International Journal of English Studies

IJES

UNIVERSITY OF MURCIA

www.um.es/engphil/ijes

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

PASCUAL CANTOS-GOMEZ*
Universidad de Murcia

ABSTRACT

This paper ainis at presenting a survey of computational linguistic tools presently available but whose potcntial has been ncither fully considered nor exploited to its full in niodern 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 dictioiiaries and electronic dictionary facilities) to DDL .

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

The paper finishcs with a number of reflections and ideas on how language technologies can be used cfficiently within the language learning context and how extensive exploration and intcgration of these tcchnologies might change and extend both modern CALL and the present language learning paradigiii..

KEYWORDS: Concordancing. corpus linguistics, data-driven learning, electronic dictionaries, Iiunian language tcchnologies, linguistic corpora, inachinc translation, morphological generation, NLP-tools, parsing, POS-tagging, speecli technology

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

I. 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 behaviour 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 behaviour and how this behaviour 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*

There is ample discussion in the literature on the merits of linguistic corpora in second language teaching and learning (Aston 1995, 1996, 1997; Ball 1995; Barlow 1992, 1995; Burnard and McEnery 1999; Celce-Murcia 1990; Collins 1999; Flowerdew 1999; Gavioli 1997; Higgins 1991a, 1991b; Johns 1986, 1988, 1991, 1993; Johns and King 1991; Leech and Candlin 1986, etc.), mainly as a result of the pioneering work of Tim Johns at Birmingham University (1986, 1988).

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 (1988) proposes *concordances* as the central idea to shift the pedagogical teaching/learning paradigm from computer as magister to computer as pedagogue. That is, from 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, Kettemann (1995) compares the treatment of the English conditional clause in a standard grammar used in Austria (Kacowsky 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 Kacowsky'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 issuc with DDL is authenticity. Widdowson (1983:30) considers that

> An authentic stimulus in tlie forrn of attested instances of language does not guarantee an authentic respoiise in tlic forin of appropriate language activity [...] we should retain the term 'authenticity' to refer to activity (i.e. process) and use tlie terin 'geiiuine' 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 languagc activity" (Tribble 1997). That is, it is likely that forcign languagc students are not necessarily niotivated by a language learning activity if the instances of language use that they are studying are cxtractcd from contexts that have little or no connection with their interests and concerns. Gcnuine examples of language in use will not necessarily lead to autlientic language use or effective language learning activities.

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

> Maiiy native speakers inake use of otliers' writiiig or speech to inodel their owii work in their native language where tlie geiirc is unfamiliar. It is tiiiie tliat this skill was brouglit out oftlie closet, aiid exploited as aii aid for learning.

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

Clearly, this, somehow, relegatcs standard, balanced and representative corpora, such as thc Brown corpus of American English (Kučera and Francis 1967), the Lancaster-Oslo-Bergen (LOB) corpus of British tcxts (Johansson 1980), or other major corpora such as the British National Corpus (BNC) (Burnard 1995), a 100 million word representative corpus of contcniporary British written and spoken texts, or tlie 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 multimcdia encyclopaedia, such as Microsoft Encarta®, among others. The latest version of Microsoft Encarta® contains more than 30,000 articlcs, between 200 and 5000 words, which count for a total of roughly 30 million words, covcring 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 multimcdia encyclopaedia virtually contain enough tcxts which niost studcnts in niost languagc classcs will find interesting and informative. In addition, with this coniprehcnsive range of topics and tcxts. it is not difficult to sclect *ad hoc* texts. focussing 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 our comprehensive and representative corpora.

II. A WOKD ON EXISTING DDL-SOFTWARE

In this section, we shall review the niain software applications used among DDL practitioners: (1) commercially available concordance progranis and (2) Tim Johns' *Context*. In addition. we sliall also present a Spanish vocabulary learning niultiniedia application. *Practica tu Vocabulario* (Sánchez and Cantos 2000), which is based on tlie electronic dictionary nietaphor and DDL-like learning/acquisition strategies.

II.1. Concondancers

Concordancers are text proccssing tools for looking at how words behave in texts. These tools allow you to find out how words are uscd in texts. Among the facilities. all concordancers allow you at least[1]:

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

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

Tim Johns has compiled nunierous cxercise cxaniples on his web page[2] using standard concordancing tools. The classroom materials that follow are extracts from his website and are a collection some participants' work of tlie *Ustí nad Labem DDL Workshop* (21st - 25th March 2000):

About and On

How do wc usc tlic prcpositioiis ABOUT aiid ON? Which oric is uscd iiiore often? Which oiic tcnds to be used in acadciiiic texts? In wliicli cases is tlierc tlic occurciice of orily onc oftliem?

