Integrating Corpus-based Resources and Natural Language Processing Tools into CALL

PASCUAL CANTOS-GOMEZ
Universidad de Murcia

ABSTRACT

This paper aims at presenting a survey of computational linguistic tools presently available but whose potential has been neither fully considered nor exploited to its full in modern CALL.

It starts with a discussion on the rationale of DDL to language learning, presenting typical DDL-activities, DDL-software and potential extensions of non-typical DDL-software (electronic dictionaries and electronic dictionary facilities) to DDL.

An extended section is devoted to describe NLP-technology and how it can be integrated into CALL, within already existing software or as stand-alone resources. A range of NLP-tools is presented (MT programs, taggers, lemmatizers, parsers and speech technologies) with special emphasis on tagged concordancing.

The paper finishes with a number of reflections and ideas on how language technologies can be used efficiently within the language learning context and how extensive exploration and integration of these technologies might change and extend both modern CALL and the present language learning paradigms.

KEYWORDS: Concordancing, corpus linguistics, data-driven learning, electronic dictionaries, human language technologies, linguistic corpora, machine translation, morphological generation, NLP-tools, parsing, POS-tagging, speech technology

* Address for correspondence: Pascual Cantos Cómez, Departamento de Filología Inglesa, Universidad de Murcia. C/ Santo Cristo 1, 30071 Murcia, Spain, e-mail: pcantos@um.es.

© Servicio de Publicaciones, Universidad de Murcia. All rights reserved.
1. INTRODUCTION

The use of concordancing in literature and linguistic analysis is nothing new. It started well before computers existed. Tribble and Jones (1990) trace the history of concordancing from the 13th century, when Hugo de San Charo enlisted 500 monks in producing a complete concordance of the Latin Bible. That is, a kind of reference work designed to assist in the exegesis of the Bible, consisting of all occurrences of terms, names, etc., that were felt to be significant, and presented these terms in a way that would help the researcher.

Of course, current applications of concordances in language and literature are not so labour intensive. The use of concordancing as a tool for language learning/teaching is relatively recent, starting in the 1980’s, when computational power began to get scaled into small, affordable personal computers.

Succinctly, a concordance is a data arrangement technique that transforms texts into lists, printing lines of text where the word or expression interested in investigating is displayed in the centre of line, know as KWIC (Key Word in Context), within an arbitrarily selected context of characters or words to its right and left. This technology permits that, for example, a language teacher or learner interested in knowing the use of the preposition of to transform a text such as:

What be more important or intriguing than our own origins? Like all animals we come from one cell that develops into an embryo which forms the adult. This embryonic development presents a fundamental problem of biological organization. From the single cell, the fertilized egg, come large numbers of cells—many millions in humans—that consistently give rise to the structures of the body. How do these multitudes of cells become organized into the structures of, for example, our body—nose, eyes, limbs, and brain? What controls their individual behavior so that a global pattern emerges? And how are the organizing principles, as it were, embedded or encoded within the egg? It is remarkable that a cell as overtly dull and structureless as the fertilized egg can give rise to such varied and complex forms. The answer lies in cell behavior and how this behavior is controlled by genes.

Into the following format:

... presents a fundamental problem [of] biological organization.  
... egg, come large numbers [of] cells—many millions in ...  
... rise to the structures [of] the body. How do ...  
... How do these multitudes [of] cells become organized into ...  
... organized into the structures [of], for example, our body ...  

Clearly, what this technique does is to make the invisible visible for teachers and learners. Patterns and structures that would else hardly be immediately recognisable, spring to the eye.
II. A WORD ON DATA-DRIVEN LEARNING


Johns (1988) states that the use of concordancing in language learning: (a) interjects authenticity (of text, purpose, and activity) into the learning process; (b) learners assume control of that process; and (c) the predominant metaphor for learning becomes the research metaphor, as embodied in the concept of data-driven learning (DDL), which builds learners’ competence by giving them access to the facts of linguistic performance.

Higgins (1985) proposes concordences as the central idea to shift the pedagogical teaching/learning paradigm from computer as magister to computer as pedagogue. That is, from a mere process-control model of language instruction to an information-resource model, where learners explore, hypothesize and learn the language for themselves and the role of instruction is to provide tools (concordance programs) and resources (texts or corpora) for doing so. Similarly, Cobb (1997) considers that DDL has a specific learning effect that can be attributed to the use of concordance software by language learners. He concludes that computer concordances might simulate and potentially rationalize off-line vocabulary acquisition by presenting new words in several contexts.

Stevens (1995) accounts that many teachers feel that concordancers are the type of software that most closely approaches fulfilling the potential of computers in language learning. In a sense, they are working approximations of expert systems. They bring cognitive and analytic skills in students to bear on the manipulation of comprehensive databases for the purpose of solving real-language problems.

The effectiveness of concordances becomes also apparent not just in teaching/learning, but also in linguistic research. By means of this technique, Ketteman (1995) compares the treatment of the English conditional clause in a standard grammar used in Austria (Kacovsky 1987) with the evidence of authentic usage (corpora), and his comparison showed that an important type of English conditional with present tense in both main clause and conditional clause, accounting for one third of all instances in the data, is ignored in Kacovsky’s grammar.

