RO-JRC-ACQUIS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of the Romanian version of the Acquis Communautaire, the common set of laws of the European Union member states. There are 10704 documents in which 34234437 tokens occur. Out of these, 27968652 are words and the rest, punctuation.

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

The corpus is represented in XML Corpus Encoding Standard (XCES) format which is compliant with the XCES Schema revision 0.4 (2003)

1.3 Character encoding

The characters are UTF-8 encoded in the Latin 2 character set. A special mention is to the Romanian diacritics “ș” and “ț” with their upper case variants ”Ș” and ”Ț” which are not the (incorrect) ones from the Latin 2 character set (“ş” and “ţ” and “Ş” and “Ţ” respectively).

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan Tufiș,
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 RACAI’s 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

The archive that will be uploaded on the MetaShare platform will contain 47 different folders out of which 46 will contain XCES XML files of the respective laws grouped by year in the interval 1958-2006 except 1959-1961. One folder called 'XCES-Schema' contains the XCES schemas against which the validation of the XML files is ensured.

3.2 Data structure of an entry

An entry is a XCES encoded XML file.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 34234437 tokens including punctuation and 27968652 words. Out of the archive it needs about 2.8 GB for disk storage on a Windows 7 computer with the NTFS file system in place.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, POS tagged, lemmatized, chunked (shallow parsed) corpus and word sense disambiguated (for selected words – words from the domain)

4.2 The natural language(s) of the corpus

The language of the corpus is standard Romanian, orthography being compliant with the current Romanian Academy norms. The diacritical signs are in place (Tufiș and Ceaușu, 2008).

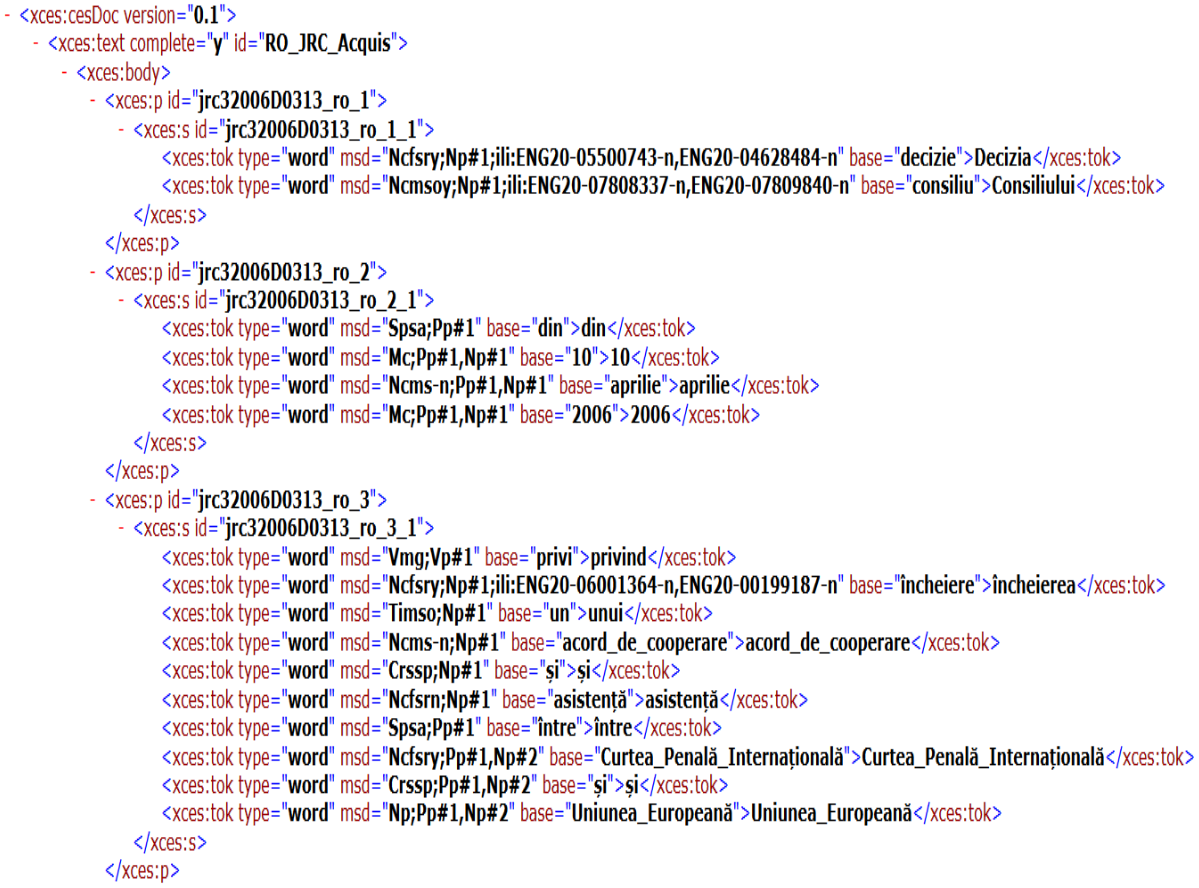
4. 3 Domain(s)/register(s) of the corpus

