English-Romanian JRC-ACQUIS Corpus

1 BASIC INFORMATION

1.1 Corpus composition

The corpus represents the Romanian-English part of the Acquis Communautaire (JRC-Acquis), the common set of laws of the European Union member states. There are 8702 XML documents (30,598,470 tokens, including punctuation), together with associated files describing the sentence alignment of each document.

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. 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 46 different folders comprising:

- XCES XML files of the European laws grouped by year in the interval 1958-2006 except 1959-1961.
- Align files corresponding to each xml file, containing the sentence alignment for the respective document;

3.2 Data structure of an entry

An entry is a XCES encoded XML file plus its associated align 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 bilingual, POS tagged, lemmatized, chunked (shallow parsed) and word sense disambiguated (for selected words – words from the domain)

4.2 The natural language(s) of the corpus

English and Romanian; the language of the Romanian part 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)

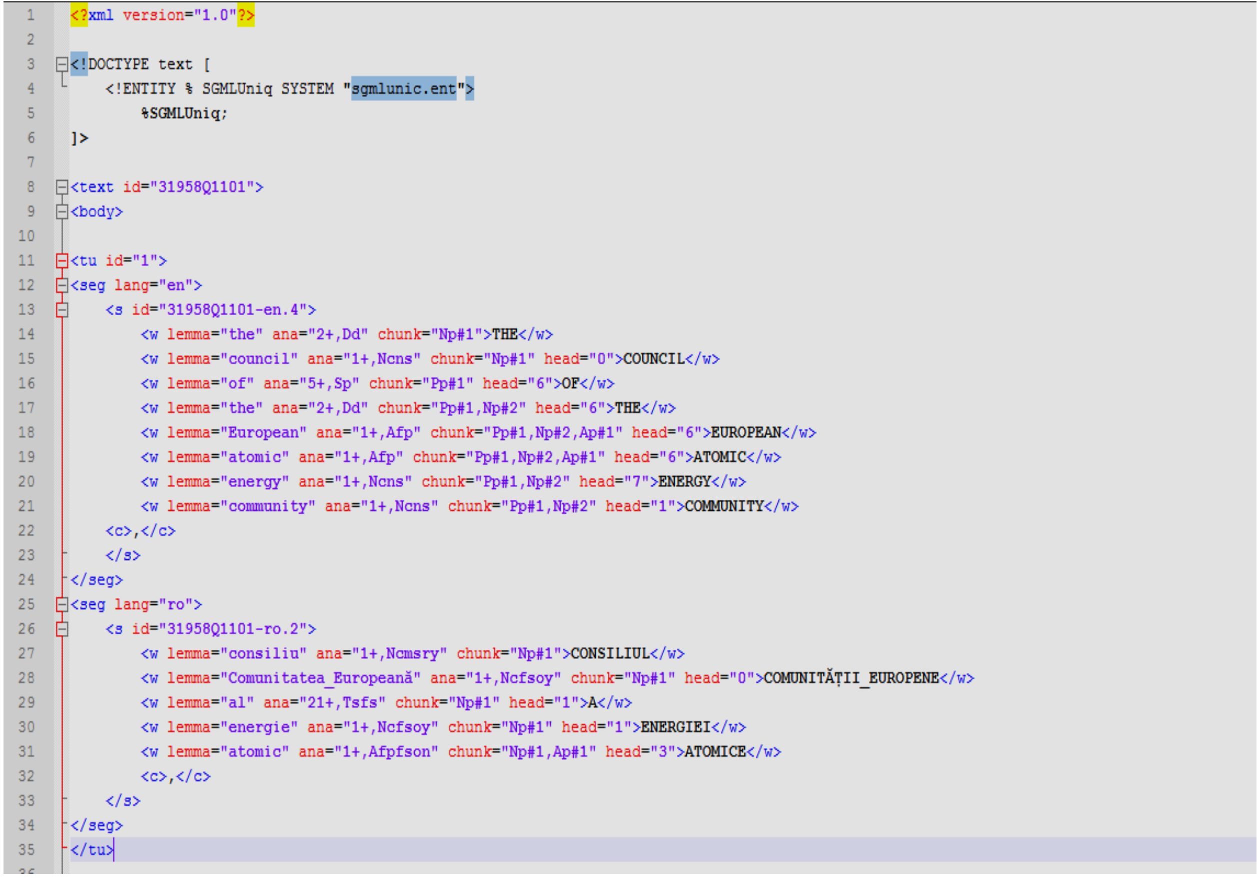
Each document contains more translation units (annotated with <tu>) and each translation unit contains 2 segments, a Romanian and an English one. The fact that these two segments are in the same <tu>

signifies that they are reciprocal translations or that they are aligned at the sentence level.