Tribble (1997) stresses the potential and usefulness of DDL to language learning/teaching even with few corpus resources or small, specific corpora. Corpora and corpus-based exercises are useful because they favour learning by discovery—grammar, vocabulary, etc. (Tribble and Johns 1990:12).
A further related issue with DDL is authenticity. Widdowson (1983:30) considers that an authentic stimulus in the form of attested instances of language does not guarantee an authentic response in the form of appropriate language activity [...]. We should retain the term 'authenticity' to refer to activity (i.e. process) and use the term 'genuine' to refer to attested instances of language.

In this sense a corpus may contain millions of "attested instances of language", but there is nothing to guarantee that you can use data from that corpus as a stimulus for "appropriate language activity" (Tribble 1997). That is, it is likely that foreign language students are not necessarily motivated by a language learning activity if the instances of language use that they are studying are extracted from contexts that have little or no connection with their interests and concerns. Genuine examples of language in use will not necessarily lead to authentic language use or effective language learning activities.

So the question is: which is the best corpus for language learners? Flowerdew (1993:309) thinks that

Mainly native speakers make use of others' writing or speech to model their own work in their native language where the genre is unfamiliar. It is likely that this skill was brought out of the closet and exploited as an aid for learning.

Similarly, Bazerman (1994:131) considers that the most useful corpus for learners of English is the one which offers a collection of expert performances in genres which have relevance to the needs and interests of the learners. These texts might exemplify the results and models of the desired forms of language behaviour that language learners want to achieve and might, therefore, be motivating starting points for language learning and language using activities.

Clearly, this, somehow, relegates standard, balanced and representative corpora, such as the Brown corpus of American English (Küchera and Francis 1967), the Lancaster-Oslo-Bergen (LOB) corpus of British texts (Johansson 1980), or other major corpora such as the British National Corpus (BNC) (Burnard 1995), a 100 million word representative corpus of contemporary British written and spoken texts, or the Bank of English at Birmingham University (Sinclair 1991), for language learning purposes. Tribble (1997) points towards non-standard corpora for DDL and draws his attention to multimedia encyclopaedia, such as Microsoft Encarta®, among others. The latest version of Microsoft Encarta® contains more than 30,000 articles, between 200 and 5000 words, which count for a total of roughly 30 million words, covering different domains and topics, such as art, geography, history, language, life science, literature, philosophy, physical science, religion, social science, sports, etc. The data provided by this multimedia encyclopaedia virtually contain enough texts which most students in most language classes will find interesting and informative. In addition, with this comprehensive range of topics and texts, it is not difficult to select ad hoc texts, focusing on students' particular needs and motivations.
Of course, our aim here is not to advertise any particular multimedia encyclopaedia but much more to encourage language students and teachers to use the vast range of language texts, corpora or data, in general, which is available in electronic form, in CD-ROMs and/or Internet, rather than urging them to construct their own comprehensive and representative corpora.

II. A WOKD ON EXISTING DDL-SOFTWARE

In this section, we shall review the main software applications used among DDL practitioners: (1) commercially available concordance programs and (2) Tim Johns' Context. In addition, we shall also present a Spanish vocabulary learning multimedia application, Practicatu Vocabulario (Sánchez and Cantos 2000), which is based on the electronic dictionary metaphor and DDL-like learning/acquisition strategies.

II.1. Conondancers

Concordancers are text processing tools for looking at how words behave in texts. These tools allow you to find out how words are used in texts. Among the facilities, all concordancers allow you at least:

- To list all the words or word-clusters in a text, set out in alphabetical or frequency order.
- To see any word or phrase in context (concordances), so that you can see what sort of company it keeps.

This text processing tool is generally used for lexicographic work, for preparing dictionaries, and by researchers investigating language patterns.

Tim Johns has compiled numerous exercise examples on his web page using standard concordancing tools. The classroom materials that follow are extracts from his website and are a collection some participants' work of the Usti nad Labem DDL Workshop (21st - 25th March 2000):
About and On

How do we use the prepositions ABOUT and ON? Which one is used more often? Which one tends to be used in academic texts? In which cases is there the occurrence of only one of them?

Book
1. Be Good, a British Medical Association book about a potential risk to human health in the Western market. It was published as coffee-table book about art and behaviour called The Ace, which urges smoking 16%. It was based on a book about the terrible fate awaiting humanity by
4. in life. This is yet another one-thrill book about forensic science, this time based on
5. title intrigued me: At last, I thought, a book about the personal relationships that action

(by Kvita Rychtarova)

'Great', 'Big', 'Large', 'Huge' and their Collocations

Task 1 - 'Great' and its collocations
Try to spot the typical cases of 'great' and its collocations in the basis of the following examples.

1. Whether recent discoveries in the Great Pyramid of Cheops have anything to do
   with restoring the Great Pyramid and its Great Mosque to their former throne
   king themselves comfortable. And the great London clade, with their toasting real fires

2. narratives of Los Mirambales and Great Expectations. What could be more comic
   she agreed to make a film in the Great Journey series, the theme of which was

Task 3 - 'Great' and missing nouns
Try to predict the words which are missing.

