MULTILINGUAL NEWS CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

4622 documents.

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

The corpus is represented in XCES format.

1.3 Character encoding

The documents 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 an archive.

2.3 Copyright statement and information on IPR

The resource is free, license-based, for research purposes

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 format). The Xml Schema Definitions can be found in the folder “XCES-Schema” located in the Root folder. The file “mling-news-general-metadata.xml” contains some general information about the corpora (license, author, etc.) and “mling-news-text-metadata.xml” contains annotation metadata (languages, number of tokens, annotation mode etc.).

3.2 Data structure of an entry

The documents are plain text UTF8 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). Examples:

euronews\en-xces\1_EN.xml euronews\ro-xces\1_RO.xml euronews\fr-xces\1_FR.xml
europarl1\en-xces\1_EN.xml europarl1\ro-xces\1_RO.xml europarl1\fr-xces\1_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".

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

- ec.europa.eu (set 1 of files): 137 documents for each language (total 411 documents)
- Euronews (set 2 of files): 506 documents for each language (total 1518 documents)
- europarl1 (set 3 of files): 492 documents for each language (total 1476 documents)
- europarl2 (set 4 of files): 500 documents for each language (total 1500 documents)
- europarl3 (set 5 of files): 212 documents for each language (total 636 documents)

The number of tokens (words) is 1334942 for English, 659031 for Romanian and 1480103 for French.

The size on disk is 277 MB.

4 CONTENT INFORMATION

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

This is a multilingual comparable corpus.

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 documentation for details.

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

The corpus is POS tagged.

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

The corpus is aligned at document level.make

4.4.4 Attributes and their values (if annotated)

See XCES documentation 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