A PREPROCESSED AND SENTENCE ALIGNED EN-FR-RO STRONGLY COMPARABLE NEWS CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

4623 documents in all languages: 1848 English files, 966 Romanian files and 1809 French files. There are also 1586 English-French sentence aligned files, 859 English-Romanian and 844 French-Romanian sentence aligned files.

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

The corpus is encoded in the XCES format (http://www.xces.org/).

1.3 Character encoding

The documents are UTF-8 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 an archive and can also be provided upon request made through the MetaShare platform.

2.3 Copyright statement and information on IPR

The resource is free, license-based, for research purposes (see the metadata on the MetaShare platform).

3 TECHNICAL INFORMATION

3.1 Directories and files

The corpora contains 5 sets of data grouped in separate folders ("ec.europa.eu", "euronews", "europarl1", "europarl2", "europarl3"). Each folder has 3 subfolders named "en-xces", "ro-xces" and "fr-xces" for English, Romanian and French documents (in XCES XML format). The XML Schema definitions can be found in the folder "XCES-Schema" located in the root folder. There are also 3 directories containing sentence aligned descriptor files: “sent-align-en-fr” with XCES compliant sentence alignments between English and French files and “sent-align-en-ro”, “sent-align-fr-ro” with similar information.

3.2 Data structure of an entry

The corpus documents are XCES files, UTF-8 encoded. They are grouped together by their language. The “en-xces” folder contains documents in English, “fr-xces” contains the French documents and “ro-xces” contains the Romanian documents. The filenames for comparable entries start with the same unique identifier (either a numeric value or a randomly generated GUID) and end with the character ‘_’ and their language code (e.g. 1_EN.xml). Example:

euronews\en-xces\7_EN.xml euronews\ro-xces\7_RO.xml euronews\fr-xces\7_FR.xml

The unique identifier is relative to each set (europarl1, europarl2, euronews etc.) meaning that “euronews\en-xces\1_EN.xml” is not the same document as “europarl1\en-xces\1_EN.xml”.

Each corresponding English-French, English-Romanian and French-Romanian document pair has been automatically sentence aligned using Moore’s sentence aligner (Moore, 2002) with a threshold of 0.8. Manually validating a subset of the alignments for each language pair, revealed a very good quality of the resulting sentence alignments.

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

- ec.europa.eu: 137 documents for each language (total 411 documents)
- euronews: 506 documents for English, 491 for French and 198 for Romanian (total 1195 documents)
- europarl1 (set 3 of files): 492 documents for English, 478 for French and 203 for Romanian (total 1173 documents)
- europarl2: 501 documents for English, 491 for French and 216 for Romanian (total 1208 documents)
- europarl3: 212 documents for each language (total 636 documents)
- sent-align-en-fr: 1586 files corresponding to 1586 aligned document pairs. In all document pairs, there are 67899 aligned sentence pairs generated by the sentence aligner
- sent-align-en-ro: 859 files with 27045 aligned sentence pairs
- sent-align-fr-ro: 844 files with 23767 aligned sentence pairs

The number of tokens (words and punctuation) in the English-French corpus is 766996 in English and 870146 in French, for the English-Romanian corpus is 284444 for English and 276377 in Romanian and for French-Romanian corpus is 225647 in French and 197386 in Romanian.

The size on disk is 306 MB.

4 CONTENT INFORMATION

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

This is a multilingual strongly comparable corpus with an automatically extracted parallel sub-corpus. It is sentence split, POS tagged, lemmatized and chunked.

4.2 The natural language(s) of the corpus

The languages for the corpus are: Romanian, English, and French.

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

The text registers represented into the corpus are: journalistic language as used in the daily newspapers and 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)

See XCES specifications for details.

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

The corpus uses the Multext-East MSD tags (http://nl.ijs.si/ME/V4/msd/html/index.html).

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

The corpus is aligned at document level and sentence level.

4.4.4 Attributes and their values (if annotated)

See XCES specifications for details.

4.5 Intended application of the corpus

Multilingual applications (MT, CLIR)

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

The annotations are automatically generated.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Moore, Robert C. 2002. *Fast and Accurate Sentence Alignment of Bilingual Corpora*. In Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, 2002, pp. 135-244.