1. Fairest who has been turned into Windsor Great __________ with antlers on his head
2. Bonanza is one of Europe's great __________ , like Rodin's an all-round
3. who had become known as Great __________. The Times, the Daily Mail and
4. In 1670 Frederick II the Great __________ of Prussia
5. the world was dominated by the great European __________ and, since the 1850s,

(by Sarka Canova and Jarka Ivanova)
Adjectives ending in -ic, -ical

1. Look carefully at the following citations. What difference in the meaning of the adjectives classic/classical can you spot?

And when Sinatra was making his classic albums for Capitol in the 1950s, Songs For Swinging ... Gramophone, the classical music magazine, has not written about her single or her album.

II. How fill in the blanks with appropriate adjectives.

1. I remember listening to all the ________ Motown music and the Philadelphia soul stuff.
2. Once it became clear that she could not continue through the next two acts, Deane decided to replace both dancers, as in ________ ballet one partner may not be physically suited to perform with a stand-in.
3. ... we always dine at Café Des Arts. A ________ bistro, run by a consortium of charming ladies, stylish, innovative food.
4. The area where the dirt collects is transparent. All our detritus is paraded on the outside, turning the ________ designs inside out. Why do we need to see it?
5. ... architect Richard Norman Shaw, capable of turning out gthic, Queen Anne and strict designs, made a valuable contribution to the Arts and Crafts movement.

(by Zuzana Šaffková and Vladislav Smolka)

II.2. Ready-made DDL-software: Context

Probably one of the best known and most used DDL-software is Context. This program encourages language learner to investigate how words are used in context in English, and is designed to supplement classes. It is based on short contexts (extracted from the database by means of the computer program MicroConcord) illustrating the use of important key items from a database of over 3 million words of text in English.

The program starts offering the user a list of headings: Top Menu (parts-of-speech and topics; see Figure 1). In addition, it is also possible to view a more detailed index of all the keywords available to the program, together with the names of the files in which each key word is stored and to select a file of contexts (by keywords defined by parts-of-speech, keywords defined by topic or morphemes — prefixes or suffixes; Figure 2). Once the user has selected the file of context, the program displays the list of key items in the bottom of the screen (among other facilities) in order to investigate the set of contexts for any particular key item (Figure 3). The Quiz Screen challenges students to guess what the missing keyword is (Figures 4 and 5). After students have finished the Quiz, they can see an analysis of their performance.
### Figure 3. Concordance-data Window

<table>
<thead>
<tr>
<th>Verbs</th>
<th>Show</th>
</tr>
</thead>
<tbody>
<tr>
<td>adopt</td>
<td>confirm</td>
</tr>
<tr>
<td>add</td>
<td>convey</td>
</tr>
<tr>
<td>add</td>
<td>convey</td>
</tr>
<tr>
<td>appeal</td>
<td>count</td>
</tr>
<tr>
<td>beg</td>
<td>cross</td>
</tr>
<tr>
<td>behave</td>
<td>destroy</td>
</tr>
<tr>
<td>blow</td>
<td>differ</td>
</tr>
<tr>
<td>borrow</td>
<td>divide</td>
</tr>
<tr>
<td>challenge</td>
<td>earn</td>
</tr>
<tr>
<td>confer</td>
<td>escape</td>
</tr>
</tbody>
</table>

### Figure 4. Quiz Window (1)

**Help**

**Guess the answer**

**Another Context**

Press C for another context!

Succinctly, electronic dictionaries are commercial derived products of standard paper dictionaries. The main differences between paper and electronic dictionaries is that in the latter you can find words immediately, even if you are unsure of the exact spelling. They can also be run as a stand-alone program and can be used in conjunction with most word-processing software; you can browse through entries, view adjacent entries, or travel swiftly between entries. In addition, some electronic dictionaries keep track of your searches so that you can very easily return to words you have previously looked up. You can also print extracts or definitions and copy them to the clipboard. Electronic dictionaries can be monolingual, bilingual or multilingual (Figures 6 and 7).
Figure 6. Example of a monolingual electronic dictionary

Figure 7. Example of a bilingual dictionary
Based on the electronic dictionary metaphor, Sánchez and Cantos designed a DDL-like software: Practicatn Vocabulario (PTV). PTV is a Spanish lexicon learning software, containing the 4500 most frequent types occurring in the CUMBRE Corpus — linguistic corpus of contemporary Spanish (Sánchez et al. 1995). All 4500 items:

- Are translated into English, French, German, Portuguese and Italian. By just clicking on the desired flag, students will get the words translated in that language. However, students might change translation language any time at will (Figure 8).
- Can be accessed, using standard electronic search facilities: term search, window scroll or thumb index (Figure 9).
- Are illustrated with a real example — full concordance sentence, extracted from the CUMBRE Corpus (Figure 10).
- Are recorded and can be heard by the students.

© Servicio de Publicaciones, Universidad de Murcia. All rights reserved.
Figure 9. Search Facilities

Figure 10. Visualising Concordance-sentence and Translated Term

© Servicio de Publicaciones, Universidad de Murcia. All rights reserved. IJES, vol. 2 (1), 2002, pp. 129-165
The option Ejercicios offers three types of exercises:

*Listen, repeat, record and check your pronunciation.* On selecting this exercise, students will choose the number of words they wish to work with by clicking on the button with the number of examples which will, at random, be the basis of the exercise. Next, the random-selected words will appear. and by clicking on the loudspeaker icon, the blue-highlighted word can be heard. Finally, students click on the microphone icon and record the highlighted word. A click on the right loudspeaker reproduces the model recording followed by the student’s recording and the student can contrast both outputs (Figure 11).

