FreeLing 1.3: Syntactic and semantic services in an open-source NLP library

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, M. Padró

TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
{batalla,bcasas,ecomelles,mgonzalez,padro,mpadro}@lsi.upc.edu

Abstract

This paper describes version 1.3 of the FreeLing suite of NLP tools. FreeLing was first released in February 2004 providing morphological analysis and POS tagging for Catalan, Spanish, and English. From then on, the package has been improved and enlarged to cover more languages (i.e. Italian and Galician) and offer more services: Named entity recognition and classification, chunking, dependency parsing, and WordNet based semantic annotation.

FreeLing is not conceived as end-user oriented tool, but as library on top of which powerful NLP applications can be developed. Nevertheless, sample interface programs are provided, which can be straightforwardly used as fast, flexible, and efficient corpus processing tools.

A remarkable feature of FreeLing is that it is distributed under a free-software LGPL license, thus enabling any developer to adapt the package to his needs in order to get the most suitable behaviour for the application being developed.

1. Introduction

We present a demo of FreeLing, an open-source library which provides basic NLP services (lemmatizing, POS tagging, chunking, NE recognition, etc.) to NLP application developers. The package also includes a front-end that enables the final user to analyze unrestricted texts.

FreeLing was first presented at LREC'04 (Carreras et al., 2004), as a suite of analysis tools released as free software, under GNU Lesser General Public License (Free Software Foundation, 1999). Version 1.2 was released on October 2004, and from then to February 2006, near 1,000 downloads were registered, which yields an average of almost 60 downloads per month.

The first FreeLing version provided morphological analysis and POS tagging for Spanish, English and Catalan. Version 1.3, presented here, extends functionalities existing up-to-date, and incorporates chart-based chunking/parsing, quantity recognition (currency, ratios, physical magnitudes...), sense annotation, named entity classification, and dependency parsing. Also, new languages –namely, Italian and Galician– have been included in the package.

In this paper, we present the LREC-2006 demo of the latest FreeLing release (1.3), which extends significantly the capabilities the library had so far. On the one hand, the morphological level is improved with better expression recognizers (e.g. Physical magnitude detection, enhanced suffix management) and syntactic processing is enhanced with a dependency parsing module capable of annotating syntactic functions. On the other hand, semantic processing is introduced with modules performing Named Entity Classification and WordNet1.6-based (Fellbaum, 1998; Vossen, 1998) sense anotation, as well as a simple most-frequent-sense semantic disambiguator.

FreeLing 1.3 maintains its open and flexible philosophy, and as previous versions, enables fast and accurate linguistic processing of English, Spanish, and Catalan texts. Additionally version 1.3 covers also Italian and Galician (includes morphological analysis and POS tagging for both of them, and syntactic processing for the later).

In the case of Spanish and Catalan, the inclusion of WordNet-based semantic annotation turns FreeLing into the first semantic resource for those languages publicly available under an open-source license.

In addition, we want to remark the Free Software condition of FreeLing, which is distributed under LGPL. This feature facilitates its portability to new languages, and the customization to special user needs. So, we believe that this system constitutes a valuable resource for NLP community (as the number of downloads for version 1.2 prove), both for research (all improvements made to the analyzers will be available to the community), and for commercial usage (LGPL license enables the use of the analyzers as a library component in commercial systems).

2. FreeLing Architecture

In (Carreras and Padró, 2002) we presented a client-server architecture for NLP applications aiming to ease the integration of language analysis services into the development of higher level application.

This architecture consists of a simple two-layer, client-server approach: A basic linguistic service layer which provides analysis services (morphological analysis, tagging, parsing, ...), and an application layer which, acting as a client, requests the desired services from the analyzers.

In this scenario, integrating the basic analyzers in a new NLP application is reduced to three simple steps:

- Convert the data from application internal representation to the service API data structures.
- Call the service and obtain the results.
- Convert the results to the application internal representation.

The advantages of this architecture are:

- It enables to use the analyzer as a function call from any NLP application, not as a separate software pack-

age. This is a crucial issue for modern NLP, specially for high level application development.

