Portuguese Definitions Corpus

I. Basic Information

1.1. Corpus information

The corpus presented here is a collection of several tutorials and scientific papers in the field of Information Technology with 651 annotated definitions from Portuguese. The texts were collected from the Web at the beginning of the 2006 and they are organized in 32 files of three different sub-domains with 268,064 tokens: Information Society (91,825), Information Technology (80,483), and e-Learning (94,756).

In this corpus, a definition is assumed to be a sentence containing an expression (the *definiendum*) and its definition (the *definiens*) and a connector between them. We identify three different tipology definitions corresponding to three different connectors, that is the verb “to be” (“ser”), all other verbs other than “to be” and punctuation mark such as “:”, finally a last class, covering all definitions not covered by the previous classification (see Del Gaudio, 2007c and 2009a). The following table displays the distribution of the different types of definitions in the corpus.

Type	*IS*	*IT*	*e-Learning*	***Total***
is_def	68	40	17	125
verb_def	80	77	66	223
punct_def	9	89	35	133
other_def	31	52	39	122
Total	**188**	**258**	**157**	**603**

This corpus was collected in the context of Language Technologies for eLearning project (www.lt4el.eu) founded by European Union whose main goal is to improve e-Learning systems by using multilingual language technology tools and semantic web techniques.

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

The corpus is represented in a variant of the XCES format described by DTD file (see LT4ELAnaProjectv3.4.dtd).

1.*3. Character encoding*

The characters are in UTF8 code.

II. Administrative Information

2.1. Contact person

Name: António Branco
Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087

Fax: +351 217 500 084
E-mail: antonio.branco@di.fc.ul.pt

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

This resource is available through META-SHARE.

2.3. Copyright statement and information on IPR

This resource is licensed for research purposes only, with no redistribution, nor derivatives allowed.

III. Technical Information

3.*1. Directories and files*

The archive that can be uploaded on the Meta-Share is a .zip file with 33 files: 32 XML and 1 DTD.

3.2. Data structure of an entry

For each text file with a set of sentences, the data is divided into paragraphs, which has the respective sentences segmented by tokens.

3.3. *Corpus size (nmb. of tokens, NB occupied in disk)*

The corpus is composed by 268,064 tokens with 2.7 *MB* compressed (25.6 *MB* uncompressed) for disk storage.

IV. Content Information

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

This is a monolingual and annotated corpus.

4.*2. The natural language(s) of the corpus*

The language of the corpus is Portuguese with pre-spelling reform of 1990[1].

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

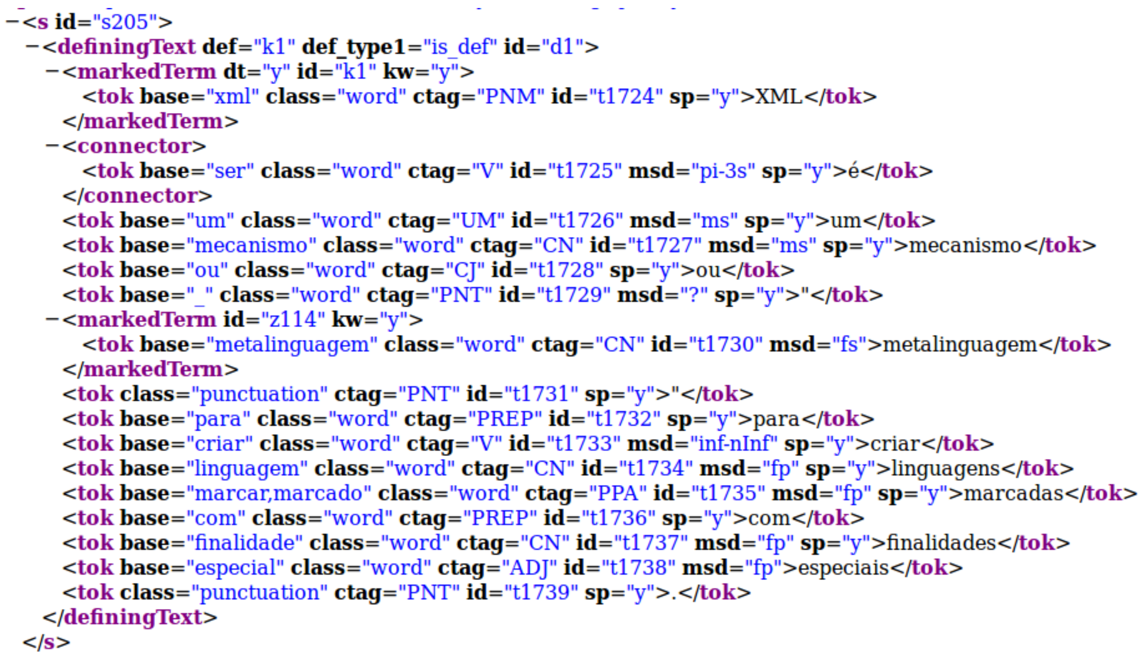
Concerning the Information domain, there are three sub-domains in this corpus: Information Society, Information Technology for non-experts, and e-Learning.

4.4. Annotation in the corpus (if an annotated corpus)

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

The corpus was pre-processed in order to convert it into a common XML format, conforming to a DTD derived from the XCES DTD for linguistically annotated corpora (see Ide and Suderman, 2002). The following example shows more detailed the structure of a DTD format, containing the lemma of each word, the attribute *ctag*, the POS information, and the *msd* with the morpho-syntactic inflection.

