# Using a Spontaneous Spoken Corpus in a Bilingual Lexicon of Prepositions [French-Spanish]

Doaa Samy,[1] Ana Valverde Mateos[1]
and Concha Sanz Miguel[2]

**Abstract**

Corpora have proved to be a potential resource for studies undertaken in nearly all branches of linguistics. This work explores the benefits of integrating examples from a spoken spontaneous corpus in a bilingual lexicon of prepositions (French-Spanish). The paper presents a detailed description of the hierarchal structure of the lexicon implemented in XML. For the integration of the examples from the spoken corpus, criteria and difficulties encountered are particularly emphasized. A wide range of applications can benefit from this resource especially Second Language Acquisition and Language Engineering.

## Introduction

Alongside the developments in language research, corpora have been always considered a potential resource in lexicography and computational linguistics. For many years, lexicographers have looked up the corpora to enrich their lexical entries with real examples and empirical evidences of the use of the language. On the other hand, automatic extraction of lexicons from corpora is a common practice in language engineering and computational linguistics. Despite the numerous studies which address the benefits of corpora in these aspects, there is little evidence of the use of spoken corpora in lexicons and dictionaries.

This paper presents a bilingual electronic lexicon of prepositions enriched with examples from a spontaneous spoken corpus C-Oral-ROM (French and Spanish versions) (Cresti & Moneglia, 2005). The novelty of our approach lies in the use of spoken corpora as a valuable resource, highlighting the importance of the contextual information through examples not only from the written language but also from the spoken language. Moreover, a multi-level structure including lexical, morphological, semantic and pragmatic information offers a comprehensive overview of the use of prepositions in both languages.

The paper is organized along a number of inter-related issues that we regard as highly important for the domain of the study. The first section discusses main premises behind our work outlining the main theoretical framework. According to this framework, our approach points out the importance of spontaneous speech corpora as an added value that fits in with the functional perspective of the lexicon. Besides, it reflects the importance of prepositions in written and spoken registers and their role as basic constituents in multi-word expressions.

[1] Computational Linguistics Laboratory, Linguistics Department, Universidad Autónoma de Madrid
*e-mail*: doaa@maria.lllf.uam.es, anaval@maria.lllf.uam.es
[2] French Philology Department, Universidad de Castilla La Mancha
*e-mail*: concha.sanz@uclm.es

The second section describes the methodology adopted for developing the electronic lexicon. Based on a traditional text book of prepositions in French and Spanish (Sanz Miguel, 1999), two main aspects are highlighted. First, the technical part concerning the design of the hierarchical structure of entries in XML, describing how information from different linguistic levels is marked up. Second, the criteria and difficulties encountered in the selection of examples from the spoken corpus (French and Spanish) and how examples are integrated in each entry. The interactive user interface provides the possibility to search the lexical information. The final output includes results of the query together with real examples from textual sources and from the spoken corpora through transcriptions and links to audio files.

The third section focuses on conclusions and future uses of this lexicon as a resource in Teaching Spanish/French as a Foreign Language and the possible integration of multi-modal resources. Furthermore, from the Language Engineering perspective, it can be integrated in POS taggers or Translation memories.

## 1. Prepositions, Lexicons and Spontaneous Spoken Corpora: an intersection point

The subject of the present work 'a lexicon of prepositions enriched with examples from a spontaneous spoken corpus' represents an intersection area which can be studied from different overlapping perspectives. In this section, we will discuss our approach regarding the following perspectives:

- Prepositions as a grammatical category addressed from the classical approach of linguistics;
- Computational lexicography within the field of computational linguistics and language engineering; and finally,
- Spoken spontaneous corpora as a subset of corpora exhibiting particular features of the spoken register of the language.

As mentioned earlier these research lines are not exclusive. The present approach lies in the interface between the three areas: lexicons, corpora and prepositions. Moreover, this area of intersection can be regarded from two different points of view: Applied Linguistics and Language Engineering Resources. The set of relations are reflected in the following diagram. As shown, it is difficult to provide clear cut boundaries to the present work. However, among the set of intersections and overlapping interfaces, lies the novelty of our approach trying to capture the features of these relations by bringing them together into a comprehensive, mutely-level structured hierarchy.

For the purpose of this study and within the general framework of computational lexicography and corpora, prepositions are considered the focal point linking the areas in concern, thus, they are addressed through their relation between corpora on one hand and lexicons on the other. In addition, this relation is discussed taking into consideration both dimensions: the Language Engineering and the Applied Linguistics perspective.

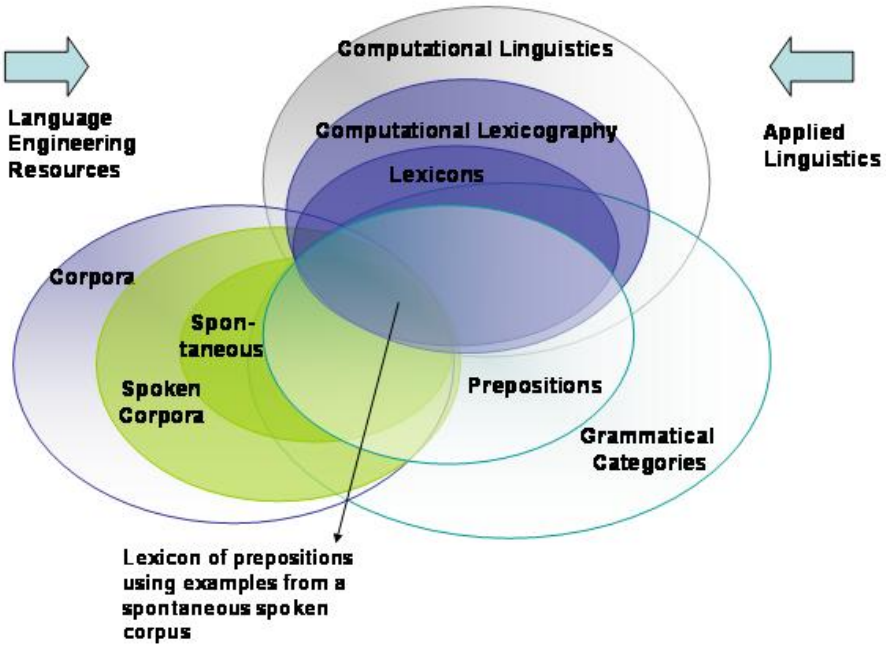


**Diagram 1**: Prepositions, Lexicons and Corpora

### Corpora and Computational Lexicography

The relation between corpora and Computational Lexicography is a bi-directional relation dating back to the first approaches in corpus linguistics. Automatic extraction of lexicons from corpora or using corpora as a resource for providing examples for lexical entries are common practices in the field. In the first case, corpora are the starting point from which lexicons are derived, while in the second case, lexicographic entries are the starting point and corpora are looked up for examples and evidences of use.