The text register represented into the corpus is the official language as used in legal documents.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is annotated at paragraph, sentence, constituent group and word levels, providing morpho-lexical and syntactic information. The following example shows the detailed structure with all tags and attributes used in the annotation. For more details about the XCES format, see www.xces.org.



```
- <xces:cesDoc version="0.1">
  - <xces:text complete="y" id="RO_JRC_Acquis">
    - <xces:body>
      - <xces:p id="jrc32006D0313_ro_1">
        - <xces:s id="jrc32006D0313_ro_1_1">
            <xces:tok type="word" msd="Ncfsry;Np#1;ili:ENG20-05500743-n,ENG20-04628484-n" base="decizie">Decizia</xces:tok>
            <xces:tok type="word" msd="Ncmsoy;Np#1;ili:ENG20-07808337-n,ENG20-07809840-n" base="consiliu">Consiliului</xces:tok>
          </xces:s>
        </xces:p>
      - <xces:p id="jrc32006D0313_ro_2">
        - <xces:s id="jrc32006D0313_ro_2_1">
            <xces:tok type="word" msd="Spsa;Pp#1" base="din">din</xces:tok>
            <xces:tok type="word" msd="Mc;Pp#1,Np#1" base="10">10</xces:tok>
            <xces:tok type="word" msd="Ncms-n;Pp#1,Np#1" base="aprilie">aprilie</xces:tok>
            <xces:tok type="word" msd="Mc;Pp#1,Np#1" base="2006">2006</xces:tok>
          </xces:s>
        </xces:p>
      - <xces:p id="jrc32006D0313_ro_3">
        - <xces:s id="jrc32006D0313_ro_3_1">
            <xces:tok type="word" msd="Vmg;Vp#1" base="privi">privind</xces:tok>
            <xces:tok type="word" msd="Ncfsry;Np#1;ili:ENG20-06001364-n,ENG20-00199187-n" base="încheiere">încheierea</xces:tok>
            <xces:tok type="word" msd="Timso;Np#1" base="un">unui</xces:tok>
            <xces:tok type="word" msd="Ncms-n;Np#1" base="acord_de_cooperare">acord_de_cooperare</xces:tok>
            <xces:tok type="word" msd="Crssp;Np#1" base="și">și</xces:tok>
            <xces:tok type="word" msd="Ncfsrn;Np#1" base="asistență">asistență</xces:tok>
            <xces:tok type="word" msd="Spsa;Pp#1" base="între">între</xces:tok>
            <xces:tok type="word" msd="Ncfsry;Pp#1,Np#2" base="Curtea_Penală_Internațională">Curtea_Penală_Internațională</xces:tok>
            <xces:tok type="word" msd="Crssp;Pp#1,Np#2" base="și">și</xces:tok>
            <xces:tok type="word" msd="Np;Pp#1,Np#2" base="Uniunea_Europeană">Uniunea_Europeană</xces:tok>
          </xces:s>
        </xces:p>
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with our high accuracy TTL tagger (Ion, 2007; Tufis et al., 2008) which implements the tiered tagging methodology (Tufiș, 1999; Tufiș & Dragomirescu, 2006). About 20% of the MSD have been manually checked, validated and, where the case, corrected (Tufiș and Irimia, 2006).

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.4.4 Attributes and their values (if annotated)

The *xces:p*, *xces:s* and *xces:tok* tags identify the level of the text under the tag: paragraph, sentence and token. *id* specifies the position of the textual unit in corpus:

- ‘jrc32006D0313_ro_1’ for the paragraph level
- ‘jrc32006D0313_ro_1_1’: the first part (jrc32006D0313) is the document identifier in the JRC Acquis corpus (the CELEX code). Then the language code follows (‚ro’), the id of the paragraph (the first integer) and the id of the sentence (the second integer);

Under each <xces:tok> tag can be found three attributes and a word form:

- *base,* whose value is the dictionary form of the word form;
- *msd:* which combine the MSD code associated to the word form, the chunk information, separated by semicolon; (ex: msd="Np;Np#1") and list of Princeton WordNet synset identifiers which are the most likely senses of that word; the WSD procedure is described in Ion (2010b).
