Cross-Language Unit Elicitation alignments (CLUE)

1. BASIC INFORMATION

1. Corpus composition

This corpus consists of a set of manual alignments of 400 parallel sentences from the Europarl corpora [1] in four languages (pt, en, es, fr), being considered the following pairs: en-es, en-fr, en-pt, es-fr, pt-es. This work deeply extends the corpus detailed in [2].

2. Representation of the corpora (flat files, database, markup)

The corpus is composed of several txt files, namely:

a) four files containing the 400 parallel sentences in each language;
b) a file for each pair en-es, en-fr, en-pt, es-fr, pt-es and pt-fr containing the word alignments;
c) a file for each pair en-es, en-fr, en-pt, es-fr, pt-es and pt-fr containing the multiword units alignments.

3. Character encoding

Characters are encoded in ISO-8859-1 (Latin1).

2. ADMINISTRATIVE INFORMATION

1. Contact person

Name:Luísa Coheur
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Assistant Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: luisa.coheur@inesc-id.pt

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.

3. TECHNICAL INFORMATION

1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain one folder with 4 files with extensions pt, es, fr and es (one for each considered language), 6 files with extension wa (word alignments) and 6 files with extension mwu (multiword units' alignments).

2. Data structure of an entry

Each file containing the sentences in the different languages (extensions pt, en, fr and es) has a sentence per line.

Each file with the extension wa contains, in each line, three digits and a S or a P (ex: 1 14 17 P). The first digit represents the sentences/lines in the parallel files that are being aligned; the second an third digits correspond to the position of the words being aligned; S represents a sure alignment and P a possible one. Details about this can be found in [2]. Very detailed guidelines are included in the corpus.

Each file with the extension mwu contains, in each line, five digits and an S or a P. The first digit represents the sentences/lines in the parallel files that are being considered; the second and third digit represent the positions of the word units being aligned in the source language (the fourth and the fifth digit represent the same for the target language); S represents a sure alignment and P a possible one.

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

Each parallel file contains 400 sentences. The .wa files (word alignments) totalize 48130 alignments and the .wmu files (multi-word units) totalize 22,099 alignments. The whole set of files occupies 1.3 MB.

4. CONTENT INFORMATION

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

This corpus is multilingual, parallel and lightly annotated.

2. The natural language(s) of the corpus

The languages of the corpus are Portuguese, English, French and Spanish.

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

The corpus has sentences from the European parliament sessions [1].

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)

Only the files with extensions pt, en, fr and es are annotated (in xml).

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

Each sentence of the previously mentioned files is marked with the tag <s snum=N>, where N represents the sentence number.

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

As previously stated, in order to establish the alignments between the different languages, for each considered languages' pair, two files are provided. The one with extension wa represents the word alignments; the one with extension mwu represents the multiword units alignments. The used notation is explained in Section 3.2.

4.4.4 Attributes and their values (if annotated)

Not relevant.

5. Intended application of the corpus

This corpus can be used as a gold collection for word alignments.

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

The alignments were manually built.

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5.

[2] João Graça and Joana Paulo Pardal and Luísa Coheur and Diamantino António Caseiro, **Building a golden collection of parallel Multi-Language Word Alignment**, *The 6th International Conference on Language Resources and Evaluation, LREC 2008*, May. 2008