- The clients requesting analysis services may be not only NLP applications, but also other service-providing modules (e.g. a parsing module might request a POS tagging service). This enables the construction of increasingly more complex language analysis servers.
- It becomes unnecessary to define data interchange formats between analyzers. Each application can choose its own representation, provided it knows how to map it to the necessary data structures or parameters when requesting a service.
- Conversions are performed between client application data structures and server library data structures, being unnecessary to define data interchange formats between analyzers, and dramatically reducing the overhead caused by the reading, writing, parsing, and transmitting of text-based representations such as XML, SGML. Note that this doesn't mean that the client application has to adapt its input/output formats or internal representations. Provided the library is accessed via its API, the client application may handle the data at will.
- The linguistic processors do not need to be initialized for each piece of text to be analyzed.
- The application may decide how and when to invoke each analyzer, and on which text segment (i.e. there is no need of a whole-text pipelined processing).
- The client-server approach enables the interaction between objects via some standard distributed object middleware, such as CORBA (Common Object Request Broker Architecture) (Object Management Group, 2001), which makes it possible to distribute applications over a network, activate several instances of the same service, if necessary, as well as executing on any platform client applications written in any programming language.

In (Carreras et al., 2004) we presented FreeLing, the first version of our open-source suite of basic language analyzers following the above described philosophy. In this demo, we present version 1.3, with new languages and new linguistic services.

3. FreeLing 1.3 Features

Version 1.3 of the suite, presented in this paper, provides the following features:

- Tokenization.
- Sentence splitting.
- Morphological analysis, with advanced suffix handling (diminutive, appreciative, clitic pronouns, etc.)
- Date-time expression recognition.
- Currency expression recognition.
- Numerical expression recognition (numbers, quantities, percentages, ratios, etc.).
- Physical magnitude expression recognition: Speed (e.g. 120 Km/h), length (e.g. 23 cm.), pressure (e.g. 12.3 in/ft2), frequency, density, power, etc.
- Part-of-Speech tagging. Two algorithms are provided: a HMM trigram model following (Brants, 2000), and a relaxation labelling model based on (Padró, 1998) which enables the use of hand-written rules together with the statistical models.
- Retokenization after POS tagging. Some words in latin languages can be splitted once their Part-of-speech is known. For instance, the word *vela* in Spanish may be a noun (*candle*) but it also may mean *see her* if it is interpreted as an imperative form of the verb *ver* (to see) plus the enclitic pronoun *la*. The suffix handler is now able to detect this cases, and enrich the analysis with the necessary information. After tagging, when the category is known, the word may be splitted to ease the syntax steps, or simply to explicit the information.
- Chart Parser, a reimplementation of (Atserias and Rodríguez, 1998).
- Dependency parser, as described in (Atserias et al., 2005).
- Sense annotator based on WN1.6 for English, Spanish, and Catalan, as well as most-frequent-sense word sense disambiguation.
- Named Entity detection and classification. The classification module is based on Machine Learning Techniques, namely, the AdaBoost-based system winner of CoNLL'02 (Carreras et al., 2002).
- Inclusion of Italian and Galician.

3.1. The Machine Learning components

The Named Entity Classification in FreeLing 1.3 is based on Machine Learning techniques, namely, the AdaBoost algorithm (Schapire and Singer, 1999) as used in (Carreras et al., 2002). This algorithm, as most ML based methods, requires the representation of the sentence to be annotated into a feature vector representation, which is achieved via a general feature extraction module based on Relational Generation Functions (Cumby and Roth, 2003).
This services are also accessible to the application using FreeLing. So, one application could use the library not as a language analysis server, but as a feature extraction and Machine Learning services layer.
We think that this service is relevant enough for general purposes as to be offered in a near future from a standalone library, which could be enriched with more ML methods and richer feature management.

3.2. The inclusion of new languages

We have repeatedly claimed that FreeLing architecture, which tries to keep program code and linguistic data as independent as possible, makes it possible to easily integrate new languages in the library.
In this version, we have proved this claim, integrating two new languages at a very low labour cost.
Both Italian and Galician were included in the suite, following the steps:

- Obtain a freely available morphological dictionary. For italian, we used Morph-it![1], and for Galician,

[1] `http://sslmitdev-online.sslmit.unibo.it`

the morphological dictionary developed by Seminario de Lingüística Informática[2] at Universidade de Vigo for the OpenTrad[3] project. Both dictionaries are distributed under an open Creative Commons license.

