Project Monitoring Sessions, 2007
National Programme on Computer Technology (TIN)

Development of Support Services for Linguistic Research over the Internet (Project TIN2004-03988)

Octavio Santana Suárez[*]
Main Researcher of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

José R. Pérez Aguiar[**]
Research Member of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Francisco Javier Carreras Riudavets
Research Member of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Zenón J. Hernández Figueroa
Research Member of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Juan Carlos Rodríguez del Pino
Research Member of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Margarita Díaz Roca
Research Member of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Juan de Dios Duque Martín de Oliva
Research Member of the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Gustavo Rodríguez Rodríguez
Collaborator on the Project
Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Abstract

The objective of this project is to place a set of remote services and clients at the disposal of the international community over the Internet in order to computationally solve linguistic phenomena of the Spanish language. The implemented services are as follows: a remote service of morphological analysis, a remote service of information on morpholexical relationships and a remote service of functional disambiguation. These services allow access to any authorized remote application by means of the inclusion of the corresponding definition document. Additionally, a client of morphosyntactic analysis of texts and a morpholexical client of information recovery have been developed. Both clients are end-use tools that put at stake the potentiality of services.

Keywords: natural language processing, computational linguistic, morphology, syntax.

[*] E-mail: osantana@dis.ulpgc.es
[**] E-mail : jperez@dis.ulpgc.es

1 Aims of the project

Remote services and their clients are innovative technology, which are based on the use of open standards so as to promote cooperative development by taking full advantage of the potentiality of the Internet. In accordance with point 3.6 of the National Programme on Computer Technology, this project aims at developing:

S1. A remote service of morphological analysis.
This service enables the lemmatization of any word of the Spanish language when identifying its canonical form, grammatical category and the inflection or derivation that produces it. With regard to verbs, the service considers the simple and compound conjugation, enclitic pronouns, the inflection of the participle as a verbal adjective (gender, number) and the diminutive form of the gerund. As far as non-verbal forms are concerned, the service considers the following: gender and number for nouns, adjectives, pronouns and articles; heteronyms due to a change of gender for nouns, superlative degree for adjectives and adverbs, adverbialization of the superlative for adjectives; derivation for nouns, adjectives and adverbs; invariable forms such as prepositions, conjunctions, exclamations, words from other languages and idiomatic expressions or phraseology. Prefixation is taken into consideration when applicable.

S2. A remote service of information on morpholexical relationships.
This service enables the recognition, generation and manipulation of the morpholexical relationships from any given word. It includes the retrieval of all its lexicogenetic information until reaching the primitive, the management and control of the affixes in the treatment of its relationships and regularity in the established relationship. It conveys a global perspective of the behaviour and productivity of Spanish words in the main word-formation processes (suffixation, prefixation, parasynthesis, suppression, regression, zero-modification, apocopation, metathesis and other non-classifiable processes that generate alternative graphical forms).

S3. A remote service of functional disambiguation.
This service offers the grammatical function of every voice in its respective context. It minimizes the possibilities due to its treatment of the syntactic structures.

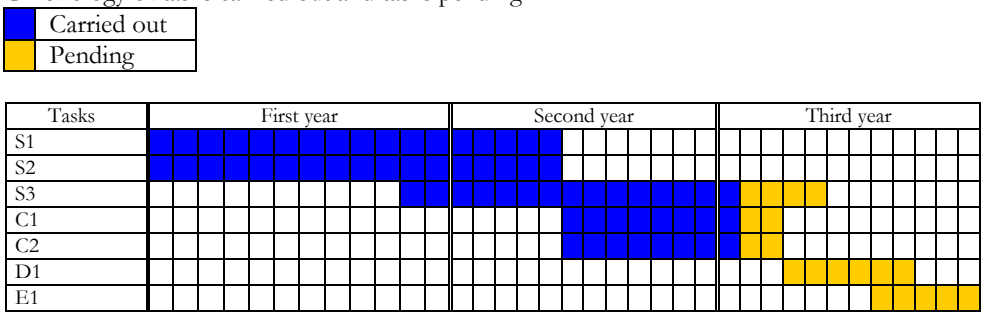
C1. A client of morphosyntactic analysis of texts.
Using the previous services and a user-friendly interface, this client allows users to obtain the morphosyntactic analysis of the chosen text; some statistical measures of its characteristics; the neologisms; and the location of grammatical co-occurrences, verbal periphrasis, lexical collocations and other linguistic phenomena.