[1] This means that the orthography rules used are those that are described by the Orthography Reform of 1945. The orthographic agreement of 1990 was adopted just in may of 2009 and is being implemented until 2012.



```
-<s id="s205">
  -<definingText def="k1" def_type1="is_def" id="d1">
    -<markedTerm dt="y" id="k1" kw="y">
        <tok base="xml" class="word" ctag="PNM" id="t1724" sp="y">XML</tok>
      </markedTerm>
    -<connector>
        <tok base="ser" class="word" ctag="V" id="t1725" msd="pi-3s" sp="y">é</tok>
      </connector>
      <tok base="um" class="word" ctag="UM" id="t1726" msd="ms" sp="y">um</tok>
      <tok base="mecanismo" class="word" ctag="CN" id="t1727" msd="ms" sp="y">mecanismo</tok>
      <tok base="ou" class="word" ctag="CJ" id="t1728" sp="y">ou</tok>
      <tok base="_" class="word" ctag="PNT" id="t1729" msd="?" sp="y">"</tok>
    -<markedTerm id="z114" kw="y">
        <tok base="metalinguagem" class="word" ctag="CN" id="t1730" msd="fs">metalinguagem</tok>
      </markedTerm>
      <tok class="punctuation" ctag="PNT" id="t1731" sp="y">"</tok>
      <tok base="para" class="word" ctag="PREP" id="t1732" sp="y">para</tok>
      <tok base="criar" class="word" ctag="V" id="t1733" msd="inf-nInf" sp="y">criar</tok>
      <tok base="linguagem" class="word" ctag="CN" id="t1734" msd="fp" sp="y">linguagens</tok>
      <tok base="marcar,marcado" class="word" ctag="PPA" id="t1735" msd="fp" sp="y">marcadas</tok>
      <tok base="com" class="word" ctag="PREP" id="t1736" sp="y">com</tok>
      <tok base="finalidade" class="word" ctag="CN" id="t1737" msd="fp" sp="y">finalidades</tok>
      <tok base="especial" class="word" ctag="ADJ" id="t1738" msd="fp">especiais</tok>
      <tok class="punctuation" ctag="PNT" id="t1739" sp="y">.</tok>
    </definingText>
  </s>
```

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

The corpus was automatically annotated with morpho-syntactic information using the LX-Suite[2] (see Silva, 2007). This is a set of tools for the shallow processing of Portuguese with state of the art performance. This pipeline of modules comprises several tools, namely a sentence chunker (99.94% F-score), a tokenizer (99.72%), a POS tagger (98.52%), and nominal and verbal featurizers (99.18%), and lemmatizers (98.73%).

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

It does not apply.

4.4.4. Attributes and their values (if annotated)

In each sentence with a definition structure, the term defined, the definition, and the connector were manually annotated using a different XML tag, *markedTerm*, *connector* (only for copula definitions*)*, and *markedTerm*, respectively, with the assignment of the information about the type of definition. The definition typology is made of four different classes whose members were tagged with *is_def*, for copula definitions, *verb_def*, for verbal non copula definitions, *punct_def*, for definitions whose connector is a punctuation mark, and finally *other_def*, for all the remaining definitions.

4.5. Intended application of the corpus

Concerning that the main goal of this corpus was to test a tool for supporting glossary construction in a automatic way in e-Learning management systems for Portuguese (see Del Gaudio, 2009b and 2009c), it also compose a reference corpus for various comparative analysis in specialized language for Portuguese and between languages. This definitions corpus is also important in the context of Question Answering (QA), ontology learning, dictionary and glossary construction, among others.

[2] Available at http://lxcenter/services/en/LXServicesSuite.html.

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

Firstly, the corpus was automatically annotated with LX-Suite tools with high accuracy (see 4.2.). In a second phase, human experts annotators marked the definitions structures.

V. Relevant References and Other Information

Silva, João, 2007. *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. MSc thesis, University of Lisbon. Published as Technical Report DI-FCUL-TR-07-16.

Ide, Nancy and Keith Suderman, XML, Corpus Encoding Standard, Document XCES 0.2. Technical Report, Department of Computer Science, Vassar College and Equipe Langue ed Dialogue, New York, USA and LORIA/CNRS, Vandouvre-les-Nancy, France, 2002.

Del Gaudio, Rosa and António Branco, 2007b, “Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach”. In J. Neves, M. Santos and J. Machado (eds.), *EPIA2007 - 13th Portuguese Conference on Artificial Intelligence*, LNAI 4874, Berlin, Springer, pages 659-670.

Del Gaudio, Rosa and António Branco, 2007c, “Supporting e-Learning with Automatic Glossary Extraction: Experiments with Portuguese”. In *Proceedings of the Workshop on Natural Language Processing and Knowledge Representation for e-Learning Environments*, RANLP2007 - International Conference on Recent Advances in Natural Language Processing.

Del Gaudio, Rosa and António Branco, 2009a, “Evaluating a Learning Management System improved with Language Technology”. In *Proceeding of the 12th International Conference Interactive Computer Aided Learning (ICL)*. Villach, Austria.

Del Gaudio, Rosa and António Branco, 2009b, “Extraction of Definitions in Portuguese: An Imbalanced Data Set Problem”. In *Proceedings of Text Mining and Applications (TEMA 2009)*, pages 501-512.

Del Gaudio, Rosa and António Branco, 2009c, “Improving e-Learning Experience with Language Technology: Evaluation Results”. In *Proceeding of the International Conference Interactive Computer Aided Blended Learning (ICBL)*. Florianopolis, Brazil.