- *type:* which values can be either “word” or “punctuation”; (ex: type="word">*mult*</xces:tok> or type="punctuation">.</xces:tok>)

The MSDs follows the Multext-East specifications (Erjavec, 2004). For Romanian there are 614 different MSDs (Tufis et al. 1997). They have been slightly modified (new tags for named entities have been added) are largely described in (Tufis and Ion, 2006).

4.5 Intended application of the corpus

Due to the mark-up accuracy, the corpus can be used for building robust statistical language models. It can also be used as a reference corpus for Romanian in various corpus specific types of investigation: quantitative analysis, collocation extraction, grammar induction, etc.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The annotations are highly reliable. The paragraph and sentence mark-up has been fully validated. The MSD tagging accuracy is at least 98%. The chunking annotation has been achieved based on a regular grammar defined over the MSD tags. The reliability of chunking mark-up is therefore similar to the tagging accuracy (cca. 98%). The WSD annotation is around 80% accurate given the fact that the most 2 labels have been assigned (to selected words).

5 RELEVANT REFERENCES AND OTHER INFORMATION

Alexandru Ceauşu. 2008. Colectarea şi procesarea documentelor româneşti ale corpusului JRC-Acquis. In Diana Maria Trandabăţ, Dan Cristea, Dan Tufiş (eds.), *Lucrările atelierului Resurse*

Lingvistice şi Instrumente pentru Prelucrarea Limbii Române, Editura Universităţii „Al. I Cuza", Iaşi.

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th LREC Conference*, LREC'04, Lisabona, pp. 1535 - 1538,

Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, Elena Irimia and Verginica Barbu Mititelu. 2010. *A Trainable Multi-factored QA System*. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, Giovanna Roda (eds.) *Multilingual Information Access Evaluation*, Vol. I Text Retrieval Experiments, pp. 257—264, Lecture Notes in Computer Science, Volume 6241/2010, Springer-Verlag.

Radu Ion, and Dan Ştefănescu. 2010b. *RACAI: Unsupervised WSD Experiments @ SemEval-2, Task 17*. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval-2, pp. 411—416, Uppsala, Sweden, July 2010. (C) Association for Computational Linguistics. ISBN: 978-1-932432-70-1.

Radu Ion. 2007. Word Sense Disambiguation methods applied to English and Romanian. Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, 153 pages.

Dan Tufiş and Alexandru Ceauşu. 2008. DIAC+: A Professional Diacritics Recovering System. In *Proceedings of the 6th LREC Conference*, Marrakech.

Tufiş, D. 1999."Tiered Tagging and Combined Classifiers". In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33

Dan Tufiş, Liviu Dragomirescu. 2004. Tiered Tagging Revisited. In *Proceedings of the 4th LREC'04 Conference*, Lisabona, pp. 39-42

Dan Tufiş, Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. 1997."Corpora and Corpus-Based Morpho-Lexical Processing". In Dan Tufiş, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, pp. 35-56.

Dan Tufiş, Radu Ion. 2007. Specificații pentru clasa de etichete folosite în adnotarea morfo-lexicală a limbii române. Raport de cercetare, iunie, Institutul de Cercetări pentru inteligență artificială, 24 pages.

Dan Tufiş, Elena Irimia. 2006. RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC Conference*, Genoa, pp. 869-872

Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefănescu. 2008. RACAI's Linguistic Web Services. In *Proceedings of the 6th LREC Conference* – LREC'08, Marrakech.