IULA Spanish LSP Treebank

1 BASIC INFORMATION

1.1 Corpus composition

This resource consists of 40,000 sentences taken from the IULA technical corpus. The sentences have been selected to be representative of the original corpus distribution both in number of sentences per domain and its length. Such sentences have been automatically annotated with POS information, and semi-automatically annotated with syntactic information. Finally, they were automatically converted to dependencies in the CONLL format.

The IULA technical corpus consists of a number of specialized texts in following domains: Law, Economics, Computer Science domain, Environment and Medicine. This LSP corpus has been created with articles from specialized publications, PhD theses, etc.

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

The corpus is represented in vertical CoNLL format and as a single text file.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Jorge Vivaldi,
Address: Roc Boronat, no. 138, 08018, Barcelona, Spain
Affiliation: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra
Telephone: +34 935422332
Fax: +34 935422321
e-mail: jorge.vivaldi@upf.edu

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

The resource will be available on the MetaShare platform as a single archive.

2.3 Copyright statement and information on IPR

The resource is free, license-based with restrictions, for research purposes and fee license-based for commercial purposes. CC_BY

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive that will be available on the MetaShare platform will contain a single text file.

3.2 Data structure of an entry

Each entry consists of a sentence separated from by a single empty line. Each sentence is segmented into tokens (equivalent to word or named entities), each token occupy a single line that contains all relevant information.

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

The corpus contains 40,000 sentences that contain about 550,000 tokens. It needs about 37 MB for disk storage (4.2 MB compressed).

4 CONTENT INFORMATION

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

This corpus is a balanced monolingual, heavily annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is Spanish.

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

The sentences belong to Language for Specific Purposes texts from the following domains: Law, Economics, Computer Science domain, Environment and Medicine. Such sentences have been taken from specialized publications like PhD theses, books, scientific articles, etc.

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)

The sentences are annotated at word level with morpho-lexical and syntactic information, dependency relations following the standard CONLL format that is described in table 1. Figure 1 shows an actual example of a sentence tagged using CONLL format. Sentences are separated by empty lines.

Field number:	Field name:	Description:
1	ID	Token counter, starting at 1 for each new sentence.
2	FORM	Word form or punctuation symbol.
3	LEMMA	Lemma.
4	CPOSTAG	Coarse-grained part-of-speech tag
5	POSTAG	Fine-grained part-of-speech tag.
6	FEATS	Unordered set of syntactic and/or morphological features, separated by a vertical bar (|), or an underscore if not available.
7	HEAD	Head of the current token, which is either a value of ID or zero ('0').
8	DEPREL	Dependency relation to the HEAD.
9	PHEAD	Projective head of current token, which is either a value of ID or zero ('0'), or an underscore if not available.
10	PDEPREL	Dependency relation to the PHEAD, or an underscore if not available.

Table 1. CONLL format description

1	Gráficas	Gráfica	n	NCFP000	postype=Comun| gen=Femenino| num=Plural	0	ROOT	0	_
2	de	de	s	SPS00	postype=Preposicion| form=Simple	1	MOD	1	_
3	Hill	NP00000	n	NP00000	postype=Propio	2	COMP	2	_
4	de	de	s	SPS00	postype=Preposicion| form=Simple	1	MOD	1	_
5	la	el	d	DA0FS0	postype=Articulo| gen=Femenino| num=Singular	6	SPEC	6	_
6	mioglobina	mioglobina	n	NCFS000	postype=Comun| gen=Femenino| num=Singular	4	COMP	4	_
7	y	y	c	CC	postype=Coordinada	4	COORD	4	_
8	de	de	s	SPS00	postype=Preposicion| form=Simple	7	CONJ	7	_
9	la	el	d	DA0FS0	postype=Articulo| gen=Femenino| num=Singular	10	SPEC	10	_
10	hemoglobina	hemoglobina	n	NCFS000	postype=Comun| gen=Femenino| num=Singular	8	COMP	8	_
11	HbA	NP00000	n	NP00000	postype=Propio	10	MOD	10	_
12	.	.	f	Fp	punct=period	11	punct	11	_

Figure 1. Example of a sentence tagged according the CONLL format

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 by the Freeling tagger. The MSD annotations are based in the PAROLE standard. See detailed information at:
http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

While the complete set of syntactic categories and their associated meaning are listed below:

- ROOT: Root
- DO: Direct Object
- IO: Indirect Object
- OBLC: Oblique Object
- BYAG: By agent complement
- ATR: Attribute
- PRD: Predicative complement
- OPRD: Object predicative complement
- PP-LOC: Locative prepositional complement
- PP-DIR: Directional prepositional complement
- SUBJ-GAP: Subject in a gapping construction
- COMP-GAP: Complement in a gapping construction
- MOD-GAP: Modifier in a gapping construction
- VOC: Vocative
- IMPM: Impersonal marker
- PASSM: Passive marker
- PRNM: Pronominal marker
- COMP: Complement (of N, ADJ, ADV, PREP)
- MOD: Modifier
- NEG: Negation
- SPEC: Specifier
- COORD: Coordination
- CONJ: Conjunction
- PUNCT: Punctuation

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

Not relevant

4.4.4 Attributes and their values (if annotated)

Not relevant

4.5 Intended application of the corpus

Due to the mark-up accuracy, this is a resource useful for linguistic research (quantitative analysis, collocation extraction, etc.) as well as training statistical based natural language processing tools (grammar induction, etc.). IULA have used this resource to train a MALT parser (more information at w3.msi.vxu.se/~nivre/research/MaltParser.html) for Spanish that can be queried at: http://ws04.iula.upf.edu/soaplab2-axis/#syntactic_tagging.malt_parser_row.

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

The annotations are highly reliable. The sentence segmentation mark-up has been manually validated. The MSD tagging accuracy is at least 98%. The syntactic annotations has been manually validated.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Marimon, Montserrat; Fisas, Beatriz; Bel, Núria; Arias, Blanca; Vázquez, Silvia; Vivaldi, Jorge; Torner, Sergi; Villegas, Marta; Lorente, Mercè (2012). "The IULA Treebank" in Calzolari, Nicoletta (et al.) (ed.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA). Pages 1920-1926. ISBN 978-2-9517408-7-7.

Oepen S, Flickinger D, Toutanova K, Manning CD (2002) LinGo Redwoods. "A Rich and Dynamic Treebank for HPSG". In: Proceedings of TLT 2002, Sozopol, Bulgaria.

Copestake A (2002). "Implementing Typed Feature Structure Grammars". CSLI Publications, Stanford.

Marimon, M. (in press). "The Spanish DELPH-IN Grammar, Language Resources and Evaluation. DOI: 10.1007/s10579-012-9199-7.

Vivaldi, J. (2009). "Corpus and exploitation tool: IULACT and bwanaNet" in Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) A survey on corpus-based research = Panorama de investigaciones basadas en corpus. Proceedings of the I Congreso Internacional de Lingüística de Corpus (CICL-09), May 7-9 2009, Universidad de Murcia]. Murcia: Asociación Española de Lingüística del Corpus. Pages 224-239. ISBN 978-84-692-2198-3

Padró L. (2011). Analizadores Multilingües en FreeLing. Linguamatica. Vol. 3, No. 2, Pages 13-20