Both approaches are feasible depending on the coverage and the target application that would benefit from the lexicon. For example, in applications where lexicography is not the main research area, it is most probable to automatically extract lexicons. On the other hand, in case the main goal is to build up a dictionary, special attention is given to the lexicographic organization of the entries and, therefore, corpora are relegated to a secondary role. The first approach is more common in Language Engineering and Natural Language Processing where linguistic knowledge is integrated in a wide range of technological applications such as Information Retrieval, Information Extraction, Automatic Summarization, Question Answering, Data mining, etc. The second approach reveals the Applied Linguistics perspective where computers and other electronic resources are considered as powerful means that could play an important role in enhancing the study of the Language in its representation at the different levels of analysis: morphological, lexical, syntactic, semantic, etc.

However, most of the research concerned with corpora and lexicons has focused on the written corpora representing “written” registers of the language. The only manifestation of speech in the majority of modern electronic dictionaries and lexicons is

limited to the integration of audio files for pronunciation purposes, reflecting, in this way, the phonetic representation of each lexical entry. Despite of the utility of this feature from the phonetic point of view, it is a mere representation of separate entries completely out of context where their real use in the spoken register is not taken into account.

Recently, "spoken" resources reflecting characteristics of the oral registers have been gaining an increasing attention. Research carried out on spoken corpora addresses different features concerning grammatical, syntactic, semantic and pragmatic issues. Though deemed as a highly valuable resource, spoken corpora have not been widely exploited for lexicographic purposes. Their integration into lexicographic resources could help to gain insight into the structure of the spoken registers of the language and its use in different situations. Moreover and if we consider the "spontaneous" feature of some corpora, this could guarantee a real representation of language in use emphasizing the pragmatic and the phonetic aspects in different situations.

## Prepositions and Corpora

Prepositions are grammatically considered as function words or linking elements that help establishing dependency relations between a word and its complement. From a semantic point of view, such dependencies are essential in expressing the relations between the different concepts.

According to Alicia María Zorrilla (2002), the Royal Academy of Spanish Language defines prepositions as "[...] inflectional words linking any syntactic element to a nominal complement [...] They are proclitic particles occurring before nominal compliments subordinating them to a certain word".

Though considered by some linguists as "empty" words, their role is crucial in the Language in a way that some consider it impossible to reach a deep knowledge of a language without a complete understanding of its prepositional system (García Yebra, 1988 in Zorrilla, 2002). This is justified by the fact that prepositions are in many cases ambiguous and their use alters the significance since they can denote a number of relations: spatial, temporal, mode, instrument, means, etc. Besides, the prepositional paradigm includes different kinds of prepositions: simple, complex and prepositional clauses. The use of the different kinds of prepositions changes from one language to another and from one register to another, within the same language.

Taking into consideration the above features, prepositions have been subject of numerous studies in different branches of Linguistics and in a variety of languages. Many of these studies based their research on samples extracted from corpora (monolingual, bilingual or multilingual) and, thus, taking advantage of the availability of corpora and concordance tools (Mindt and Weber, 1989) (Tanimura et al., 2004) (Granath and Wherrity, 2005).

As expected, most studies regarding prepositions focus on written corpora reflecting aspects of written register of the language. The state-of-the-art, however, reveals a lack of studies addressing the prepositions within the context of spoken corpora. Given these facts, an electronic resource which compiles a number of prepositions together with their use not only in written registers, but also in spoken registers of two languages (French and Spanish) fills an important gap in the present research panorama.

Opposite to the trends paying particular attention to prepositions, in many cases, statistical quantitative approaches to Language Processing have discarded the

prepositions. This is explained in terms of Zipf's Law where prepositions lose relevance due to their high frequency. Consequently, many statistical approaches include prepositions in a list of stop-words to avoid noise during the processing stages. Nevertheless, recently prepositions have been gaining a growing attention from the NLP community, especially in areas concerning Natural Language Understanding as we will discuss in the following section dedicated to Prepositions and Lexicons.

### 1.3. Prepositions and Lexicons

To complete the outline of our theoretical framework, this section tackles the relation between prepositions and lexicons from the linguistic and the computational perspectives.

Linguistic studies concerning prepositions dates back to the Greek philosophers in their first attempts to establish a typology of grammatical categories. Since then, investigating the prepositional systems in the different languages has continued to analyze the prepositions from different dimensions. As a result of these attempts, literature review reveals a number of monolingual dictionaries and studies dedicated to prepositions in Spanish (Zorrilla, 2002) (Fernández López, 1999) (García Yebra, 1988) and French (Vandeloise, 1986) (Borillo, 1992) (Cadiot, 1997).

Despite the fact that statistical approaches ignore prepositions, recent computational approaches highlight the role of prepositions in understanding the syntax and the underlying semantics of Natural Language. During the last years, many projects and many studies have targeted the prepositions as essential constituents of other units such as propositions. PropBank and FrameNet are examples of these projects (Kingsbury and Palmer, 2003) where the use of prepositions is governed by verbs or nouns.