*Listen and write.* This is a word dictation practice; students will hear randomly chosen words and will have to write them correctly. The program allows three guesses before displaying the correct spelling. To facilitate the writing of Spanish diacritics, PTV provides them on a small table below the text-entry window (Figure 12).

*Read in your language and translate into Spanish.* Here the program displays randomly selected words in the target language chosen and students have to write the translation for each word into Spanish (Figure 13).

![Figure 11. Exercise Type 1: Listen, Repeat, Record and Check your Pronunciation](image)
Figure 12. Exercise Type 2: Listen and Write

Figure 13. Exercise Type 3: Read in your Language and Translate into Spanish
III. INTEGRATING DDL AND LANGUAGE TECHNOLOGIES

In recent years, a new term has been coined by the CALL community: Human Language Technologies (HLT). This term embraces a wide range of research and development areas within the area of Language Engineering or Language Technologies.

The field of human language technology covers a broad range of activities with the eventual goal of enabling people to communicate with machines using natural communication skills. Research and development activities include the coding, recognition, interpretation, translation, and generation of language. Advancements in human language technology offer the promise of nearly universal access to on-line information and services. Since almost everyone speaks and understands a language, the development of spoken language systems will allow the average person to interact with computers without special skills or training, using common devices such as the telephone. These systems will combine spoken language understanding and generation to allow people to interact with computers using speech to obtain information on virtually any topic, to conduct business and to communicate with each other more effectively. (Cole 1996)

III.1. Some HLT tools

There are many HLT tools that have become commercial systems. Among those systems, probably the two areas that have focused most commercial and scientific motivation are Machine Translation (MT) and Speech Recognition (SR). Particularly interesting here is the possible application domain of MT and SR to CALL and more generally, language teaching and learning, and other HLT tools. Interesting in this respect are part-of-speech (POS) taggers and syntactic parsers. These two Natural Language Processing (NLP) tools might help teachers and learners to pre-process texts and highlight certain grammatical phenomena or patterns without the trouble of having to manually annotate a text.

In the following sections, we shall introduce some HLT applications and try to highlight their interest for language teachers and learners, in general, and also for non-HLT initiated CALL practitioners.
III.1.1. Machine Translation

From its earliest days, MT has been bedevilled by grandiose claims and exaggerated expectations. In present day, however, the term MT is generally the standard for computerised systems responsible for the production of translations from one natural language into another, with or without human assistance.

Although the ideal may be to produce high-quality translations, in practice the output of most MT systems is revised and edited. In this respect, MT output does not differ much from the output of most human translators which is normally revised by other translators before dissemination. MT output may also serve as rough or raw translations.

While many of the commercially available MT packages may be useful for extracting the gist of a text they should not be seen as a serious replacement for the human translator. Most interlingual translations are not half-intelligible, letting you know whether a text is worth having translated properly aid where there are many situations where the ability of MT systems to produce reliable, if not perfect, translations at high speed are valuable. Even where the quality is lower, it is often easier and cheaper to revise ‘draft quality’ MT output than translate it entirely by hand. The translation quality of MT systems depends mainly on restrictions of the translation domain, linguistic architecture and components.

Imposing restrictions on the input such as (a) limiting the texts to particular sublanguages of document type and subject field and/or (b) controlling the language (reducing ambiguities, colloquial expressions, etc.), may improve translation quality.

Regarding MT architecture, the first MT systems are generally referred to as having a direct translation approach. The main idea behind this architecture is that source language sentences can be transformed into target language sentences by shallow analysing the source text, replacing source words with their target language equivalents as specified in a bilingual dictionary, and then roughly re-arranging these order to suit the rules of the target language. The second basic type is the interlingua approach. This type assumes the possibility of converting texts to and from meaning representations common to more than one language. Translations consist of two stages or phases: (1) from the source language to the interlingua and (2) from the interlingua to the target language. The third type of MT systems, the transfer approach, involves three stages: (1) converting source texts into intermediate representations in which ambiguities have been resolved irrespectively of any other language, (2) converting these into equivalent representations of the target language, and (3) generating the target texts (translations).

Some other MT systems rely less on the approaches mentioned above. Example-based interlingual translation, for instance, does not employ mapping between languages but instead matches stored translation examples against each other using a bilingual corpus of translation pairs (Nagao 1984). An even more radical approach to MT is the statistical approach (Brown et al. 1993) which requires the use of large bilingual corpora which serve as input for a statistical...
translation model.

Regarding language teaching, MT systems can be easily and efficiently integrated into the learning process. Some potential applications are:

- Translating full texts or paragraphs. Students can translate and then read the texts in their own language, extracting the gist without teacher intervention (Figure 14). In most MT systems, users can translate sentences automatically or interactively. Automatic translation proceeds autonomously without the intervention of the user. Whereas in interactive translation, the user can intervene in the translation process and choose the best word whenever more than one translation is possible.