The corpus is annotated at constituent group (chunk) 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.

The alignment file contains entries of type:

Sentence_number word_position_en word_position_ro S (S doesn't have an explicit significance, it is a residue from the sentence aligner software).



```
<?xml version="1.0"?>

<!DOCTYPE text [
    <!ENTITY % SGMLUniq SYSTEM "sgmlunic.ent">
        %SGMLUniq;
]>

<text id="31958Q1101">
<body>

<tu id="1">
<seg lang="en">
    <s id="31958Q1101-en.4">
        <w lemma="the" ana="2+,Dd" chunk="Np#1">THE</w>
        <w lemma="council" ana="1+,Ncns" chunk="Np#1" head="0">COUNCIL</w>
        <w lemma="of" ana="5+,Sp" chunk="Pp#1" head="6">OF</w>
        <w lemma="the" ana="2+,Dd" chunk="Pp#1,Np#2" head="6">THE</w>
        <w lemma="European" ana="1+,Afp" chunk="Pp#1,Np#2,Ap#1" head="6">EUROPEAN</w>
        <w lemma="atomic" ana="1+,Afp" chunk="Pp#1,Np#2,Ap#1" head="6">ATOMIC</w>
        <w lemma="energy" ana="1+,Ncns" chunk="Pp#1,Np#2" head="7">ENERGY</w>
        <w lemma="community" ana="1+,Ncns" chunk="Pp#1,Np#2" head="1">COMMUNITY</w>
    <c>,</c>
    </s>
</seg>
<seg lang="ro">
    <s id="31958Q1101-ro.2">
        <w lemma="consiliu" ana="1+,Ncmsry" chunk="Np#1">CONSILIUL</w>
        <w lemma="Comunitatea_Europeană" ana="1+,Ncfsoy" chunk="Np#1" head="0">COMUNITĂȚII_EUROPENE</w>
        <w lemma="al" ana="21+,Tsfs" chunk="Np#1" head="1">A</w>
        <w lemma="energie" ana="1+,Ncfsoy" chunk="Np#1" head="1">ENERGIEI</w>
        <w lemma="atomic" ana="1+,Afpfson" chunk="Np#1,Ap#1" head="3">ATOMICE</w>
        <c>,</c>
    </s>
</seg>
</tu>
```

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 automatically 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)

The corpus is aligned at sentence and word level.

4.4.4 Attributes and their values (if annotated)

The DTD annotation schema:

```
<!DOCTYPE text [
<!ELEMENT text (body)>
<!ATTLIST text id CDATA #REQUIRED>
  <!ELEMENT body (tu+)>
  <!ELEMENT tu (seg+)>
  <!ATTLIST tu id CDATA #REQUIRED>
  <!ELEMENT seg (s)>
  <!ATTLIST seg lang (en | ro) #REQUIRED>
  <!ELEMENT s (w | c)+>
  <!ATTLIST s id ID #REQUIRED>
  <!ELEMENT c (#PCDATA)>
  <!ELEMENT w (#PCDATA)>
  <!ATTLIST w
ana CDATA #REQUIRED
lemma CDATA #REQUIRED
        chunk CDATA #IMPLIED
        wns CDATA #IMPLIED
        head CDATA #IMPLIED
  >
<!ENTITY % ISOlat1 PUBLIC
 "ISO 8879-1986//ENTITIES Added Latin 1//EN" >
%ISOlat1;
      <!ENTITY % ISOlat2 PUBLIC
 "ISO 8879-1986//ENTITIES Added Latin 2//EN" >
%ISOlat2;
<!ENTITY % ISOnum PUBLIC
 "ISO 8879-1986//ENTITIES Numeric and Special Graphic//EN" >
%ISOnum;
<!ENTITY % ISOpub  PUBLIC
 "ISO 8879-1986//ENTITIES Publishing//EN" >
%ISOpub;
]>
```

The *ana* attribute combines the MSD code associated to the word form and the metacategory of the MSD (ex: ana="1+,Ncns"). The values of the *wns* attributes are part of a list of Princeton WordNet synset identifiers which are the most likely senses of that word; the WSD procedure is described in Ion (2010b). The head attribute specifies the head of the chunk in which the annotated word takes part.

The MSDs follow 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) and 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 sentence alignment 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.