Book

t be sent a British Medical Association book about a potential risk to numan health pe
2 erial market. 'We published o coffee-table book about ant behaviour called The Ancc, which
unds sterling 16 95 Publicised *os* a book about the terrible fate awaiting humanity f
4 in life. This is yet another 'qee-whizz" book about forensic science, this time based un
5 title intriqued me: at last, I thought, a book about the personal relationships that scien

21 een evergreen. I remember buying my first hook on planets (by Patrick Moore) hack in the 1
22 ds. How things have changed. A good book on the planets has always needed to he up-
23 fghanistan, so how did he come to write a book on Murdock? Was his choice dictated by the
24 aos I'm keeping that for myself. It's a hook on kilims - geometrically-patterned rugs fr
25 r (HarperCollins, 1990), and is writing a hook on the future of US nacional security polic

(by Kvita Rychtárová)

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

Task 1 - 'Grcat' aiid its collocations

Tiy to spot what tlic typical cases of 'grcat' and its collocations are on tlic basis of tlic following examples.

1.

Whether recent discoveries in the Great Pyramid of Cheops have anything to do
icated to restoring Al-Andalus and its Great Mosque to their former fñith
king themselves comfortable. And the great London clubs, with their roaring coal fires

2.

narratives of Les Miserables and Great Expectations. Whñt could be more comical
she agreed to make a Film in the Great Journey series, the theme of which was

Task 3 - 'Great' aiid missing nouns

Try to predict tlic words which are iriissiiig.

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

(Park, players, Race, King, Empires)

(by Sarka Canova and Jarka Ivanova)

Adjectives ending in -ic, -ical

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

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

II. how fill in tlic blanks with nppropriatc 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 hoth dancers. as in _____ ballet one partner may not be physically suited to perform with a stand-in.
3. ... we always dine at Cafe 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 _____ design inside out. Why do we need to see it?
5. ... arctiitect Kichard Norman Shaw, capable of turning out gathic, 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 tlie hest known and most used DDL-software is *Context*[3]. This program encouragcs language learner to invcstigate how words are used in context in English. and is designcd to supplement classes. It is based on short contexts (extracted from thc database by means of tlie computer prograni *MicroConcord*) illustrating the use of important key items from a database of over 3 million words of tcxt in English.

The program starts offering tlie user a list of headings: Top *Menu* (parts-of-speech and topics; sce *Figure* 1). In addition, it is also possible to view a more detailed index of all the keywords available to the prograni, together with the names of the files in which each key word is storcd and to selcct a file of contexts (by kcywords defined by parts-of-speech, keywords delined by topic or morphemes —prefixes or suffixes; *Figure* 2). Once the user has selected the file of context. the prograni displays the list of key items in the bottom of the screen (among other facilities) in ordcr to invcstigate 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 liavc finished tlic Quiz, they can see an analysis of their performance.