In addition to propositions and framenets, prepositions are basic elements in multi-word expressions. The term "multi-word expression" in fact does not indicate a grammatical category. It is a term that arose within the Natural Language Processing field to denote elements formed up from more than one word. These elements include compound prepositions, phrasal verbs, idioms, collocations, semi-fixed or fixed phrases. In all these elements, prepositions play a crucial role, and, thus, they highly affect the interpretation of these units. The need for lexicons of multi-word expressions increases due to their importance in a variety of applications such as Information Retrieval, Machine Translation, Question Answering, etc. (Calzolari et al., 2002).

Given the theoretical framework, our proposed lexicon of prepositions using spontaneous spoken corpus constitutes a contribution to the state-of-the-art considering both the linguistic and the computational points of views. We adopt a comprehensive approach that brings together real evidence from written and spoken corpora, offering, in this way, a broader scope covering the pragmatic and phonetic aspects of the use of prepositions.

Once we settled the theoretical basis of the study, in the following section we will present a detailed description of the lexicon structure focusing on the technical phases of development together with the challenges of integrating examples from the spoken corpus.

## 2. The Electronic Lexicon

The electronic version of the lexicon of prepositions is based, mainly, on the textbook titled "El libro de las preposiciones" (Sanz Miguel, 1999). The main goal of the book is to provide a thorough and illustrated resource for the difficulties of use concerning the prepositions in French and Spanish. It consists of sixteen chapters distributed among two main parts. The first part includes the first five chapters concerned with the translation of five basic prepositions from French into Spanish. The five prepositions are: *à, chez, de, en, voici-voilà*. The second part deals with the translation of Spanish prepositions into French and it covers the remaining eleven chapters where each is dedicated to a preposition. The set of Spanish prepositions includes the following: *a, ante, con, de, desde, en, entre, hasta, para, por* and *tras* (to/for, at, with, of, from, in, between, till, for, through, after). The book includes approximately 2,500 entries. These entries represent cases where literal translation from French into Spanish and vice versa is not applicable and the use is different between both languages such as in the case of "false friends". Entries represent different types of prepositions varying from simple prepositions, compound prepositions or prepositional phrases. In addition, it includes uses of prepositions governed by verbs, adjectives and nouns such as *rêver de, oblige de, proche de,* etc. Examples from written resources are provided in the textbook to illustrate the use of the prepositions. Most of these examples are selected from literary texts. However, other resources include, press, songs, advertisements, booklets, user manuals, etc.

Two electronic versions were developed: a version in HTML for browsing the dictionary on-line and a version in XML. The HTML version follows the structure of the original textbook, except for minor changes regarding the appearance and the formal presentation. However, for the XML version, a set of tags and features were added to enhance the organization of the lexicon and to allow an effective search and retrieval of the information. Moreover, the set of added features helps expanding the range of the applications that could benefit directly or indirectly from the lexicon. Integrating examples from the spontaneous spoken corpus C-ORAL-ROM is considered one of the major contributions among the set of added features.

C-ORAL-ROM stands for the *Corpus Oral de lenguas Romances,* Integrated Reference Corpora for Spoken Romance Languages (Cresti and Moneglia, 2005). It is the main deliverable of the C-ORAL-ROM project funded by the Fifth Framework Programme of the European Union within the Information Society Technology Programme. The corpus is made up of four sub-corpora in four romance languages: Spanish, French, Italian and Portuguese. The four sub-corpora are comparable recordings reflecting different registers of the spoken language. A number of criteria were adopted to classify the different recordings. The following are examples of the different typologies within each sub-corpus:

- Number of participants: dialogues, monologues and conversations
- Formal versus informal, where formal includes a sub-typology of news broadcast, political, preaching, interviews, etc. Informal includes familiar conversations, public speeches, etc.
- Channel: transmission (telephone/other media) or face-to-face

C-ORAL-ROM consists of 772 spoken texts and 121:43:07 hours of speech by 1,427 different speakers. Each sub-corpus consists of approximately 300,000 words. Each recording includes the following data (Moneglia, 2005: 2):

- Session metadata providing basic information of speakers, recording situation, acoustic quality, source and content of each session;
- Orthographic transcription, in standard format, enriched by the tagging of terminal and non-terminal prosodic breaks;
- Text-to-speech synchronisation, based on the alignment with the acoustic source of each transcribed utterance;
- Part of Speech tag of each form in the transcribed texts and the corresponding frequency list of forma and lemmas.

The Spanish sub-corpus was developed by the group of the Computational Linguistics Laboratory *Laboratorio de Lingüística Informática-Universidad Autónoma de Madrid* (LLI-UAM) (Moreno-Sandoval et al., 2005).The French sub-corpus was developed by the group of DELIC-Université de Provence, France (Campione et al., 2005). For the present work, we used the French and the Spanish corpora, together with the tools provided for browsing the corpora, such as the concordance of the French Corpus 'Contextes', 'WinPitch Corpus' for handling the audio files and the interface of the Spanish corpus available at the LLI-UAM group.

Counting on these resources, we undertook the task of converting the preposition textbook into an electronic lexicon enriching it with examples from the Spanish and the French corpora. The description of the technical development of the lexicon could be divided into the following modules:

- Basic processing for the conversion from document to HTML and XML format
- Developing the hierarchal XML tagset for the automatic tagging of the lexicon
- Integrating examples from the C-ORAL-ROM
- Developing a web user interface to access the lexicon

**Basic processing**

Based on the Word document files of the textbook, we developed a number of simple scripts in PERL to extract the textual content of the documents. In this stage the input consists of Word document file and the output is an HTML file. For each chapter file we generated a HTML version. At the end, all the HTML files were merged in one HTML file. Changes in the conversion only affected the representation of the content. For the web version, hyperlinks, tables, fonts and colours were added to facilitate browsing the book in a more interactive way. The following is a screen shot of one of the chapters after being converted into HTML.

