative and accusative |

Genders:

| m  | f  | n  | f-n |
|----|----|----|-----|
| masculine | feminine | neuter | feminine and neuter |

Noun and Adjective Article:

| art | neart |
|-----|------|
| with enclitic | without enclitic |

Pronominal Forms:

| acc | neacc |
|-----|------|
| stressed | unstressed |

Verbal Polarity:

| neg |
|-----|
| negated form |

Disjunction operator: / = or (e.g. s/adj = noun or adjective).

A MSD includes such labels joint by dots. Labels are unambiguous, so that their position in MSD is irrelevant.

Accents are not used in the Romanian writing. If one wants to apply the dictionary on Romanian written text, the stressed vowels in dictionary have to be changed with the corresponding unstressed vowels, in the following manner: á > a, é > e, ó > o, ü > u, å > a, ì > i.

3.2.2. The XML format

The XML variant of RoMorphoDict is more informative than the previous one. Besides the morphosyntactic description, it provides information about the paradigm number, about roots and about the correspondence between roots and inflected forms. In (2), a verb example is given, where the elements and attributes have the following meaning:

**entity** – is the XML entry of the dictionary describing the whole paradigm of a word. Its attribute **type** specifies the word part of speech.

**parad** – indicates the flexion class of the word, for verbs, nouns and adjectives.

**vform** – is the element containing the inflected form the morphosyntactic description of which is described by the attribute **mood** = verbal mood, **tense** = verbal tense, **pers** = person, **nr** = number, **pol** = polarity, **gen** = gender, **rid** = root identifier. The values of these attributes are labels presented in section 3.2.1.

**glos** – is a slot for different notes referring to the entry.

(2) <entity type="verb">
  <parad>a-10</parad>
  <vform mood="inf" rid="0">uscă</vform>
  <vform mood="ind" tense="prez." pers="1" nr="sg." rid="1">ușcă</vform>
  ……
  <root rid="1">uscă</root>
  <root rid="0">ușcă</root>
</entity>

Other examples of entries in XML dictionary are given in (3) for nouns and adjectives and in (4) for pronouns and determiners.

(3) <entity type="noun">
  <parad>nul-2</parad>
  <nform pos="s." gen="n." nr="sg." art="neart." case="n-a." rid="0">căsuș</nform>
  ……
  <root rid="0">căsuș</root>
</entity>

(4) <entity type="pronoun">
  <pform pos="det." gen="m-n." nr="sg." forma="acc" case="n-a." rid="0">acă</pform>
  ……
  <glos>determinator antepus</glos>
</entity>

Entities of type **noun** contain the elements **parad**, **nform**, **glos** and **root**, while those of type **pronoun**, **pform** and **glos**.

The elements **nform** and **pform** describe nominal flexion by the attributes **gen** = gender, **nr** = number, **case** = case. In addition, they have particular attributes and values:

**nform** – has the **pos** attribute with the values **s** or **adj** and the attribute **art** = article;

**pform** – has the **pos** attribute with the values **det** or **pron** and the attribute **forma** with the values **acc** or **neacc**.

XML entries for non inflectional parts of speech, namely adverb, preposition, conjunction and interjection have the
simple description in (5).
(5)  <entity type="adverb">
       <form>&#250;nđe</form>
</entity>

The proper part of speech is indicated as the value of the attribute type of the element entry.

3.2.3. Diacritics
The diacritics and vowels marked with an accent are represented in Unicode encoding, with the following decimal codes:
&\#225; = â – without accent becomes ‘a’.
&\#7845; = å – without accent becomes ‘å’ (&\#226;).
&\#237; = ĩ – without accent becomes ‘i’.
&\#250; = ũ – without accent becomes ‘u’.
&\#7855; = â – without accent becomes ‘ê’ (&\#259;).
&\#238; = ĕ – without accent becomes ‘e’.
&\#243; = õ – without accent becomes ‘o’.
&\#259; = ā
&\#226; = ā
&\#238; = ĭ
&\#315; = ş
&\#355; = ţ

For using the XML dictionary on Romanian written texts, one has to delete the accent marks as it was shown upper. For this XML description, there is already an interrogation tool on CD.

4. The syllable dictionary: RoSyllabiDict

4.1 Building procedure
Building the Romanian syllable dictionary was a continuation of the morphological dictionary by that the previously inflected forms were then syllabified. For syllabifying, we used the following resources:
• a program implementing Romanian syllabification rules;
• the syllabification information that DOOM provides;
• an inventory of Romanian diphthongs and triphones.

The critical points in (Romanian) syllabification are sequences of vowels which can be pronounced as diphthongs/triphthongs or hiatus. In many cases, the pronunciation type cannot be inferred from the context, see Dinu (2003). For some entries, DOOM specifies the vowels in hiatus, for example, like this: adăugă (sil. -da-u-). Sometimes this information is given only for lemma, sometimes only for some inflected forms. There are a lot of hiatus situations which are not specified in DOOM. This description inconsistency was a source of errors in automatic processing.

We have applied our procedure on forms without accent, because our syllabification resources were like that and Romanian writing does not mark accents. But it is worth mentioning that one can get better results if the syllabification procedure takes into account stress information, since this reduces the number of diphthongs/hiatus ambiguities. For instance, the sequence -ei- can be a diphthong or a hiatus, but -ei- is always a hiatus. We have done some post-processing improvements, related to accent information, as well as some partial checkings of work.