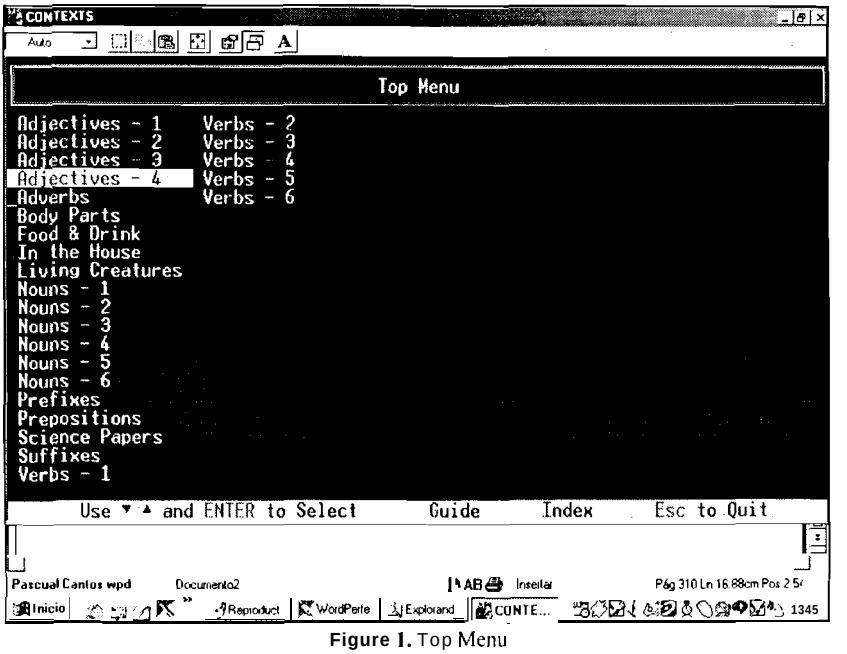


Figure 1. Top Menu

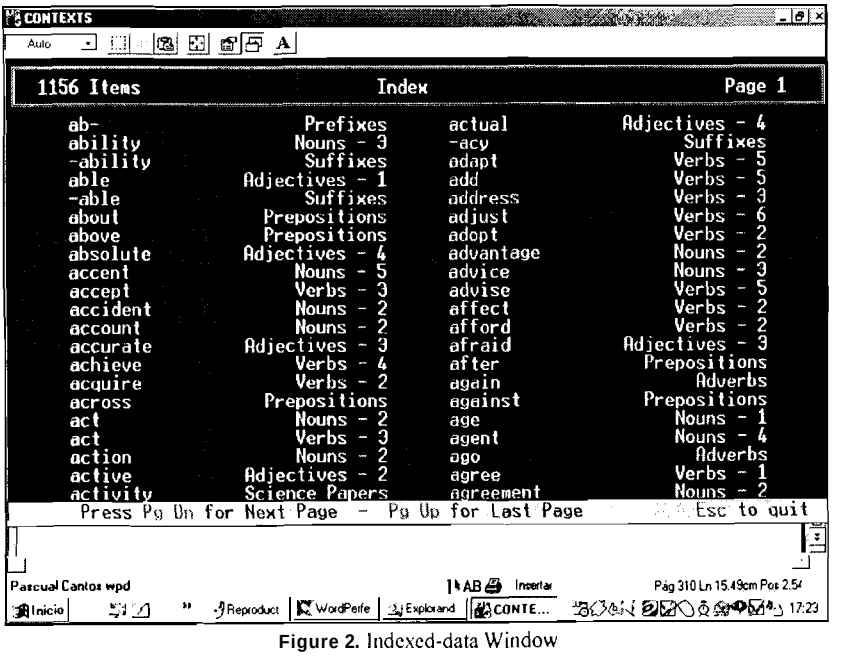


Figure 2. Indexed-data Window

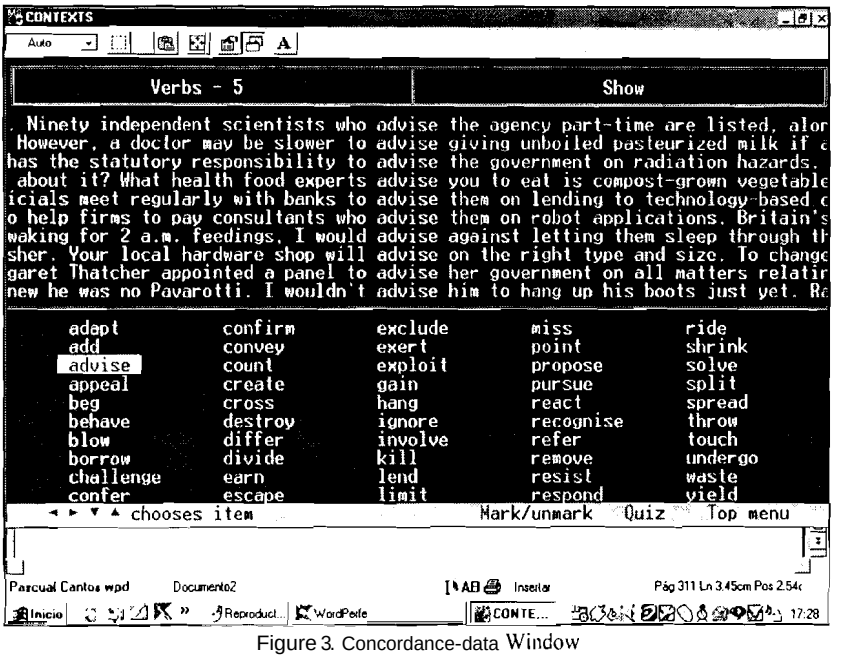