- Obtain some POS tagged disambiguated corpus. The authors of the dictionaries kindly provided us such corpus. We used 100,000 words for Italian and 25,000 for Galician.
- Program some scripts to map the original morphological information into POS tags that can be used in FreeLing. We followed PAROLE standard, as for Spanish and Catalan.
- Use those scripts to map the morphological dictionaries, and the tags in the training corpus.
- Train the taggers using the corrected corpus.
- Include the dictionaries and the tagging statistics into FreeLing package.
- Adapt to each languages the rules that control FreeLing modules behaviour (tokenizer, multiword recognizer, suffix handler, etc.)

The integration of Galician and Italian costed about 10 work days of a computer engineer familiar with FreeLing but with no knowledge of Italian nor Galician (although native speaker of Catalan and Spanish). Sufixation rules for Galician were written by a linguist in 1 work day.

3.3. Service Homogeneity across Languages

The extensions incorporated in version 1.3 are not homogeneous, and –as in any free software project– depend greatly on external collaboration, and often, more on legal than technical constraints.

For instance, WN-based semantic annotation is only availble for Spanish, Catalan and English, but not for Italian or Galician, since there is no freely availble version of WN for these langauges (even a reduced version, as is the case of Catalan and Spanish).

Similarly, some features are not available for some languages due to a lack of man power, resources, or simply collaboration from the community. E.g., chunking is not available for English nor Italian, since nobody wrote or adapted the necessary context free grammar, and named entity classification is not available in most languages due to the lack of annotated training corpus.

Nevertheless, the present version constitutes a powerful and easily customizable and extendable language analysis tool, which we are proud to present to LREC-2006.

4. FreeLing 1.3 in Figures

The Spanish and Catalan morphological dictionaries are rather smaller than the others, but since they contain the most frequent lemmas, they are expected to cover all closed category tokens plus over 80% of open–category tokens of unrestricted text.

- The English dictionary was automatically extracted from WSJ, with minimum manual post-edition, and thus may be a little noisy. It contains over 160,000 forms corresponding to some 102,000 different combinations lemma-POS .
- The Spanish and Catalan dictionaries are hand build, and contain the 6,500 most frequent open-category lemmas for each language, plus all closed-category lemmas. The Spanish and Catalan dictionaries try to maintain the same coverage (that is, the same lemmas are expected to appear in both dictionaries). The Spanish dictionary contains over 81,000 forms corresponding to more than 7,100 different combinations lemma-POS , and the Catalan one contains near 67,000 forms corresponding to more than 7,400 different combinations lemma-POS .
- Italian dictionary contains over 355,000 forms corresponding to over 36,000 lemma-POS combinations.
- Galician dictionary contains more than 90,000 forms, corresponding to near 7,400 lemma-POS combinations.

In all cases, unknown words are handled via conditional probabilities of POS tags given word suffixes, following the proposal of (Brants, 2000), so that the most suitable POS tags are proposed for each word not included in the morphological dictionary. Each unknown words is assigned 2.5 tags in average, and over 99% of them get the right tag among those proposed.

The basic POS tagger is a classical trigram HMM tagger in the style of (Cutting et al., 1992; Brants, 2000), trained on WSJ for English, and on 100,000 words of hand-disambiguated corpus for Spanish, Catalan and Italian, and on a 25,000 word corpus for Galician. The tagger provides a precision near 97% for Spanish, Catalan, English and Italian, and about 95% for Galician.

Also, another tagger is provided, based on (Padró, 1998). Although the performances are similar, this second tagger offers the possibility of merging hand–written rules with the statistical model.

The system is able to morphologically analyze a text at a speed near 6,000 words/second in a P4 2.8 GHz processor. The POS tagger disambiguates the morphological analyzer output at a speed of 3,100 words/sec. When performing both tasks simultaneously on the same processor, the speed is 2,300 words/sec.

The Named Entity Recognition module is a very naive pattern recognizer, which searches for capitalized words, allowing some functional words to be considered as part of a NE. The F_1 of this module is about 90%. The Named Entity Classification module has an accuracy about 91% when applied over a perfect NE detection. When both modules are combined, the NER+NEC perfomance is about $F_1 = 82\%$.

5. Main External Contributions

Many people apart from the original authors contributed to by reporting problems, suggesting various improvements, submitting actual code, extending linguistic databases, or simply, allowing us to use their linguistic data.

This is a list of the most remarkable of these contributions. The order is not relevant.