C2. A morpholexical client of information recovery.
Using the previous services, this client can locate the documents that meet the specified search requests on the Internet, for instance, specific words affected to a greater or lesser extent by the various existing word transformation mechanisms in the Spanish language, or grammatical characteristics or linguistic phenomena that may appear in the document.

D1. Technical and user documentation.
The technical and user documentation are of utmost importance in order to offer the service or distribute the client to potential users with the assurance that they will be able to use it effectively.

E1. Control of the service or client under real conditions.
Control of the service or client under real conditions. Once the service is offered or the client is distributed to its users, based on the comments they make, a feedback phase focusing on any of the previous tasks will take place. This will entail adjustments and corrections to its operation. Moreover, this phase lasts indefinitely, although, for the purposes of this project, an initial phase of limited duration, during which said feedback will be stronger, is estimated.

Chronology of tasks carried out and tasks pending.



2 Success level achieved in the project

Of the aforementioned objectives, those concerning the remote service of morphological analysis and the remote service of information on morpholexical relationships have been met. Work is currently underway on the remote service of functional disambiguation and on the two clients that have a high percentage of performance. Still pending is the technical and user documentation that was initially going to be prepared at the end of every service but which is now going to be moved to the final year of the project in order to homogenize it and draft it when all services and clients are operating. Accordingly, human resources have been allocated to other more complex tasks in order to complete them within the intended timeline and to efficiently solve the problems encountered.

2.1 Remote service of morphological analysis (S1)

The objective of the morphological service is to support the morphological study of words over the Internet. It is developed for the ".NET" platform as an "ASP.NET" web service that uses C# as the development language in the Microsoft Visual Studio setting. This allows for any authorized remote application to access the service by means of the inclusion of the corresponding definition document in a *Web Services Description Language* (WSDL) format. This can considerably encourage advances in research, since other groups whose work needs to rely on the morphology of the

Spanish language can develop their own applications by using this service, without having to develop their own morphological tools

This service is based on a morphological engine, implemented in C++ as a *Dynamic Link Library* (DLL) of MS-Windows, whose potential is endorsed by the 4,951,802 inflected words taken from all entries in the Dictionary of the Spanish Language (DRAE), the General Dictionary of the Spanish Language (VOX), the Dictionary of Current Spanish Usage (CLAVE SM), the Dictionary of Synonyms and Antonyms (ESPASA CALPE), the Ideological Dictionary of the Spanish Language (JULIO CASARES) and the Dictionary of Contemporary Spanish Usage (MANUEL ALVAR EZQUERRA). Furthermore, in the region of 1240 proper nouns not included as entries in the consulted sources are added. These are related to nationality adjectives and other adjectives and nouns, such as 'follower of' or 'doctrine' —Marxist, Marx's Marxism— among other meanings. In the region of 9000 adjectives derived from verbal participles that have not been included in said sources have also been added. The universe contemplated by the service is made up of 195,743 canonical forms —181,593 non-verbal and 14,150 verbal— that have generated 4,951,802 inflected or derived forms. Moreover, this does not contemplate the extension inherent to the incorporation of prefixes and enclitic pronouns —more than 4000 million words— that the morphological analysis system also includes at a success rate of 100%. With the exception of the recognition of prefixes and enclitic pronouns, it is implemented on the basis of data and not rules. The database system can easily be updated to include possible changes in language or errors resulting from manual processing. The service operates with a word, supplied by a remote client, that is lemmatized in order to return its morphological characteristics by means of a SOAP (*Simple Object Access Protocol*).

2.2 Remote service of information on morpholexical relationships (S2)

As regards the construction philosophy, this service has the same characteristics as the morphological analysis service described in the previous section: an "ASP.NET" service written in C# that is supported on a DLL of morpholexical relationships written in C++.