Figure 3. Concordance-data Window

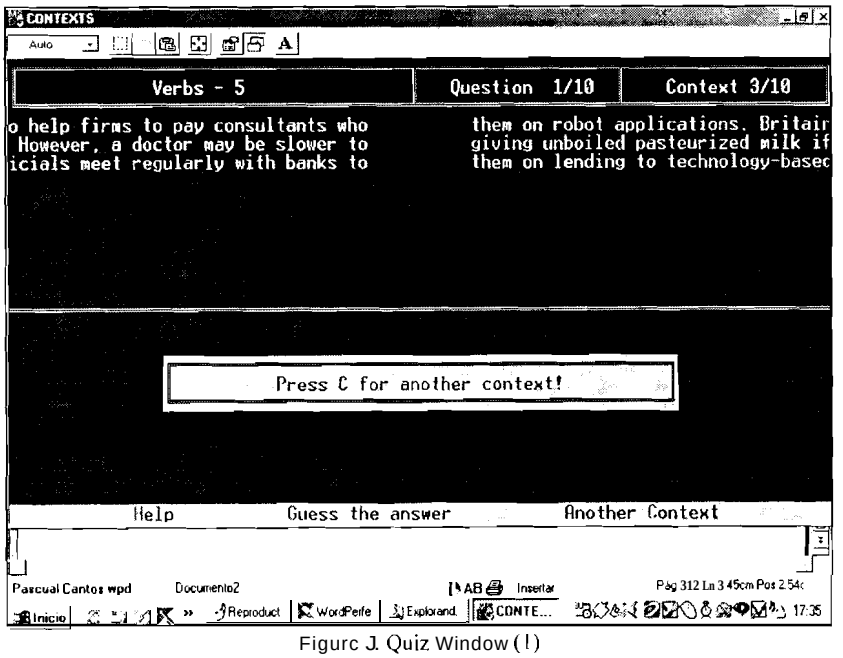


Figure 4. Quiz Window (1)

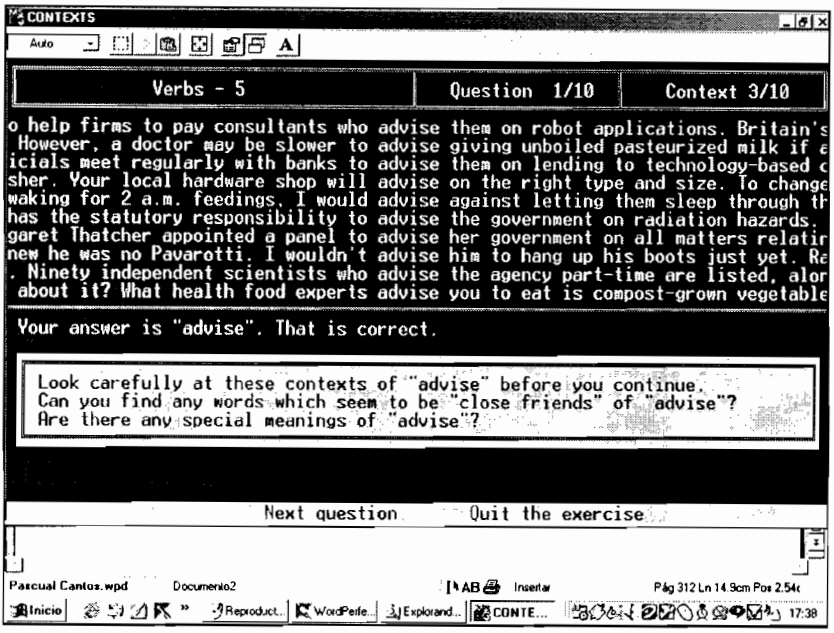
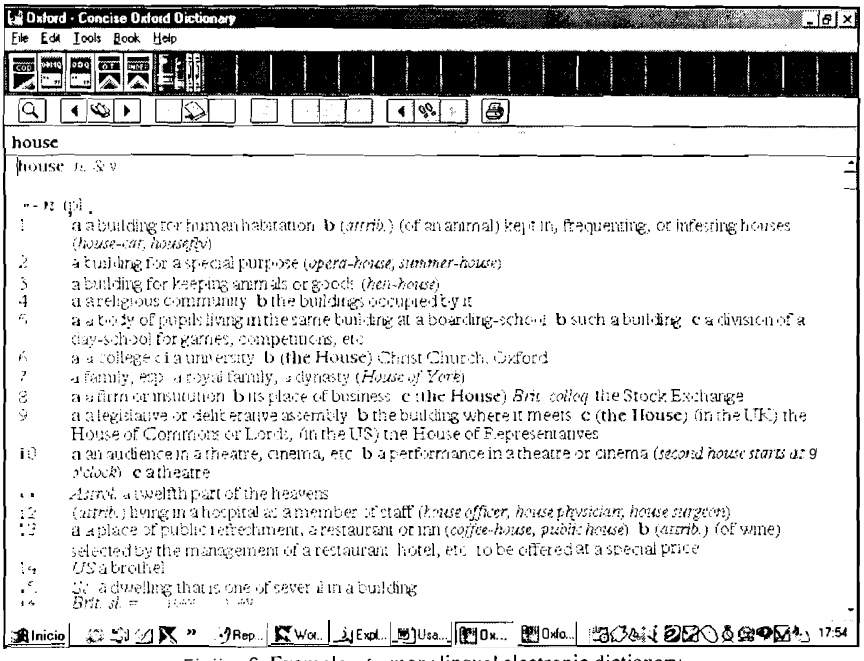