- Translating sentence-by-sentence and print the source and target texts in a line-by-line format. This layout can be useful for comparing the original and translated text. This allows students to explore for equivalents between the source and target language, look for erroneous translations/false friends and assist in their own translations. (Table 1).

- Studying or writing in a foreign language.

- Looking up words (dictionary) and their inflections (Spanish grammar) (Figure 15).

![Figure 14: Automatic Text Translation (English-Spanish)](image)

© Servicio de Publicaciones, Universidad de Murcia. All rights reserved. IJES, vol. 2 (1), 2002, pp. 129-165
1. The use of concordancing in literature and linguistic analysis is nothing new.

2. It started well before computers existed.

3. Tribble and Joies (1990) trace the history of concordancing from the 13th century, when Hugo de Sancto Caroli enlisted 500 monks in producing a complete concordance of the Latin Bible.

4. That is, a kind of reference work designed to assist in the exegesis of the Bible, consisting of all occurrences of certain words, etc., that were felt to be significant, and presented in a way that would help the researcher.

Table 1. Line-by-line Printed Translation

<table>
<thead>
<tr>
<th>MicroTec Spanish Assistant - [Spanish Word Indications]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
</tr>
<tr>
<td>Imperfect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Preterite</td>
<td>yo querer</td>
<td>él quiso</td>
<td>habíamos querido</td>
<td>nosotros queríamos</td>
<td>ellos querrán</td>
<td></td>
</tr>
</tbody>
</table>

Figure 15. Inflection Look-up
III.1.2. Part-of-Speech Tagging

Word frequency lists derived by computers from corpora have clear shortcomings. These computer counts and sorting of word forms somehow bury or distort important facts about the language: variant inflected forms of nouns and verbs especially would be treated as entirely different word types. For example, be, am, are, is, was and were would be accounted for completely different linguistic items. Similarly, the frequency count of the number of occurrences of the word form light in a corpus would include the noun, verb, adjective and adverb.

Because manual annotation of each word token with its parts-of-speech (POS) in the corpus would be too expensive, the solution adopted has been to design computer programs, known as POS-taggers, to annotate automatically every word in the corpus with a tag to show the POS it belongs in context.

TAGGIT* was the first computer program designed and implemented to annotate a major corpus and assigned 87 tags to the word forms in a corpus. Subsequent developments in POS-taggers found necessary to expand the tagset and to modify the rule-based approach of TAGGIT.

Other tagging systems, such as CLAWS*, are based on probabilistic principles and are remarkably robust. In particular, CLAWS uses 133 basic word and punctuation tags and gets a minor error margin of just 3.4%.

Another extension of automatic POS-tagging is the combination of rules and stochastic or probabilistic principles. This is found in eTiKeT@". Actually, this HLT tool is not a tagger but a tagger-generator. It has not been designed for any specific language but, in principle, for any language. It starts from scratch: with an empty lexicon (data base), without any linguistic information (rules) nor probabilistic data and uses just 14 tags (Table 2). The user's task is to train or "teach" it for the language desired. All sessions are stored and the manual tagging is compared with the system's performance (Tables 3a, 3b and 4). Once a satisfactory success rate has been achieved, the system can be left to perform automatically without human intervention.

To speed up the initial human tagging phase, the user can alternatively feed the system's lexicon with stoplist items. That is, high-frequent non-ambiguous types, mostly close-class items, such as pronouns, prepositions, conjunctions, articles, auxiliary and modal verbs, etc.

The program tags on a sentence-by-sentence basis and outputs the results either in a database mode (Figure 16) or as running ASCII text with the tags attached to the tokens in the text (Figure 17). Additionally, the user can also consult the patterns and statistics the system has inferred so far (Figure 18).
<table>
<thead>
<tr>
<th>ID</th>
<th>POS</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Noun</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>Verb (lexical)</td>
<td>V</td>
</tr>
<tr>
<td>2</td>
<td>Verb (aux)</td>
<td>Aux</td>
</tr>
<tr>
<td>3</td>
<td>Verb (modal)</td>
<td>Mod</td>
</tr>
<tr>
<td>4</td>
<td>Adjective</td>
<td>Adj</td>
</tr>
<tr>
<td>5</td>
<td>Adverb</td>
<td>Adv</td>
</tr>
<tr>
<td>6</td>
<td>Preposition</td>
<td>Pre</td>
</tr>
<tr>
<td>7</td>
<td>Particle</td>
<td>Par</td>
</tr>
<tr>
<td>8</td>
<td>Conjunction</td>
<td>Con</td>
</tr>
<tr>
<td>9</td>
<td>Interjection</td>
<td>Int</td>
</tr>
<tr>
<td>10</td>
<td>Determiner</td>
<td>Det</td>
</tr>
<tr>
<td>11</td>
<td>Pronoun</td>
<td>Pro</td>
</tr>
<tr>
<td>12</td>
<td>Punctuation</td>
<td>Pun</td>
</tr>
<tr>
<td>13</td>
<td>Other</td>
<td>Oth</td>
</tr>
</tbody>
</table>

Table 2. Tag-set of eTiKeTe

<table>
<thead>
<tr>
<th>SessionCode</th>
<th>FileCode</th>
<th>CorrectGuess</th>
<th>WrongGuess</th>
<th>LeftContext</th>
<th>RightContext</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>2187</td>
<td>1407</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>63</td>
<td>50</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3a. Information on Session Performance and POS-Disambiguation Context selected