There is a morpholexical relationship between two Spanish canonical forms when one has been formed from the other by means of a word formation process in the Spanish language: suffixation, prefixation, parasynthesis and others. There is a suffixal morpholexical relationship between two words when one has been formed from the other by adding a suffix; thus its semantic and functional aspect is generally altered —footballer has a semantic and functional relationship with football. There is a prefixal morpholexical relationship between two words when one has been formed from the other by adding a prefix; thus its semantic aspect is slightly altered. Incorporating prefixes does not usually alter the functionality of the root word —antechamber has a semantic and functional relationship with chamber. There is a parasyntetic morpholexical relationship between two words when one has been formed from the other by adding an affix in front of the word and another one behind it —usually a prefix and a suffix—; thus the semantic and functional aspect of the new word varies with respect to the root word —irrationalism has a semantic and functional relationship with rational. Another kind of morpholexical relationship is established if two words have undergone one of the following processes: suppression, regression, zero-modification, apocopation, metathesis and other non-classifiable processes that generate alternative graphical forms. The information engine of morpholexical relationships —implemented according to data— is capable of obtaining information on approximately 70,147 suffixal morpholexical relationships,

11,113 prefixal morpholexical relationships, 3812 parasyntetic relationships and 4694 morpholexical relationships of other kinds.

This service enables selecting between three options that give access to various degrees of information on words that are morpholexically related to some data. The service operates with a word and its grammatical category that can be obtained from the morphological analysis service. The result is a list of words which, according to the chosen option, will only be formed by the nearest ones, those with an intermediate degree of relationship or all those that are in some way related.

2.3 Remote service of functional disambiguation (S3)

As far as the construction philosophy is concerned, this service that is still being developed has the same characteristics as the service described in the previous section: an "ASP.NET" service written in C# that is supported on a DLL of disambiguation and on a DLL of morphological analysis that uses the remote service of morphological analysis.

There is a considerable amount of words in the Spanish language that can have various grammatical functions and, consequently, a textual analysis could produce a great amount of combinations unless the function of each word within the context in which it appears is considered. Functional disambiguation consists of eliminating the results that do not correspond to their function in the text.

This service uses a functional disambiguation method that reduces the size of the answer of the morphological processor into two phases. In the first phase, a functional disambiguation method based on local syntactic structures is applied and those grammatical functions that are incompatible with the immediate context of each word in the phrase are ruled out. In the second phase, a structural functional disambiguation is carried out and the combination of grammatical functions of the phrase that do not produce a valid syntactical representation tree are ruled out.

Studies are currently focusing on whether applying the second phase is worthwhile, since the loss of performance inherent to the application of the second phase does not seem to justify its inclusion even though a disambiguation percentage of 87% can be achieved with the first phase. This figure reaches 96% after applying the second phase.

2.4 Client of morphosyntactic analysis of texts (C1)

The client of morphosyntactic analysis of texts is developed as an "ASP.NET" web application that interacts with the user on the one hand, and with the remote services described above on the other.

This application offers the user two possibilities: writing a text directly in the window for that purpose or extracting it from a file. Once the text to be analysed has been selected, the application can locate recognised and non-recognised words, indicating where they appear and their occurrences; look for co-occurrences of various words within a given radius; find occurrences of known idiomatic expressions; and analyse the occurrence of text segments with a specified minimum degree of repetition.

2.5 Morpholexical client of information recovery (C2)

The objective of this client is make it possible for information to be recovered from the Internet, selected by means of a search pattern that includes inflective and derivative morpholexical characteristics, offering far greater scope than that offered by most search engines whose inflective capacity is reduced.

The client is developed as an MS-Windows application that can be installed onto the user's computer and connected to the previously described services over the Internet. Carrying out a search requires a configuration of two different areas: the field and the pattern of the search. The field is understood to be the set of websites from where the documents to be analysed will be taken; those documents fitting to the specified pattern will be selected, for instance, the search can be focused on documents that fulfil a given pattern, specifying as the field of the search all Spanish universities, just the University of Las Palmas de Gran Canaria or even a department of said University. The pattern of the search can vary from an exact word to the inclusion of a great deal of characteristics based on linguistic phenomena, for instance, documents containing the sequence of words "inflation housing youth" can be located where the words of the sequence can appear inflected or derived in proportions indicated by the user and separated from one another by a certain distance.

3 Results Indicators

Two books on morpholexical relationships have been published:

- *Relaciones morfoléxicas prefijales para el procesamiento del lenguaje natural.*
  Santana, O.; Carreras, F.; Pérez, J.
  Editorial MILETO; ISBN: 84-95282-92-5. Madrid, 2005.
  Pages: 116.
- *Relaciones morfoléxicas parasintéticas para el procesamiento del lenguaje natural.*
  Santana, O.; Carreras, F.; Pérez, J.
  Editorial MILETO; ISBN: 84-95282-96-8. Madrid, 2006.
  Pages: 156.