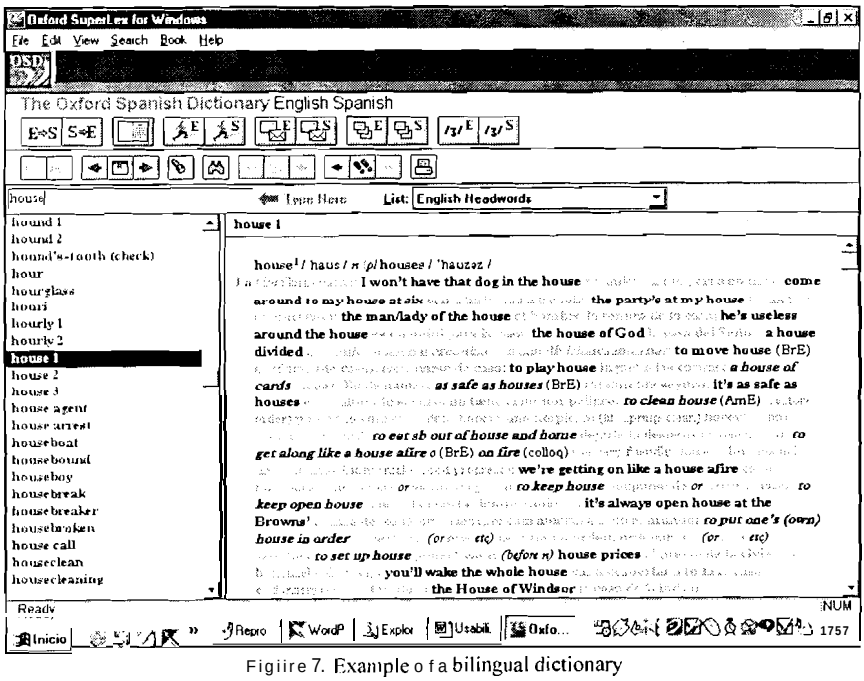
Figure 5. Quiz Window (2)

II.3. Using Electronic Dictionary Facilities for DDL

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*).



Figiire 6. Example o f a monolingual electronic dictionary



Figiire 7. Example o f a bilingual dictionary

Based on the electronic dictionary metaphor, Sánchez and Cantos[4] designed a DDL-like software: *Practica tu Vocabulario* (*PTV*). *PTV* is a Spanish lexicon learning software, containing the 4500 most frequent types occurring in the *CUMBRE Corpus* —a linguistic corpus of contemporary Spanish (Sánchez et al. 1995). All 4500 itenis:

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.