<table>
<thead>
<tr>
<th>TagLastWord</th>
<th>Date</th>
<th>FinishedSession</th>
<th>JustWords</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2001-04-19 14:51:09.36</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2001-04-19 14:52:35.21</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2001-04-19 14:54:48.02</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2001-04-19 14:57:17.85</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2001-04-19 14:58:53.97</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2001-05-10 15:25:29.61</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2002-04-25 10:20:48.49</td>
<td>YES</td>
<td>NO</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3b. Information on Tags, Date, Session, Text or Single Word Tagging and Language

© Servicio de Publicaciones, Universidad de Murcia. All rights reserved.
<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Frequency</th>
<th>SessionCode</th>
<th>CorrectGuess</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>10</td>
<td>62</td>
<td>0</td>
<td>YES</td>
<td>0</td>
</tr>
<tr>
<td>about</td>
<td>6</td>
<td>17</td>
<td>7</td>
<td>YES</td>
<td>0</td>
</tr>
<tr>
<td>above</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>YES</td>
<td>0</td>
</tr>
<tr>
<td>accept</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>acceptable</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>achieve</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>achieved</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>achieves</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>achieving</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>across</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td>YES</td>
<td>0</td>
</tr>
<tr>
<td>actual</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>YES</td>
<td>0</td>
</tr>
<tr>
<td>additional</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>address</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>addressed</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>administration</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>admiral</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>advance</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>affect</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>NO</td>
<td>0</td>
</tr>
<tr>
<td>after</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>YES</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4. Data Base Extract (Types, Tag, Frequency, Session, First-Time Guessing of the Type and Language)

Figure 16. eTikEti (Data Base Layout)
Integrating Corpus-based Resources and Natural Language Processing Tools into CALL.

Figure 17. ASCII Layout

Figure 18. Information on Inferred Patterns and Statistics

O Servicio de Publicaciones. Universidad de Murcia. All rights reserved.  IJES, vol. 2 (1), 2002, pp. 129-165
III.1.3. Lemmatisation

The distinction between words (tokens), the word forms (types) and base forms (lemmas) is important. Consider the following word sequence: play, playing, played, play, plays, playing, played and played, where we have nine words (tokens), four word forms (types) and one lemma, namely play. As mentioned above, standard concordancers would process inflected forms (tokens) of the same base form (lemma) as different word forms (types). A way of dealing with this and other potential problems (see POS-section), which can seriously affect the counting of linguistic items, is to classify together all the identical or related forms of a word under a common headword: lemmatisation; just as in a dictionary where the various morphological inflected and derived forms of a word are listed under a single entry. In order to handle the complexities of morphology, including irregularities, lemmatisers typically employ two different but combined processes: (1) a stem-based method to deal with irregularities, by means of rules; for example, better and best are listed and counted under the headword good; (2) a suffix stripping method; if a word form is not listed under any headword (stem-based method), then a number of suffix stripping rules are applied; for example, the plural suffix -s is taken off the word form cats, outputting the base form cat. Finally, if a word form does not appear as part of the suffix rule...
system, or is not listed as a specific exception, then the word is listed as a lemma in the same form in which it appears in the text. The lemmatisation process is normally performed automatically as part of the POS-process, producing enlarged tagged data lists: token, tag and lemma and grammatical information (gender, number, tense, etc.), i.e. TL-Tag (Figure 19).

A useful CALL application based on the lemmatisation process is Verbo...
A much more interesting and challenging application is Tagcorder\textsuperscript{1}. Tagcorder has been implemented to allow complex searches within the CUMBRE Corpus (Sánchez et al. 1995) and takes full advantage of tagged and lemmatised data. Users can invariably look for terminal nodes (types; Figure 22), non-terminal nodes (POS-tags; Figure 23) with any additional tagged grammatical information (number, person, mood, tense, etc.), base forms (lemmas; Figure 24) and/or any combinations of types. POS-tags and lemmas; Figures 25 and 26). The program itself is very interactive and flexible in its search procedure and extremely fast as it works with pre-indexed text.

![Figure 22. Type Search "ABC"](image)

\textsuperscript{1}Seguramente la razón a ABC, que desde hace varios años propone que una misma persona no pueda estar más de ocho años en la presidencia del Gobierno.
El director de la Real Academia Española ha sido nombrado doctor "honoris causa" por la Universidad de Buenos Aires, a propuesta de la Facultad de Filosofía y Letras, en reconocimiento a su destacada trayectoria en los estudios de lengua y literatura española.

El director de la Real Academia Española ha sido nombrado doctor "honoris causa" por la Universidad de Buenos Aires, a propuesta de la Facultad de Filosofía y Letras, en reconocimiento a su destacada trayectoria en los estudios de lengua y literatura española.