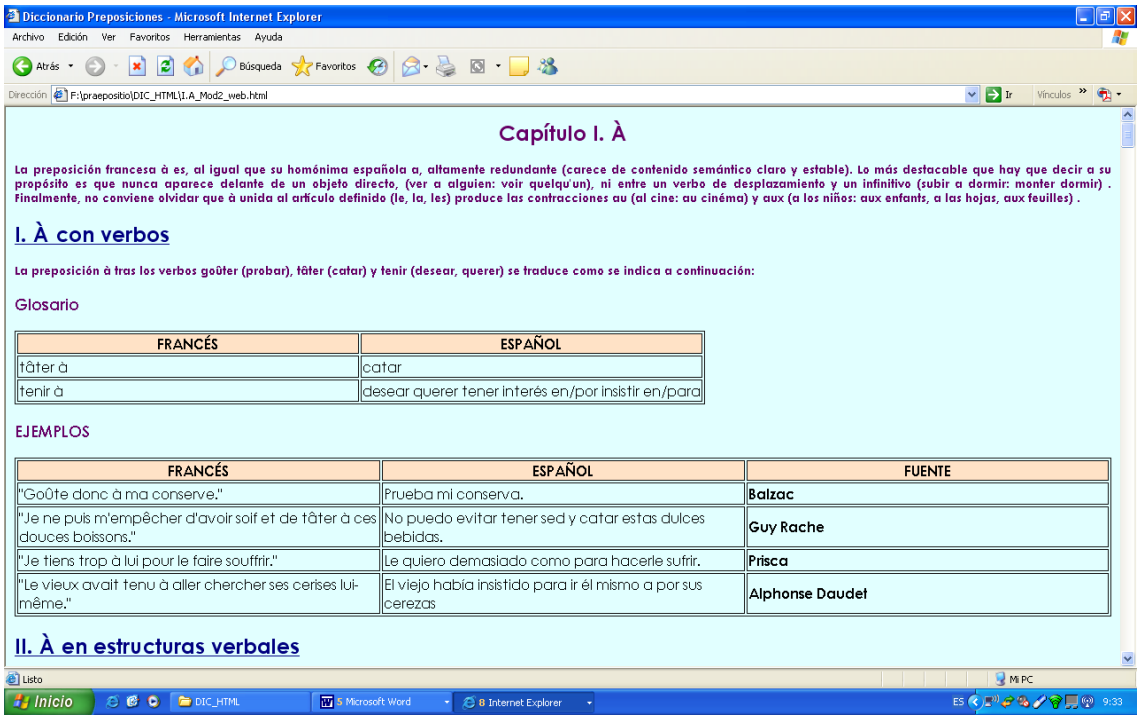


**Figure 1**: Snapshot of the HTML version

## XML Tagset and the Automatic Tagging of the Lexicon

After generating the HTML version of the lexicon, we proceeded on with generating the XML version. For generating this version, and as in the previous stage, we developed a number of scripts in PERL making use of the potential of the regular expressions.

The XML tagset is defined through the DTD file (*Data Type Definition*). The following are the main elements of the XML tagset reflecting the hierarchical structure of the lexicon:

- <dic> is the root element
- The tag <lex> represents each lexical entry where for each <lex>, there are tags indicating the language <es> and <fr>. Each lexical entry is identified by a unique number to facilitate the search and the retrieval of the information. Separating the level of the lexical entry from the language specification guarantees more flexibility allowing the introduction of other languages in the future without altering the main hierarchy.
- Each group of related lexical entries are grouped under a tag named <glos>.
- Each lexical entry may contain a tag called <ej> indicating the examples.
- At the example level, there are three main sub-elements indicating the language of the example, its corresponding translation and the source of the example.
- An attribute indicating the type of the example accepts the values of “oral” or “written” to indicate if the example is from the spoken or from the written register.
- For each example a tag indicating the source of the example is included. In case of examples from written language already present in the original version of the textbook, this element is filled with the name of the author of the

sentence used as example. In case of examples from the spoken corpus, the source was put to C-ORAL-ROM since this is the corpus we are using in this version of the lexicon. However, in case other resources are used, they could be indicated without any change in the structure of the lexicon.

- For the oral examples, the value of the attribute called file indicates the location and the id of the file from which the example was selected in the Corpus. The location is essential to link the example with the audio file. Also for restricted use, a link to the transcription file is to be included.
- Further features can be added at the level of the oral example, such as the type of recording (formal, informal, dialogue, monologue, etc.)
- A tag named <desc> can appear at the different levels if needed. Information regarding descriptions or explanations is included in this tag.
- For the semantic uses of the prepositions, an attribute indicating the "case" is included if applicable since this information is not available for all entries in the present version.
- Moreover, a tag indicating the grammatical information (optional) can be included under each lexical entry. Information concerning the Part-Of-Speech in case of compound prepositions or multi-word expressions. However, this information is not available for all entries in the present version. Further completion of the grammatical tags is considered for future work.

**Integrating Examples from the C-ORAL-ROM**

The phase of integrating examples from the C-ORAL-ROM was carried out by a bilingual speaker of French and Spanish. This is crucial since the search and the selection of examples from the Spoken language register require a level of proficiency in both languages in order to distinguish the different uses and to ensure that the examples selected reflects the use given in the lexicon.

For the French corpus, the Contextes tool, provided by DELIC group was used to look up the corpus for examples, while for the Spanish corpus; we used the interface available at the LLI-UAM.

The integration of the examples involves a number of challenges in both: searching the examples and selecting which examples to include. Thus, we will first discuss the difficulties in the finding the examples, and the criteria we adopted for the selection process.

Finding the examples was not an easy task. The preposition or the expression in concern is introduced in the concordance tool which looks up the corpus for occurrences of the query string. A number of difficulties were encountered at this stage such as the high recall number of simple searches or the absence of examples in the corpus for other searches. Once the examples are retrieved from the corpus, we encountered other types of challenges concerning the selection of the examples. Decisions have to be taken in order to select which examples are to be introduced to the lexicon and which are to be discarded. This is applicable in the following situations:

1- If the corpus contained many examples of the expression in concern
2- If the corpus lacks of any examples of the expression in concern

The number of results obtained by introducing a simple preposition was really high due to the high frequency of prepositions. For example, figure 2 shows the results

obtained (207 occurrences in the French corpus) when the preposition “chez” is searched.