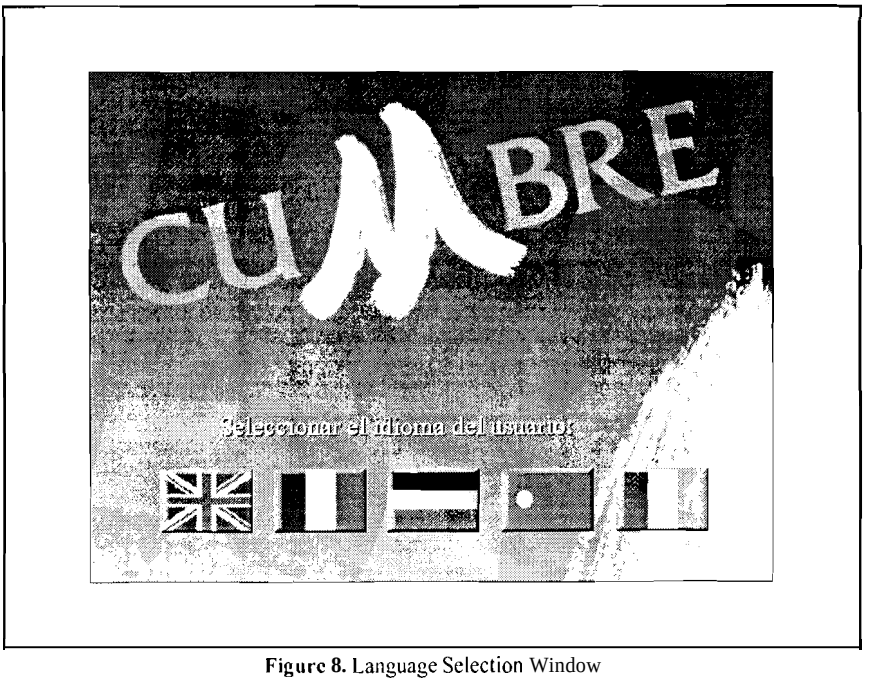


Figure 8. Language Selection Window

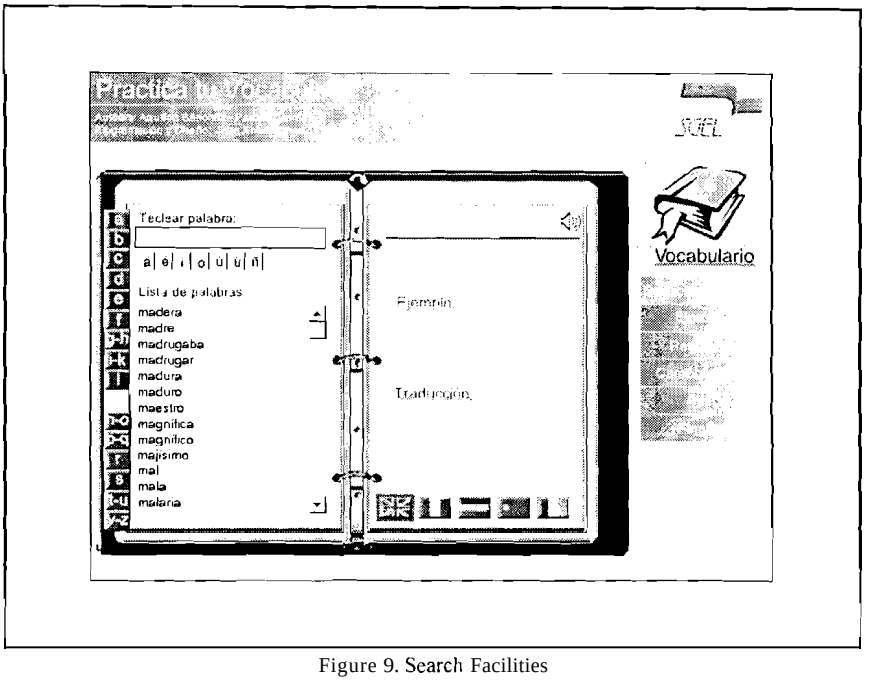


Figure 9. Search Facilities

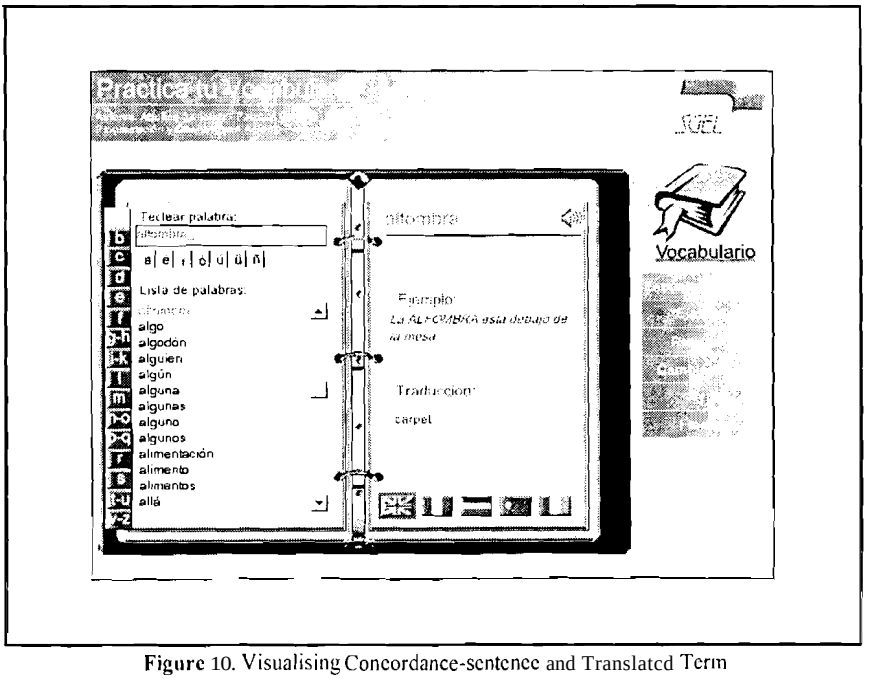


Figure 10. Visualising Concordance-sentence and Translated Term

The option *Ejercicios* offers three types of exercises:

Listen, repeat, record and check your pronunciation. On selecting this exercise. students will choose tlie nuniber 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-sclected words will appear. and by clicking on the loudspeaker icon, tlie blue-highlighted word can be heard. Finally, students click on the microphone icon and record tlie highlighted word. **A** click on tlie