[2] http://sli.uvigo.es

[3] http://www.opentrad.org

- Mikel Forcada and the InterNostrum[4] team in Universitat d'Alacant completed the Spanish and Catalan dictionaries to cover the same lemas in both languages, enlarging the dictionaries from 5,000 to 6,500 lemmas.
- TALP and CLiC research centers and the NLP research group[5] at UNED (Universidad Nacional de Educacion a Distancia), who developed the Spanish WordNet in the framework of EuroWordNet and Meaning projects, granted the distribution of the synsets for the lemmas included in FreeLing Spanish dictionary.
- TALP and CLiC, who developed the Catalan WordNet, granted the distribution of the synsets for the lemmas included in FreeLing Catalan dictionary.
- The feature extraction module is based on the code developed by Dan Roth's Cognitive Computation Group[6] at University of Illinois at Urbana Champaign (UIUC), who we thank for allowing us to distribute our modified version under LGPL.
- The English WordNet[7] was developed by the Cognitive Science Laboratory at Princeton University. Synset information is included in FreeLing under the original WordNet license terms.
- The Italian dictionary is extracted from Morph-it!, developed by Marco Baroni and his colleagues at the Scuola Superiore di Lingue Moderne per Interpreti e Traduttori[8] of the University of Bologna. These data are included in FreeLing under their original Creative Commons license.
- The Galician dictionary was obtained from the OpenTrad project, and was developed by Xavier Gómez Guinovart and the members of the Seminario de Lingüística Informática at Universidade de Vigo. These data are included in this package under their original Creative Commons license. These researchers also took an active role in the creation and debugging of the Galician morphological data and rulesets.

6. Some Internal Details

The internal architecture of the system is based on two kinds of objects: linguistic data objects and processing objects.

6.1. Linguistic Data Classes

The basic classes in the library are used to contain linguistic data (such as a word, a POS tag, a sentence, a document...). Any client application must be aware of those classes in order to be able to provide to each processing module the right data, and to correctly interpret the module results.
The linguistic classes supported by the current version are:

- `analysis`: A tuple <lemma, POS tag, probability, senses>.
- `word`: A word form with a list of possible analysis.
- `sentence`: A list of words known to be a complete sentence, it may include also a parse tree and/or a dependency tree.

Figure 1 presents a UML diagram with the linguistic data classes.

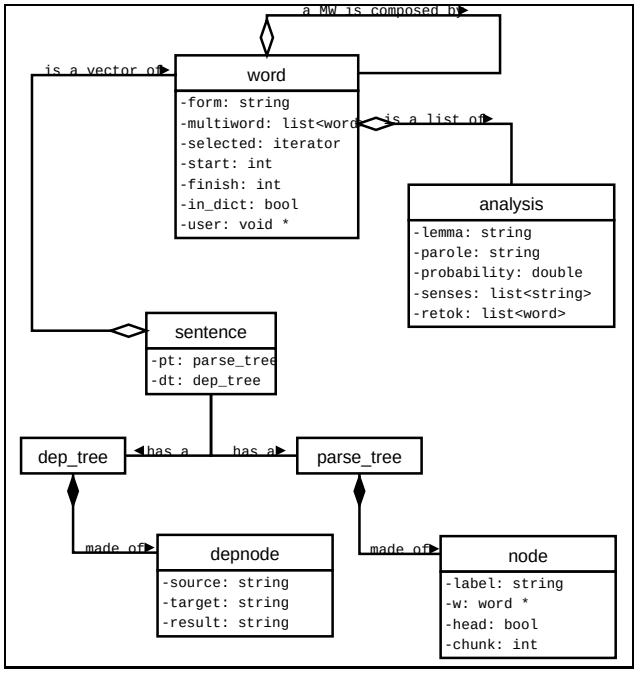

Figure 1: FreeLing-1.3 Linguistic Data Classes.