The following articles have been published in national and international journals:

- NAWeb: un navigateur et analyseur morphologique des pages web pour l'espagnol.
  Santana, O.; Hernández, Z.; Rodríguez, G.
  Cahiers de lexicologie. Revue internationale de lexicologie et de lexicographie, issue 87- 2005-2. ISSN: 0007-9871 (29/43).
  2005.
- Functional Disambiguation Based on Syntactic Structures.
  Santana, O.; Pérez, J.; Losada, L.; Carreras, F.
  Literary and Linguistic Computing, Vol. 21, issue 2, (187/197).
  2006.

The following papers have been presented at international conferences:

- Una aplicación para el procesamiento de la sufijación en español.
  Santana, O.; Carreras, F.; Pérez, J.; Rodríguez, G.

IX Simposio Internacional de Comunicación Social, *Actas*, Vol. II. ISBN: 959-7174-05-7, (623/629).
January, 2005.

- Software Application for Parasynthesis in Spanish Automatic Processing.
  Santana, O.; Carreras, F.; Pérez, J.; Rodríguez, J.C..
  The 2005 International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'05. Proceedings. ISBN: 1-932415-73-4, (46/52).
  June, 2005.
- Spanish Morphosyntactic Disambiguator.
  Santana, O.; Pérez, J.; Losada, L.; Carreras, F.
  The 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing. ACH/ALLC 2005. Conference Abstracts. ISBN: 1-55058-308-5, (207/209).
  June, 2005.
- Una Aplicación para el Procesamiento de la Prefijación en Español.
  Santana, O.; Carreras, F.; Pérez, J.; Rodríguez, G.
  4th Multi-Conference on Systems, Cybernetics and Informatics. CISCI 2005. *Memorias*, Vol. II. ISBN: 980-6560-38-8, (322/327).
  July, 2005.
- Parasynthetic Morpholexical Relationships of the Spanish: Lexical Search beyond the Lexicographical Regularity.
  Santana, O.; Carreras, F.; Pérez, J.; Rodríguez, J.
  Proceedings of the IADIS International Conference. Applied Computing. 2006. ISBN: 972-8924-09-7, (627/631).
  February, 2006.

4 Bibliography

[1] Santana, O.; Carreras, F.; Pérez, J.: *Relaciones morfoléxicas sufijales para el procesamiento del lenguaje natural*. Editorial MILETO; ISBN: 84-95282-91-7. Madrid, 2004. (178 pages).

[2] Santana, O.; Carreras, F.; Pérez, J.; Rodríguez, G.: Relaciones morfoléxicas prefijales del español. *Boletín de Lingüística*, Vol. 22. ISSN: 0798-9709, (79/123). July - December, 2004.

[3] Santana, O.; Hernández, Z.; Rodríguez, G.: DAWEB: Un descargador y analizador morfológico de páginas Web. *Procesamiento de Lenguaje Natural*, Revista Nº 30. Ed. SEPLN. ISSN: 1135-5948, (75/87). March, 2003.

[4] Santana, O.; Hernández, Z.; Rodríguez, G.: Morphoanalysis of Spanish Text: Two Applications for Web Pages. *Lecture Notes in Computer Science* 2722. Web Engineering. Ed. Springer-Verlag. ISSN: 0302-9743. ISBN: 3-540-40522-4, (511/514). July, 2003.

[5] Santana, O.; Pérez, J.; Carreras, F.; Duque, J.; Hernández, Z.; Rodríguez, G.: FLANOM: Flexionador y lematizador automático de formas nominales. *Lingüística Española Actual* XXI, 2, 1999. Ed. Arco/Libros, S.L. (253/297).

[6] Santana, O.; Pérez, J.; Carreras, F.; Rodríguez, G.: Suffixal and Prefixal Morpholexical Relationships of the Spanish. *Lecture Notes in Artificial Intelligence*, 3230. Ed. Springer-Verlag. ISSN: 0302-9743, (407/418). October, 2004.

[7] Santana, O.; Pérez, J.; Hernández, Z.; Carreras, F.; Rodríguez, G.: FLAVER: Flexionador y lematizador automático de formas verbales. *Lingüística Española Actual* XIX, 2, 1997. Ed. Arco/Libros, S.L. (229/282).
[8] Santana, O.; Pérez, J.; Losada, L.; Carreras, F.: Bases para la desambiguación estructural de árboles de representación sintáctica. *Procesamiento de Lenguaje Natural*, Revista Nº 32. Ed. SEPLN. ISSN: 1135-5948, (43/65). March, 2004.