The syllable dictionary has now 525,530 entries, whose format is shown in next section.

4.2 Entry format
RoSyllabiDict is also available in two variants: one in a text format on three columns and one in XML format.

4.2.1. Three columns format
Entries on three columns have the following structure:

```
  WORD<tab>SYLLAB<tab>OBS
```

where WORD is the inflected form of a word, SYLLAB – the syllabified form of the word in first column and OBS – remarks in cases of ambiguity. The fields OBS can miss for words which are unambiguously syllabified (6a). Ambiguity can have two reasons: different pronunciation accents (6b) or different types of syllabification (6c).

The word in first column is not marked with an accent, because this is the form in which it appears in texts. Instead, the syllabified form, in the second column, bears an accent because syllabification can differ depending on the accent of the word. For instance, the written word acceptă (‘s/he accepts’) is ambiguous whereas the corresponding spoken one is not, because if the accent is on the final syllable the word is a verb in simple perfect (v.perf.) and if the accent is on the penultimate syllable the verb is in present (v.prez.) (6b). The syllabification makes this distinction of accent. Besides, different grammatical forms can imply different syllabified forms (see example 7c in section 4.2.2.).

(6)  a. accept ac-cépt
    b. acceptă ac-cép-tâ v.perf.
        acceptă ac-cép-tâ v.prez.
    c. dezactivare de-zac-ti-vâ-re de-zactivare dez-ac-ti-vâ-re struct.

On the other hand, DOOM stipulates two types of syllabification: one, preferred, according to the pronunciation and another according to the internal structure of the word. The second one, called ‘structural syllabification’, amounts to split the word at the boundaries of the affixes it contains, like in (6c), where the prefix ‘dez’ is separated from the main word ‘activare’. The first one is considered by default. In the case of structural syllabification the word ‘struct’ appears in the field OBS.

4.2.2. The XML format
An entry in XML format of RoSyllabiDict is described with the element form, see example (7). The value of the attribute w (= word) is the word the syllabification of which is given as the content of the element form. The value of the attribute obs (= observation) indicates the situation for which the syllabification is valid, if it is the case (7b, c).

Actually, values of the attribute obs can refer to the type of syllabification or to the grammatical information. On the one hand, its value is the word ‘struct’ if the structural
syllabification has been applied. On the other hand, its value indicates the grammatical information proper for that syllabified form, in cases of homonyms. Homonyms are differentiated only if they show different syllabifications (or different accents, see (6b) upper), such as the word aburi (‘steam’) in (7c), which, as a verb, is syllabified ‘a-bu-ri’ (obs="v.inf/v.perf”), and, as a noun, ‘a-buri’ (obs="s.").

(7) a. <form w="abandona" obs=""> a-ban-do-n</form>
b. <form w="ignorant" obs=""> ig-no-r</form>
<form w="ignorant" obs="struct"> i-gno-r</form>
c. <form w="aburi" obs="v.inf/v.perf"> a-bu-r</form>
<form w="aburi" obs="s."> &amp;#225;ur</form>

Note that only syllabified form contains accent information, Unicode encoded as a vowel with an accent diacritic mark, like it is presented in section 3.2.3.

5. Conclusion and further work
The work presented here is meant to fill a void in the field of electronic resources for Romanian language. The resources will be made available on web, for free, at an address communicated by the author. The dictionaries will be enriched with new entries, corresponding to the recent edition of DOOM, in 2005. We do not intend to introduce new words from corpora, because not all the words in corpora enter the language and we want to keep our dictionaries as close as possible to normative works. Words in corpora can be registered in special dictionaries.

6. Acknowledgements
The research reported in this paper has been supported by the National University Research Counsel of Romania (CNCSIS), grant no. 33549/18A/2002 and by the Institute of Romanian Language, contract no. 8/2005.

7. References
Bîrlădeanu, A., Burciu, N. (2006), Crearea unui generator morfologic pentru verbele din limba română. In C. Forescu, D. Tufis, & D. Cristea (eds.) Lucrărilor atelierului Resurse Lingvistice și Instrumente pentru Prelucrarea limbii Române, Editura Universității “Al. Ioan Cuza”, Iași, pp. 119-122.
DOOM (1989) – Dicționarul ortografic, ortoepic și morfologic al limbii române, Ed. Academiei, 1989.
Dinu, L. P. (2003). An approach to syllables via some extensions of Marcus-contextual grammars. Grammars 6(1), pp. 1-12
Dinu, L.P. (2006). On the quantitative and formal aspects of the Romanian syllables, Revue Roumaine de Linguistique, LI (3-4), pp. 477-498.
Dumitriu, D.-M., (2006a), Grammaires de flexion des noms roumains par automates finis, Ed. Aius, coll. Infolingua, 4, Craiova.
Dumitriu, D.-M., (2006b), Grammaires de flexion des adjectifs roumains par automates finis, Ed. Aius, coll. Infolingua, 4, Craiova.
Ionescu, E. (2003) Premiseale unui dicţionar morfologic electronic al limbii române. In F. Hristea & M. Popescu (eds.) Building Awareness in Language Technology, Editura Universității din București, pp. 461-468.
Paumier, S. (2006), Unitex 1.2 User Manual, http://igm.univ-mlv.fr/~unitex/UnitexManual.pdf.