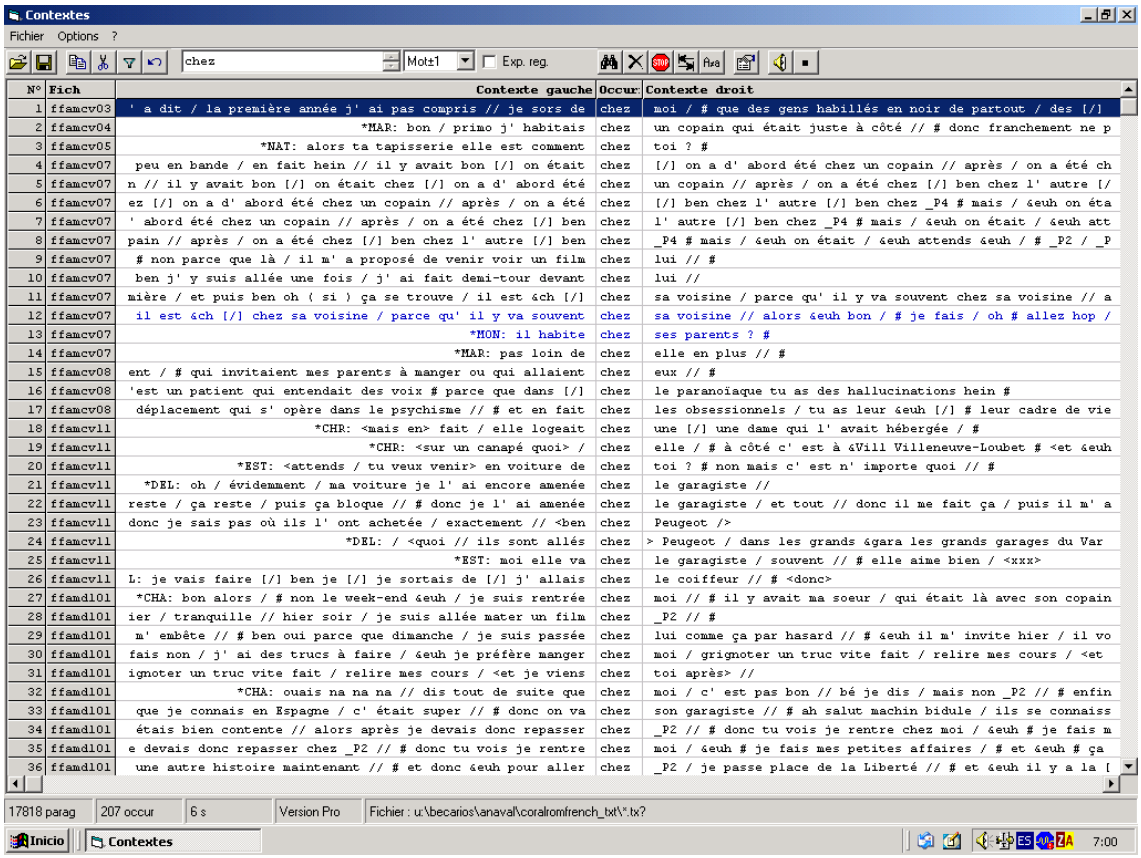


**Figure 2**: Results obtained for the preposition “chez” in the French corpus

Given the high number of recall, these examples have to be examined in order to select the example which fits with the use and the meaning given in the lexicon. This is possible by examining the transcribed text available through the tool. The transcribed text represents the whole recording together with the metadata such as the kind of speech and the participants. Accessing this information gives a better view of the occurrence and helps deciding which occurrence fits best in our lexicon.

The criteria adopted in case different numbers of examples were found consider aspects related to the clarity and simplicity, on one hand, and adequacy to the use intended in the original lexicon. Based on these features, we opted for clear, simple examples offering a better understanding of the use of the preposition or the expression. Besides and given that the basic textbook was mainly designed for the written register of the language, variations of use between the written and the spoken registers are quite common. In this way, many of the examples found in the spoken corpus represent different uses than those pretended by the book. Also, uses in the spoken register are subject to changes more frequently than uses in written registers. Among these possibilities offered by the spoken corpus, our selection criteria, in this version, tried to reflect as far as possible the main significant and use given by the primary author of the textbook. However, for posterior versions, we intend to extend the examples offering, in this way, a wider scope for the user and a better understanding of the use of the prepositions in the different situations.

Another type of difficulty is the absence of examples in the corpus which could be explained taking into account the following reasons:

- it is possible that the corpus does not contain instances of the expression;
- some expressions, especially prepositional phrases preceded by verbs, might occur in an inflected form and, thus, searching the infinitive forms would not lead to any results.

Regarding the first reason, it is already expected that corpora, in general, are limited and do not reflect all uses of the language. Besides, in dealing with spontaneous spoken corpora, we have to be aware of the fact that the spoken register of a language is more limited than the written register in terms of lexical resources. This is explained if we take into consideration that speakers adopt different strategies in the oral expression. On one hand, they tend to save effort by maximizing the use of a limited set of lexical and grammatical resources, and, on the other hand, they apply other strategies typical of the oral register such as the intonation and prosody.

The second reason can be explained considering the linguistic phenomenon of lemmas vs. types. In a dictionary or a lexicon, the entries are organized in terms of lemmas, i.e. the canonical form of the word. However, occurrences in corpora are types that do not usually adopt the canonical form. This feature has to be taken into consideration during the look-up phase. Thus, the terms of search should be adapted by introducing inflected forms to widen the area of search. For example, constructions of VERB+PREPOSITION, such as "*rêver de*", if introduced in the infinitive form, the concordance would not retrieve any results. To retrieve results, we adapted the search query by introducing the verb in the present form or in the past form. In the case of ADJECTIVE+PREPOSITION, modifying the query implied changes in number and gender. For example, "*obligé/e de*" (obliged to) was looked up by introducing the different combination of feminine/masculine and singular/plural to retrieve all occurrences of the expression.

