



**GLiCom Spanish Wordform List v.1 is a computational lexicon of inflected wordforms in Spanish.**

**This lexicon can be used in any application for Text Analysis in Spanish, in particular those in need for a lemmatizer, POS tagger, or Named Entity recogniser.**

## BACKGROUND

The GliCom research group have interests in descriptive and theoretical research in the areas of Phonetics and Phonology, Morphology, Syntax, Semantics, Pragmatics and Discourse Structure, as well as the computational strategies involved in processing them (symbolic versus statistical approaches, finite state automata, context-free and mildly-context sensitive grammars, recent advances in Categorical Grammar, Unification Grammars, Constraint-based systems, computational techniques for speech processing...).

Also in development of linguistic tools applied to machine translation, text error correction, text-to-speech, and information retrieval.

## THE TECHNOLOGY

The lexicon is distributed in two sublexicons:

1. word forms
2. verb-clitic combinations

The list of wordforms 1,152,242 entries, including (i) regular words (1,144,086), (ii) toponyms and anthroponyms (8,032), (iii) abbreviations and acronyms (775), and (iv) computational terms (124). Each entry consists of: form, lemma, morphosyntactic tag and the word type.

The list of verb-clitic combinations contains 4,283,637 entries, exhaustively covering all formal combinations (including infinitive, gerund and imperative). Note that some clitic combinations may be formally possible although semantically implausible. Each entry consists of: form, lemma of the verb and combination of morphosyntactic tags of the verb and the pronoun(s).

## STATE OF DEVELOPMENT

Fully developed. The lexicon is formatted as plain text in a tab separated file.

## INTELLECTUAL PROPERTY

© UPF 2015

## MARKET OPPORTUNITY

Database available to interested parties, both for research and commercialization purposes. The database can be distributed in parts or in a whole.



## COMMERCIAL OPPORTUNITY

Computational lexicon available for licensing with technical cooperation

## CONTACT

Marc Santandreu  
Technology Transfer Unit  
(+34) 93 542 28 96  
marc.santandreu@upf.edu

## KEYWORDS

Text analysis, wordform, lexicon, Spanish

## SEE MORE TECHNOLOGIES AT:

<http://knowledge.upf.edu/>