El director de la Real Academia Española ha sido nombrado doctor "honoris causa" por la Universidad de Buenos Aires, a propuesta de la Facultad de Filosofía y Letras, en reconocimiento a su destacada trayectoria en los estudios de lengua y literatura española.
Se enfrenta a un rival elegido por una ambigua legación. Es un cirujano que se olvida de su propia piel, su presencia en el tribunal y en la televisión del mundo llenó de su mirada al país. Milosevic para arrancárselo a la oposición y traerlo a su lado. Honestidad que se ha puesto y se va a probar a prueba ante el proceso de Máximo, para caracterizar a la oposición y averiguar a su campo incluso como nuevo primer ministro.

Figure 25. Complex Search: "UN" + NOUN (Coindexed Countable + Masculine + Singular)

Figure 26. Complex Search: VERB (Main + Two Clitics + Infinitive)
III.1.4. Parsing

Parsing involves the procedure of bringing basic morphosyntactic categories into high-level syntactic relationships with one another. This is probably the most commonly encountered form of corpus annotation after POS tagging.Parsed corpora are sometimes known as treebanks.

There are rules governing the way in which words can be put together to form syntactically well-formed or grammatical sentences; the study of syntax aims to discover them and to describe and analyse language in terms of these rules. Consider the sentence A dog chased that girl, where we find the same pattern of constituents before and after the verb, that is determiner + noun. These two words also appear to be belong together more closely than say the noun dog and the verb chased. Another way of illustrating that these words belong together is to give the girl and the dog name—names of specific items such as individual people, animals, places and so on called proper nouns—and we get, for example Henry chased Carol.

It seems clear that natural languages or human languages have a role of constituent structure. A sentence is not just a mere string of words. The words are grouped into phrases, each of which consists of a short phrase. Many of the important properties of languages are organised around constituent structure. Constituent structures (a) group words into constituents such as the rig and into the garden; (b) give names to the constituents, such as noun phrase and prepositional phrase. In turn, constituent structures are sanctioned or generated by rules, known as phrase-structure rules of this type:

\[
S \rightarrow NP \ VP \\
NP \rightarrow N \\
V \rightarrow V
\]

where S stands for sentence, N for noun phrase, VP for verb phrase, N for noun and V for verb. So the PS-rules above states that a sentence consists of a noun phrase followed by a verb phrase. In turn, the NP of an N and the VP of a single V. The tree structure derived or generated by that rule would be

![Tree Structure](image)

Parsing algorithms can proceed top-down or bottom-up. In some cases, top-down and bottom-up algorithms can be combined\textsuperscript{15}.

©Servicio de Publicaciones, Universidad de Murcia. All rights reserved. IJES, vol. 2 (1), 2002, pp. 129-165
The Visual Interactive Syntax Learning (VISL) website is particularly interesting and useful for language learners. It contains an on-line parser and a variety of other tools concerned with English grammar, including games and quizzes. The parser itself is an excellent and very transparent application that allows learners to analyse and experiment on sentences and study their structure (Figure 27).

Interesting in this respect is also the parsing of students' erroneous input. Integrated parsers into CALL software can be prepared to deal with linguistic errors in the input. So the grammar that copes with correct sentences is complemented with a grammar of incorrect sentences. The advantage of this error grammar approach is that the feedback to students' output can be very specific and is normally fairly reliable as it can be attached to a very specific rule. However, the major drawback of this approach is that individual learner errors have to be anticipated in the sense that each error needs to be covered by an adequate rule.

**Figure 27. Visual Interactive On-Line Parser**

### III.1.5. Speech Technology

CALL software has normally been restricted to written text. However, recent advances in multimedia have resulted into powerful hardware and software applications, allowing users to
attach a microphone and loudspeakers to soundcards and to record his/her own voice. Furthermore, storing these sound files is no more a problem due to the immensely increased capacity and cost reduction of hard disks, other storage devices (CD-ROM) and improved compression algorithms for this kind of data (i.e., MP3 files).

Presently, there is a wide range of speech software available. This includes (a) spoken input processing (2) or speech analysis, where speech input is analysed and represented graphically or numerically; (2) speech recognition: the transformation of spoken input into written output; and (3) speech synthesis, that is the conversion of text to speech (3); this includes no just matching characters to sounds, but also intonation and the rhythm of particular utterances have. Advances in speech synthesis technology have reached a high level of performance and robustness and some CALL applications have started considering its integration (3).

In contrast, speech recognition is far more complex than speech synthesis. Speech recognition needs an extensive analysis of speech by means of a number of parameters, which are very difficult to establish as they can be easily affected by background noise, speech speed (connected speech), particular accents or idiosyncratic individual’s speech. All this leads to complicate and interfere in the fixing and interpretation of the established parameters.

There are some commercial applications able to "understand" natural speech and can provide language students with realistic, highly effective, and motivating speech practice. One of these applications is IBM® ViaVoice®. This program runs on normal PCs and includes speaker independent continuous speech recognition engines and is able to deal with complete sentences spoken at a natural pace, not just isolated words. Though it requires a minor training period. To run the program, the user just needs to associate it with any wordprocessor, where the user's utterances will be transcribed in (Figure 28) or run Speech Pad, a simplified standard wordprocessor that includes ViaVoice.