After the selection of the examples (for the target language) is carried out, we faced another challenge at introducing the examples in the lexicon. The examples encountered in the corpus are in a transcribed format which makes use of a number of standard signs to reflect certain features of the spoken language such as the overlapping, interruptions, comments, pauses, etc. Introducing the transcribed text directly into the lexicon might result illegible for the standard user. This is the reason that we decided to modify these signs in the examples directly introduced into the lexicon opting for simple conventions that could be easily understood by the user. For other types of user, the information concerning the transcription is made available through a link to the transcription file.

Finally and to enhance the features and usability of the lexicon, a tag indicating a translation of the selected example (of the target language) was included. We opted for the translation since the spoken corpus is comparable and not a parallel one. The translation was carried out by competent bilingual speakers capable of reflecting the corresponding equivalence according to the context implied in the example. In case the source language is Spanish and the target language is French, in addition to the Spanish translation of the example in target-language (French), we decided to add one more example from the corpus in the source language (Spanish) which reflects an equivalent use in the source language.

## The web user interface

To access the lexicon, we developed a simple web user interface. Through this interface, the user can search for entries in the lexicon on-line. The user introduces his search query in a simple form.

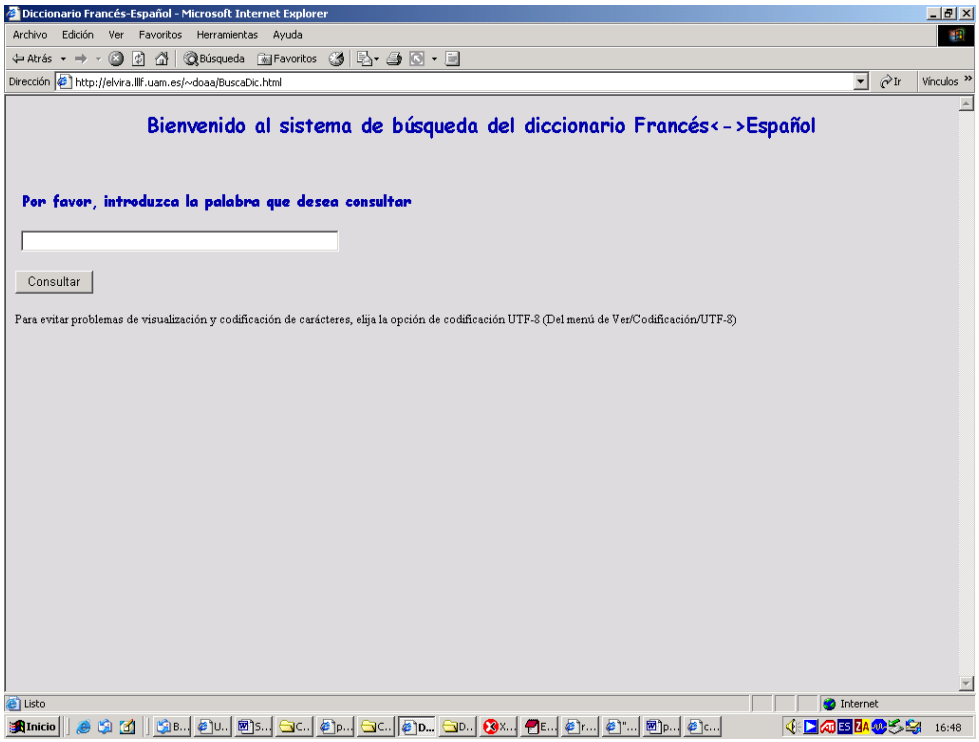


**Figure 3**: The web interface

Through a dynamic CGI-script, the query is sent to the server and the script searches the lexical entries marked with the <lex> tag. All occurrences of the query are retrieved, together with all the associated information concerning the grammatical, semantic and pragmatic features. Besides, links to the audio file and the transcription texts are also retrieved. In this way, for each lexical entry, the web user interface represents the data in the form of tables where each field represents the query in the source and the target language. Under each lexical entry, the associated information is represented. This information includes:

- examples indicating the following features:
    - written (example in target language together with its translation)
    - oral (example in target language together with its translation)
        - Links to the audio files
        - Links to the transcription files
    - the source

- grammatical features for both source and target language (though in the present, this feature is not yet included in all lexical entries) indicating the Part-of-Speech in case of compound entries and complex constructions; and finally,
- brief semantic information related to the uses of the expression, i.e. if it is used to express mode, instrument, cause, movement, etc.

In the appendix, we include the retrieved set of prepositions and expressions used with the word *mano* (hand).

## 3. Conclusion and Future Work

In this work, we have presented a primary version of a bilingual lexicon of prepositions enriched with examples from a spontaneous spoken corpus. Based on a theoretical framework, we described the phases and the challenges encountered in the development of this resource. The novelty of the approach lies in merging examples from written and spoken sources, offering, in this way, real evidences of language in use across the different registers.

A lexical resource, exhibiting these features, is highly valuable for Applied Linguistics, especially for studies dealing with Second Language Acquisition/Teaching. At the same time, this resource could be integrated in modules for Language Engineering and Natural Language Processing systems. Taking into account these applications, a number or studies can benefit from it. For future studies, we plan to undertake two main directions: Second Language Acquisition and NLP applications. For the former, we plan to conduct a study to evaluate the utility of the resource based on using it in class by the learner or by the teacher. For the later, we intend to integrate the Spanish entries of lexicon, especially the multi-word expressions, in the general lexicon used in the POS tagger and the morphological analyzer. Besides, we are developing a tool to convert the set of examples and its corresponding translations into the translation memory standard TMX format (Translation Memory Exchange). This conversion allows direct integration of the examples into translation memory systems.

On the other hand and regarding the lexicon itself, we consider quantitative and qualitative enhancements for future versions. Quantitative enhancements imply expanding the coverage of the lexicon by including more entries and more examples, while for qualitative aspects; we esteem appropriate including further features at the grammatical, semantic and pragmatic levels. At the grammatical level, we plan to include Part-of-Speech tags for all the lexical entries and all the examples in both languages. Regarding the semantic features, more detailed description of semantic cases is considered. Finally, for the pragmatic features, we plan to include further information regarding the spoken examples such as the participants, their age, the place where the speech is taken place, etc.