6.2. Processing Classes

Apart from classes containing the linguistic data, the library provides classes able to transform those data:

- `tokenizer`: Receives plain text and returns a list of `word` objects.
- `splitter`: Receives a list of `word` objects and returns a list of `sentence` objects.
- `morfo`: Receives a list of `sentence` and morphologically annotates each `word` object in the given sentences. In fact, this class applies a cascade of specialized processors (number detection, date/time detection, multiword detection, dictionary search, etc.) each of which is in turn a processing class:
  - `locutions`: Multiword recognizer.
  - `dictionary`: Dictionary lookup and suffix handling.
  - `numbers`: Numerical expressions recognizer.
  - `dates`: Date/time expressions recognizer.
  - `quantities`: Ratio and percentage expressions and monetary amount recognizer.
  - `punts`: Punctuation symbol annotator.
  - `probabilities`: Lexical probabilities annotator and unknown words handler.
  - `np`: Proper noun recognizer.
- `tagger`: Receives a list of `sentence` objects and disambiguates the POS of each `word` object in the

[4] `http://www.internostrum.com`
[5] `http://nlp.uned.es`
[6] `http://l2r.cs.uiuc.edu/~cogcomp`
[7] `http://wordnet.princeton.edu`
[8] `http://www.ssit.unibo.it`

given sentences. If the selected analysis carries retokenization information, the word may be splitted in two or more new words.

- NE `classifier`: Receives a list of `sentence` objects and classifies all `word` objects tagged as proper nouns in the given sentences.
- `Sense annotator`: Receives a list of `sentence` objects and enriches with synset information the `analysis` chosen by the tagger for each `word` object.
- `chunk parser`: Receives a list of `sentence` objects and enriches each of them with a `parsed_tree` object.
- `dependency parser`: Receives a list of parsed `sentence` objects and enriches each of them with a `dependency_tree` object.

Figure 2 presents a UML diagram with the procesing classes.

The client application is free to decide in which format wants to input, output or store its linguistic data, and only has to translate it to the classes described above when interacting with the library. Also, the client application is free to decide for which processing steps the library is going to be used –e.g. the application may require a tagger for Spanish but not for Catalan, or may want to call directly the morphological analyzer skipping tokenization and splitting steps, or may want to instance only a date/time expressions recognizer, without using any other functionality, etc.

7. Conclusions and Further Work

We have presented FreeLing 1.3, the most recent version of this Open Source Language Analysis Suite.

The previous version (1.2) had near 1,000 downloads between October 2004 and February 2006, averaging 60 downloads/month, which indicates that FreeLing already constitutes a valuable resource for NLP community.

With the inclusion of Named Entity Classification and WN-based semantic annotation, we introduce semantics in the library, and offer the first open-source general-purpose semantic resource for Spanish and Catalan.

Also, the integration of Italian and Galician at a very low cost, prove that our approach is really flexible and allows the easy extension with new languages, provided a morphological lexicon and a POS tagging training corpus are available.

Future versions of the analyzer library will provide more functionalities and improve the already existing features. We are specially interested on improving semantic processes such as Word Sense Disambiguation or Semantic Role Labelling, as well as in developing syntax and dependency parsing for English.

Also, the current ML engine will probably be build as a standalone general purpose library, and extended to include more algorithms than those currently supported, and a more flexible feature management.

8. Acknowledgments

This work has been partially funded by the European Union through the MEANING project (IST-2001-34460) and by the Spanish Industry Department through the OpenTrad project (FIT-340101-2004-0003, FIT-340001-2005-2). We also want to thank all FreeLing users for their valuable feedback and contributions.

9. References

Jordi Atserias and Horacio Rodríguez. 1998. Tacat: Tagged corpus analizer tool. Technical report lsi-98-2-t, Departament de LSI. Universitat Politècnica de Catalunya.

Jordi Atserias, Eli Comelles, and Aingeru Mayor. 2005. Txala: un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, (35):455–456, September.

Thorsten Brants. 2000. Tnt - a statistical part- of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing, ANLP*. ACL.

Xavier Carreras and Lluís Padró. 2002. A flexible distributed architecture for natural language analyzers. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC*, Las Palmas de Gran Canaria, Spain.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2002. Named entity extraction using adaboost. In *Proceedings of CoNLL Shared Task*, pages 167–170, Taipei, Taiwan.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

C. Cumby and Dan Roth. 2003. Feature extraction languages for propositionalized relational learning. In *IJCAI Workshop on Learning Statistical Models from Relational Data*. http://l2r.cs.uiuc.edu/ danr/Papers/CumbyRo03a.pdf.

Doug Cutting, Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. 1992. A practical part–of–speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, ANLP*, pages 133–140. ACL.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.

Free Software Foundation. 1999. Lesser public general license. License conditions, Free Software Foundation. See http://www.gnu.org/licenses/licenses.html.

