Romanian Word-form Lexicon: tbl.wordform.ro

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc)

This is a wordform lexicon containing statistical information extracted from the Romanian Balanced Corpus

1.2 Representation of the lexicon (flat files, database, markup)

The lexicon is a flat file, one entry per line, fields being tab separated

1.3 Character encoding

The characters are UTF8 encoded

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan Tufis,
Address: Calea 13 Septembrie, no. 13, 050711
Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
Position: Director
Telephone: +4021 3188103
Fax: +40 21 3188142
e-mail: tufis@racai.ro

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as zip archive

2.3 Copyright statement and information on IPR

The resource is free license-based for research purposes and fee license-based for commercial purposes

3. TECHNICAL INFORMATION

3.1 Directories and files

There is only one file named **tblwordform.ro**

3.2 Data structure of an entry

Each entry is a four fields line, tab separated:
<wordform><tab>lemma<tab><msd><tab><frequency> where:
- <wordform> is the occurrence form in the ROMBAC corpus
- <lemma> is the lemma of the wordform or “=”, if the word form is the lemma form
- <msd> is a morpho-syntactic tag compliant with the Multext-East specification
- <frequency> is the of the wordform in the ROMBAC corpus;

Only word forms that at least 5 occurrences have been retained in the lexicon

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

There are 97372 entries and the lexicon requires 2,6 MB disk space

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the lexicon is Romanian

4. 2 Entry Type

Each of the 14 grammatical types defined by the Multext-East specifications is represented in the wordform lexicon.

As Multext-East specifications are supposed to cover many languages, some attributes might be irrelevant for a specific language. In the linear encoding of morpho-syntactic information for a wordform, the position of the attribute that is irrelevant is filled in with the special character '-'.

4.3 Attributes and their values

The MSD is a linear attribute value representation with fixed positions for each part of speech. Each position corresponds to a specific attribute and it is filled in by one character code. If the respective attribute is not relevant for the combination of the other attribute-values the position of the attribute that is irrelevant is filled in with the special character '-'.

For instance, a singular (s) masculine (m) common (c) noun (N) definite form (y) and in an oblique case –genitive or dative (o) will be encoded as **Ncmsoy**; the code **Vmip2s** describes a main (m) verb (V) indicative mode (i), present tense (p) second person (2) singular (s).

4.4 Coverage of the lexicon

The lexical entries cover general language as reflected in ROMBAC (Romanian Balanced Corpus)

4.5 Intended application of the lexicon

The lexicon is meant for all types of basic language processing (tokenization, tagging, lemmatization)

4.6 POS assignment

The MSD (extended POS) have been manually assigned by trained linguists

4.7 Reliability (automatically/manually constructed)

Highly reliable

5. RELEVANT REFERENCES AND OTHER INFORMATION

Dan Tufiş, Radu Ion - Specificaţii pentru clasa de etichete folosite în adnotarea morfo-lexicală a limbii române. Research Report, RACAI, June 2007

Radu Ion, Elena Irimia, Dan Stefănescu, Dan Tufiş. ROMBAC: The Romanian Balanced Annotated Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation - LREC 2012*, Istanbul, Turkey, 21-27 May, 2012

Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefănescu. RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May 2008. ELRA - European Language Resources Association.

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation - LREC 2004*, Lisabona, pp. 1535 – 1538.