## Acknowledgements

This research has been supported by the Praepositio project, Universidad de Castilla La Mancha and by the grant TIN2004-07588-C03-02 (Spanish Ministry of Education and Science).

**References**


Borillo, A. (1992) : "Le lexique de l'espace : prépositions et locutions prépositionnelles de lieu en français". In L. Tasmowski & A. Zribi-Hertz (eds.) *Hommage à Nicolas Ruwet*, Ghent : Communication et Cognition.

Cadiot, P. (1997): *Les prépositions abstraites en français*. (Paris: Colin)

Campione, E.; Véronis, J. and Deulofeu, J. (2005) : The French Corpus. In Cresti, E. y M. Moneglia (Eds.), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, (Amsterdam: John Benjamins Publishing Company) pp. 111–34.

Calzolari,N.; Fillmore, C.; Grishman, R.; Ide, N.; Lenci, A.; MacLeod, C. and Zampolli, A. (2002): Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, pp. 1934–40.

Cresti, E. y M. Moneglia (Eds.) (2005): *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages* (Amsterdam: John Benjamins Publishing Company).

Fernández López, M. (1999): *Las preposiciones: valores y usos. Construcciones preposicionales*. (Salamanca: Ediciones Colegio de España).

García Yebra, V. (1988): *Claudicación en el uso de las preposiciones*. (Madrid: Gredos)

Granath, S. and Wherrity, M.P. (2005): Prepositions with *that*-clause complements in tagged corpora, with a special focus on *in that*. In *Proceedings of Corpus Linguistics 2005*. Available on-line from: http://www.corpus.bham.ac.uk/PCLC/GranathWherrity.doc (Accessed June 2007)

Kingsbury, P. and Palmer, M.(2003): PropBank: the Next Level of TreeBank. In *Proceedings of Treebanks and Lexical Theories*. Available on-line from: http://w3.msi.vxu.se/~rics/TLT2003/doc/kingsbury_palmer.pdf (Accessed June 2007)

Mindt, D. and Weber, C. (1989): Prepositions in American and British English, in World Englishes vol. 8, no. 2, pp. 229–38.

Moneglia, M. (2005): The C-ORAL-ROM resource. In Cresti, E. y M. Moneglia (Eds.), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, (Amsterdam: John Benjamins Publishing Company) pp. 1–70.

Moreno-Sandoval, A., G. De la Madrid, M. Alcántara, A. González, J.M. Guirao y R. De la Torre (2005): The Spanish Corpus. In Cresti, E. y M. Moneglia (Eds.), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, (Amsterdam: John Benjamins Publishing Company) pp. 135–61.

Sanz Miguel, C. (1999): *El libro de las preposiciones*. (Toledo: Azacanes)

Tanimura, M.; Takeuchi, K. and Isahara, H. (2004): From Learners' Corpora to Expert Knowledge Description: Analyzing Prepositions in the NICT JLE (Japanese Learner English) Corpus. In *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 139–47. Available on-line from: http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/1405/1/16.pdf (Accessed June 2007)

Vandeloise, C. (1986). *L'espace en français : sémantique des prépositions spatiales* (Paris: Seuil)

Zorrilla, A.M. (2002): *Diccionario de las preposiciones españolas: Norma y uso*. (Capital Federal: e.d.b).

**Appendix**: Example of the results retrieved

**Ha buscado la(s) palabra(s): mano**

**Total resultados=17 en glosario y 7 en ejemplos**

**Resultado 1.**

| ES: tener a mano | FR: avoir sous la main <EJ< font> |
|---|---|

**Ejemplo de lenguaje escrito**

"Nous n'avons pas de lion sous la main." No tenemos ningún león a mano.

Fuente: Jacques Prévert

**Ejemplo de lenguaje oral**

Tu as pas son numéro sous la main ? Je vais l'appeler. ¿No tienes su número a mano? Voy a llamarla.

Fuente: CORALROM

**Resultado 2.**

| ES: con la mano | FR: de la main |
|---|---|
| | FR: avec la main <EJ< font> |

**Ejemplo de lenguaje escrito**

"Faisant des gestes de la main, Katherine a l'air de bénir la foule." Haciendo gestos con la mano, Katherine parece bendecir a la multitud.

Fuente: Jacques-Marie Bourget

**Resultado 3.**

| ES: con las manos | FR: avec les mains |
|---|---|
| | FR: dans ses mains </L< font> |

**Resultado 4.**

| ES: con las dos manos | FR: des deux mains |
|---|---|
| | FR: à deux mains <EJ< font> |

**Ejemplo de lenguaje escrito**

"Il soulèvera le plafond des deux mains." Levantará el techo con las dos manos.

Fuente: Nicole Ciravegna

**Ejemplo de lenguaje escrito**

"Il a pris sa tête à deux mains." Cogió su cabeza con las dos manos.

Fuente: Raymond Queneau

**Ejemplo de lenguaje oral**

Tu as un geste alternatif des deux mains, l'une après l'autre.

Tienes un gesto alternativo con las dos manos, una después de la otra.

Fuente: CORALROM

**Resultado 5.**

| | |
|---|---|
| **ES: con las manos desnudas** | **FR: les mains nues** |
| | **FR: à mains nues <EJ< font>** |

**Ejemplo de lenguaje escrito**

"A mains nues, je déblayai un peu de terre."

Con las manos desnudas, yo despejé un poco de tierra.

Fuente: Christian Jacq

**Resultado 6.**

| | |
|---|---|
| **ES: con las manos en los bolsillos** | **FR: les mains dans les poches <EJ< font>** |

**Ejemplo de lenguaje escrito**

"Il y a des invités qui se présentent les mains dans les poches."

Hay invitados que se presentan con las manos en los bolsillos.

