The paradigmatic morphology of Romanian (RoPMorph)

1 BASIC INFORMATION

1.1 Resource composition

The paradigmatic morphology of Romanian has been developed for several years in different formats and variants (Tufiş, 1989), the most complete being implemented in the LISP-based ELU linguistic programming environment (Estival et al., 1994).

This unification-based implementation of the paradigmatic morphology, together with lexical repositories, associating paradigms and lexical roots to almost 35.000 of Romanian lemmas, was documented in a flat (theory neutral) attribute-value representation (FAVR, see Tufiş, Barbu, 1997).

In the context of the paradigmatic morphology theory, a word is treated as an entity made of two fundamental units: a *root* and an *ending* (built of one or more desinences and/or suffixes). The root usually carries context-free information, while the ending is a bearer of contextual information.

Some contextual information - consisting of restrictions on its use in conjunction with the specified endings - can be associated with the root if there is root alternation (for the same lemma and the same part-of-speech, the different inflected forms can share two or more roots).

The information associated with the root is stored in a dictionary (lexical repository) entry corresponding to the lemma of the corresponding root. Such an entry has the following structure:

pos
@lemma
root_1 root_2 ... root_k associated_paradigm1
root_k+1 ... associated_paradigm2
...

The information associated with the ending is stored in ROPMORPH, the file containing a complete inventory of the Romanian paradigms for verbs, nouns, pronouns, articles and adjectives.

Any lemma can be associated to one or more inflectional paradigms. An inflectional paradigm is a tree structure that identifies all the legal endings (and the associated restrictions) which can be associated to a root (or more roots) of a given lemma.

For a detailed description see 5.b and 5.c.

1.2 Representation of the resource

XML markup

1.3 Character encoding

The characters are UTF8 encoded

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

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

The resource will be uploaded on the MetaShare platform as an 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

A single xml file, named *morphaltUTF8.xml*

3.2 Data structure of an entry

```
<PARADIGM PARADIGM=value of type string >
  <TYPE TYPE= value of type string >
    <NUM NUM = value of type string >
      <ENCL ENCL= value of type string >
        <CASE CASE= value of type string >
          <TERM TERM= value of type string  ALT = value of type string />
        </CASE>
      </ENCL>
    </NUM>
  </TYPE>
  </PARADIGM
```

3.3 Resource size (nmb. of rules, MB occupied on disk)

286 entries: 286 paradigms for nouns, verbs, adjective, articles, pronouns.
Compressed:18k; uncompressed:603k.

4 CONTENT INFORMATION

4.1 Type of the resource (language (in(dependent)

Language dependent

4.2 The natural language for the resource

Romanian

4.3 Domain(s)/register(s) of the resource

Not applicable

4.4 Annotations in the resource

4.4.1 Types of annotations

Morphological Annotation

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed)

The representation is in the form of the attribute-value pairs.

4.4.3 *Attributes and their values (if annotated)*

- ✓ *Attributes for the PARADIGM tag*
  PARADIGM = the name of the paradigm
  CAT=”n/adv”) GEN="masculine/feminine”
  TYPE="manner/common"(“manner” for adverbs and “common” for nouns)
  INTENSIFY="none/diminutive/augmentative"
- ✓ *Attributes for the TYPE tag*
  TYPE= "{proper common}/common/proper"
- ✓ *Attributes for the NUM tag*
  NUM="singular/plural"
- ✓ *Attributes for the CASE tag*
  CASE="{nominative/genitive/
  dative/accusative/vocative}">
  Attributes for the HUM (human) tag
  HUM=”imperson/person”
  CASE="{nominative/genitive/
  dative/accusative/vocative}">

- ✓ *Attributes for the ENCL tag*
  ENCL="no/yes"
- ✓ *Attributes for the TERM tag*
  TERM=string; this string is an ending for a morphological form in Romanian
  ALT ="1/2" for nouns and adjectives and
  ALT="1/2/3/4/5/6/7/8/9" this shows to a morphological generator on which alternative root to apply the ending in TERM
- ✓ *Attributes for the VOICE tag*
  VOICE= "{active reflexive}/active/passive"
  GEN="masculine" – for the active VOICE
  NUM="singular" –for the active VOICE
- ✓ *Attributes for the TENSED tag*
  TENSED="yes/no" PRD="yes/no"
  For TENSED="yes"→ PRD="no" MOOD="infinitive" TENSE="present"
- ✓ *Attributes for the MOOD tag*
  MOOD="indicative/conjunctive/imperative/participle/gerund/supine/infinitive"
- ✓ *Attributes for the TENSE tag*
  TENSE="present/imperfect/simpleperfect/pastperfect"
- ✓ *Attributes for the PERS tag*
  PERS="1/2/3"

4.5 Intended application of the resource

This is a very reliable resource to be used in implementing a Romanian morphological generator. We also used this resource in developing a procedure of lexical acquisition for Part-of-Speech tagging. (see 5.d, 5.e)

4.6 Reliability of the annotations

Manually assigned.

5 RELEVANT REFERENCES AND OTHER INFORMATION

a. D.Tufiş. "It Would Be Much Easier If WENT Were GOED", in Proceedings of the 4th European Conference of the Association for Computational Linguistics, Manchester, 1989

b. Dominique Estival, Dan Tufiş, Octavian Popescu. 1994. Devélopment d'outils et des données linguistiques pour le traîtement du langage naturel. Rapport Final – Projet EST (7RUPJO38421) ISSCO, Geneve, September 1994

c. Dan Tufiş Barbu A.M. "A Reversible and Reusable Morpho-Lexical Description of Romanian". In Dan Tufiş, Andersen P. (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, 1997.

d. Elena Irimia. 2007. RomGen - A Romanian morphological generator, essential component of an English-to-Romanian MT system. In proceedings of the Doctoral Consortium, EUROLAN 2007 Summer School, Iaşi, România, July 23 - August 3, 2007 pp. 54-58.

e. Elena Irimia. 2007. ROG - A Paradigmatic Morphological Generator for Romanian. In proceedings of "The 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics", October 5-7, 2007, Poznań, Poland, pp. 408-412, ISBN 978-83-7177-407-2.

f. Dan Tufiş, Radu Ion, Elena Irimia şi Alexandru Ceauşu. 2007. Achiziţie lexicală nesupervizată pentru adnotare morfo-lexicală. Atelierul de Lucru "Resurse Lingvistice Româneşti şi Instrumente pentru Prelucrarea Limbii Române", 14-15 decembrie 2007, Iaşi, România.

g. Dan Tufiş, Elena Irimia, Radu Ion, Alexandru Ceauşu. 2008. Unsupervised Lexical Acquisition for Part of Speech Tagging. In Proceedings of LREC 2008 (Language Resources and Evaluation Conference), May 26 - June 1, Marrakech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0.