Many multimedia CALL courses already have and still include some naive direct pronunciation practice. That is, exercises which focus on pronunciation, fluency and word order, and with native speaker models which are heard immediately after a student's performance. These applications leave the learner-model comparison to student's criteria or visualise graphically both performances, indicating the success rate in %. The negative side of these exercises is that in some instances it is even for native speakers of the language very difficult, if not impossible, to achieve satisfactory success rates. Some of these applications are neither very flexible nor accurate, and sometimes students would need to repeat their utterances in several occasions before the program "understands" them correctly. In turn, this might lead to some small frustration.
V. CONCLUSION

CALL has for long been dominated by the drills and practice associated with behaviourism (cloze and gap-filling exercises, multiple-choice tests, etc.) and, eventually, the use of some basic word-processing tools. Soon, some language teachers and CALL practitioners reacted negatively and noted a lack of progress in CALL (Kaliski 1993, Last 1992, etc.), partially due to the:

- Limited CALL software available
- Reduced number of computational exercises
- Incompatibility between employed CALL techniques and current language teaching pedagogy (particularly influential in this respect has been the emergence communicative syllabus)
- Consequence of new technology being unable to fulfill teachers' expectations

Fortunately, things have changed for CALL in the 90s, partly because of the wider availability of PCs and the integration of linguistic corpora and NLP-technology. The use of linguistic corpora and NLP-applications are highly valuable tools for language description with important implications for language teaching, as they can:

- Assist language teacher in identifying relevant content of instruction (vocabulary, grammar, contexts, etc.), and
- Help in developing new pedagogical and methodological approaches to instruction (i.e. shifting the pedagogical teaching/learning paradigm from computer as magister to computer as pedagogue).

Modern CALL has changed and instead of adapting it to what software can offer, an attempt is made to get it to take account of the necessary conditions of successful language learning. Learners are given much more control over what they learn: autonomous language learning in self-paced, non interactive, meaning dominated, task-oriented activities (Kennedy

particularly interesting in this respect is the use of real and relevant text sample for students and teachers as the central pedagogical teaching/learning cornerstone. Real time manipulation of texts by students using integrated user-friendly interfaces, including word-processing tools and NLP-applications could conforni an extremely valuable pedagogical paradigm within the foreign language learning/teaching context. Teachers would be able:

- To extract, manipulate and adapt texts to students needs and language level
- To enrich plain texts with POS-tags and syntactic annotations for class work
- To extract vocabulary lists, phrase lists, concordance lists, etc. (sublanguage specific, adapted to a specific level. or domain, etc.)
- To generate automatically ad hoc exercises, depending on students' particular needs.

Similarly, students could also take advantage of this integrative CALL application:

- To explore the target language by means of concordancers with integrated taggers and parsers and/or tagged texts and treebanks
- To extract the gist of more difficult texts, using MT-software
- To check the meaning of words and phrases (electronic dictionaries)
- To generate automatically ad hoc exercise generation, depending on one's own needs
- To hear the text, selected sentences or words, using speech synthesizers
- To answer orally to some responses, dictating the solutions to the computer (speech recognition tools)

Actually, what we propose is a sophisticated CALL language processing tool\textsuperscript{31} that

- Takes full advantage of current computational advances in an integrated and unitary way:
  - Electronic dictionaries (monolingual, bilingual or multilingual ones)
  - MT systems
  - POS-taggers
  - Syntactic parsers
  - Concordancers
  - Speech production/recognition
  - Word-processors (this includes spell-checkers and grammar and style checkers)
- Goes beyond written text as it also accounts for oral production and oral recognition
- Assists both teachers and students in their respective tasks and that could contribute to new and challenging pedagogical and methodological paradigms in the area of foreign language learning.

© Servicio de Publicaciones. Universidad de Murcia. All rights reserved. IJES, vol. 2 (1), 2002, pp. 129-165
And since we have all these computational tools at our disposal, it makes no sense to renounce their application in such an important area as language pedagogy. We cannot dismiss them, we must use them ...

NOTES

3. Context can be downloaded free of charge for non-commercial purposes from Tim Johns DDL-web page: http://web.bham.ac.uk/johnst/miconc.htm
5.htm
6. Hutchiiis and Sonders (1992) and Arnold et al. (1994) provide excellent introductions to MT.
8. See Greene and Rubin (1971) for a detailed description of this tagger.
10. To get a free copy, for academic purposes only, e-mail Rafael Valencia@rafavalencia@cono.es, Rodrigo Martinez (rodrigo@cono.es) or Pascual Cantos (pcantos@cono.es).
11. Where 0 = English and 1 = Spanish.
12. TL-Tagger (Technical linguistics Project) and the CUM/RECorpus Project aid is not yet commercially available. For those interested in it contact one of the people involved in the project: Enrique Pérez de Lema (enrique@jazzfree.com), Jose Simón (jsimon@teleline.es), Aquilino Sánchez (asanchez@cono.es) or Pascual Cantos (pcantos@cono.es).
14. See ciidilato 12.
20. The Polytechnical University of Hong Kong site includes a number of text-to-speech tools: http://vce.polyu.edu.hk/text2speak.
21. We have deliberately not considered the integration of Information Technologies here. This would have inevitably expended the potential of the "tool" proposed here.
REFERENCES


O Servicio de Publicaciones, Universidad de Murcia. All rights reserved. *LJES*, vol. 2 (1), 2002, pp. 129-165.


© Servicio de Publicaciones. Universidad de Murcia. All rights reserved.