Fuente: Mariella Righini

**Resultado 7.**

| | |
|---|---|
| **ES: con una sola mano** | **FR: d'une seule main** |
| | **FR: avec une seule main </L< font>** |

**Resultado 8.**

| | |
|---|---|
| **ES: saludar con la mano** | **FR: saluer de la main </L< font>** |

**Resultado 9.**

| | |
|---|---|
| **ES: coger de la mano** | **FR: prendre par la main <EJ< font>** |

**Ejemplo de lenguaje escrito**

"Le Dieu d'hier le prenait par la main."

El Dios de ayer le llevaba de la mano.

Fuente: Roger Bodart

**Resultado 10.**

| | |
|---|---|
| **ES: Traer de la mano** | **FR: tenir par la main <EJ< font>** |

**Ejemplo de lenguaje escrito**

"Je tenais par la main ma fille." Yo llevaba de la mano a mi hija.

Fuente: Victor Hugo

**Resultado 11.**

| ES: poner la mano en | FR: mettre la main sur |
|---|---|
| | FR: poser la main sur <EJ< font> |

**Ejemplo de lenguaje escrito**

"Gabriel attendait, les mains posées sur ses genoux." — Gabriel esperaba, las manos apoyadas en sus rodillas.

Fuente: Raymond Queneau

**Resultado 12.**

| ES: en (la) mano | FR: à la main |
|---|---|
| | FR: dans la main <EJ< font> |

**Ejemplo de lenguaje escrito**

"Elle tenait à la main un bouquet de roses blanches." — Ella tenía en la mano un ramo de rosas blancas.

Fuente: Henri Troyat

**Ejemplo de lenguaje escrito**

"J'ai senti dans mes mains un animal immonde." — Sentí en mis manos un animal inmundo.

Fuente: Georges Bataille

**Ejemplo de lenguaje oral**

Vous niez avoir jamais possédé d'armes, or, on vous a vu à plusieurs reprises avec un poignard à la main. — Usted niega haber poseído jamás un arma, ahora bien, se le ha visto en varias ocasiones con un puñal en la mano.

Fuente: CORALROM

**Resultado 13.**

| ES: caer en manos de | FR: tomber aux mains de <EJ< font> |
|---|---|

**Ejemplo de lenguaje escrito**

"Un lac écarlate est tombé aux mains du vent." — Un lago escarlata cayó en manos del viento.

Fuente: Maurice Blanchard

**Resultado 14.**

| ES: tener en mano | FR: tenir en main <EJ< font> |
|---|---|

**Ejemplo de lenguaje escrito**

"J'ai un autre de mes personnages qui tient un parapluie en main." — Tengo a otro de mis personajes que tiene un paraguas en la mano.

Fuente: Michel Folon

**Resultado 15.**

| | |
|---|---|
| **ES: tener en la mano** | **FR: avoir dans la main** |
| | **FR: tenir dans la main <EJ< font>** |

**Ejemplo de lenguaje oral**

Rétina 2 c'est un appareil allemand qui se pliait avec un petit soufflet, mais qui était pas très grand, on l'avait bien dans la main.

Rétina 2 es un aparato alemán que se doblaba con un fuellecito, pero que no era muy grande, se podía tener bien en la mano.

Fuente: CORALROM

**Resultado 16.**

| | |
|---|---|
| **ES: tener en mis (tus, sus, ...) manos** | **FR: avoir en mes (tes, ses, ....) mains** |
| | **FR: tenir en mes (tes, ses, ...) mains <EJ< font>** |

**Ejemplo de lenguaje escrito**

"Il tient les clés des portes en ses mains pour trahir la ville."

Tiene las llaves de las puertas en sus manos para traicionar a la ciudad.

Fuente: Jean Maillart

**Resultado 17.**

| | |
|---|---|
| **ES: ganar por la mano** | **FR: prendre de vitesse <EJ< font>** |

**Ejemplo de lenguaje escrito**

"L'avancée des troupes les a pris de vitesse."

La avanzadilla de las tropas les ganó por la mano.

Fuente: Olivier weber

####################

**Se han encontrado los siguientes resultados en el/los siguiente(s) ejemplo(s)**

**1. Ejemplo de lenguaje escrito**

**"Ces mains n'ont pas vendu d'oranges."**

**Esas manos no han vendido naranjas.**

**Fuente: Arthur Rimbaud**

**2. Ejemplo de lenguaje escrito**

**"En voulant arracher l'assiette des mains de son frère, le roi Louis XVI jette quelques gouttes de bouillie sur la tête de Monsieur."**

**Por querer arrancar el plato de las manos de su hermano, el rey Luis XVI tira algunas gotas de papilla a la cabeza de Monsieur.**

**Fuente: Léo Moulin**

**3. Ejemplo de lenguaje escrito**

| | |
|---|---|
| "En voulant arracher l'assiette des mains de son frère, le roi Louis XVI jette quelques gouttes de bouillie sur la tête de Monsieur." | Por querer arrancar el plato de las manos de su hermano, el rey Luis XVI tira algunas gotas de papilla a la cabeza de Monsieur. |

**Fuente: Léo Moulin**

**4. Ejemplo de lenguaje escrito**

| | |
|---|---|
| "Voici ce qui est. Prenez l'affaire en main." | Esto es lo que hay. Tome usted el asunto entre sus manos. |

**Fuente: Emile Zola**

**5. Ejemplo de lenguaje escrito**

| | |
|---|---|
| "Je l'ai prise en main." | La he cogido en mano. |

**Fuente: Evelyne Madelénat**

**6. Ejemplo de lenguaje escrito**

| | |
|---|---|
| "Il tient sa tête entre ses mains." | Coge su cabeza entre las manos. |

**Fuente: Florence Saugues**

**7. Ejemplo de lenguaje oral**

| | |
|---|---|
| Il est déporté au titre du S.T.O, la main inéluctable du destin s'acharne encore sur lui, rapatrié le vingt mai dix-neuf cent quarante-cinq, il séjourne à Paris jusqu'en dix-neuf cent quarante-sept. | Se le deportó a título de S.T.O., la mano inevitable del destino se ceba de nuevo sobre él, repatriado el veinte de mayo de mil novecientos cuarenta y cinco, permanece en París hasta mil novecientos cuarenta y siete. |

**Fuente: CORALROM**