Object Management Group. 2001. Common object request broker architecture. Technical document, Object Management Group. See http://www.omg.org, http://www.corba.org.

Lluís Padró. 1998. *A Hybrid Environment for Syntax–Semantic Tagging*. Ph.D. thesis, Dep. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, February. http://www.lsi.upc.es/˜padro.

R. E. Schapire and Y. Singer. 1999. Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*, 37(3).

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

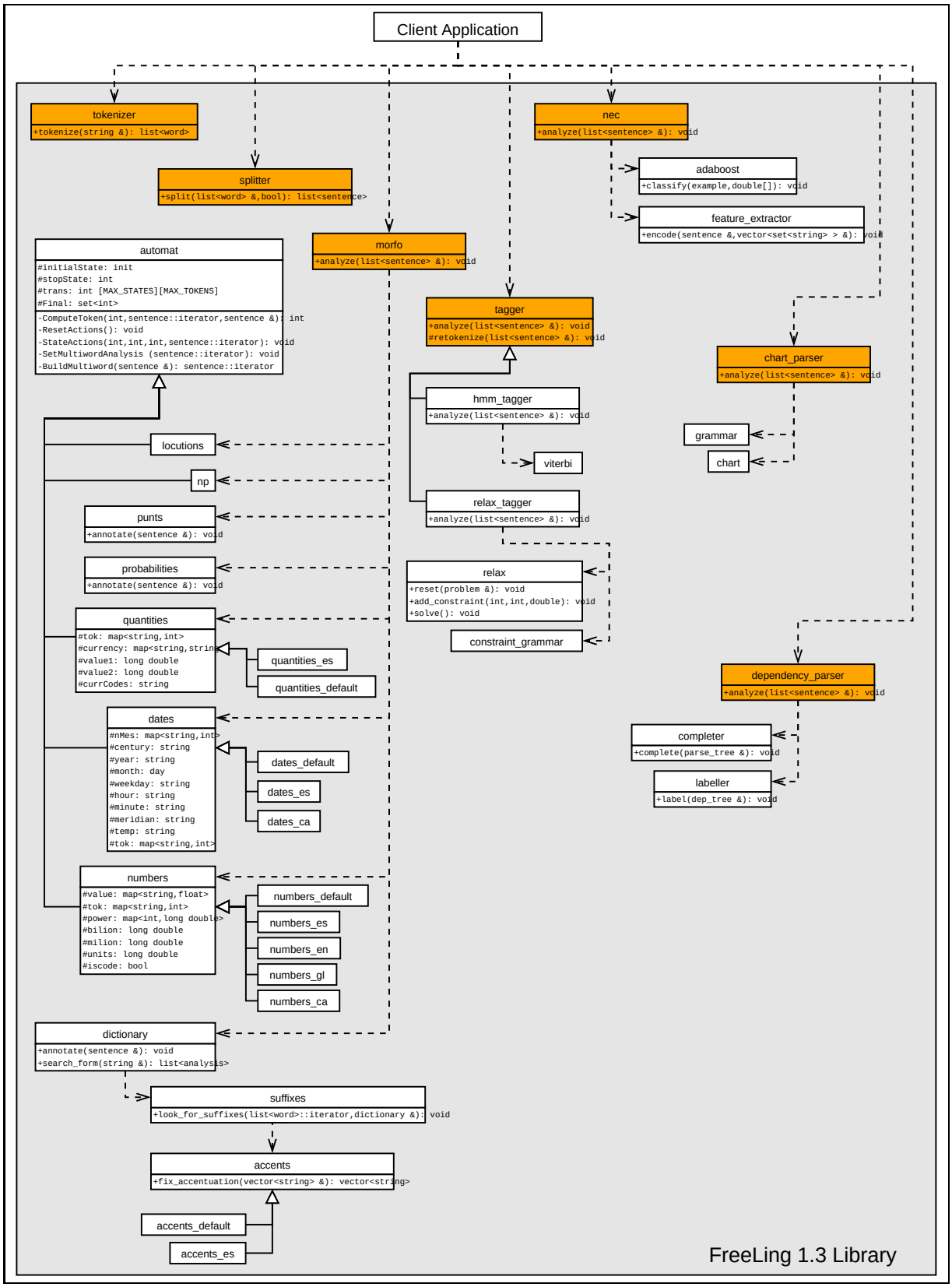

Figure 2: FreeLing-1.3 Main Processing Classes.