right loudspeaker reproduces the *model* recording followed by tlie 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 havc to write them correctly. The program allows three guesses before displaying thc correct spelling. To facilitate the writing of Spanish diacritics, PTV provides them on a small table below tlie text-entry window (*Figure 12).*
Read in your language and translate into Spanish. Here tlie program displays randomly selected words in the target language chosen and students have to write the translation for each word into Spanish (*Figure 13).*

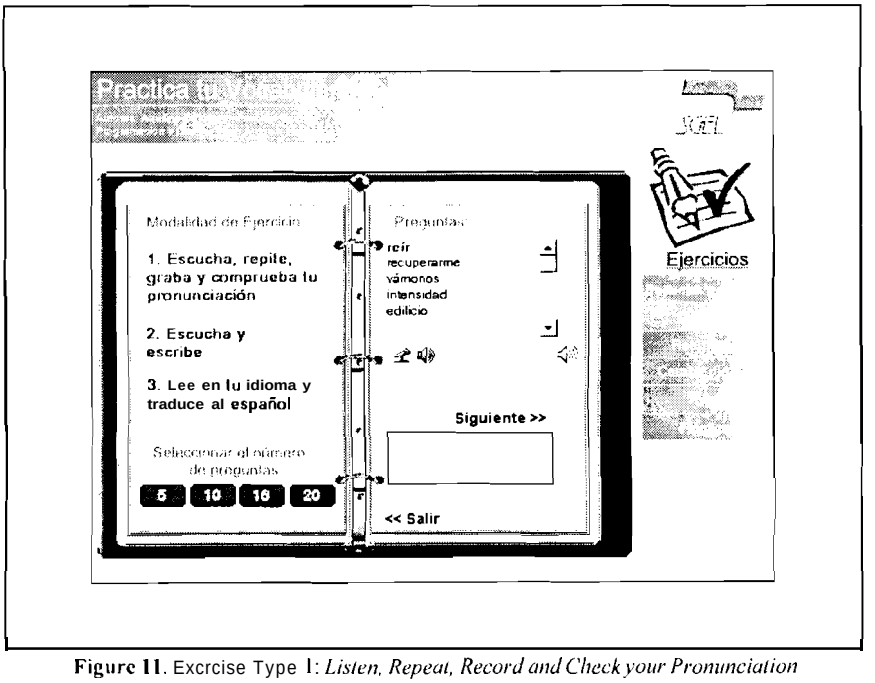


Figure 11. Excrcise Type 1: *Listen, Repeat, Record and Check your Pronunciation*

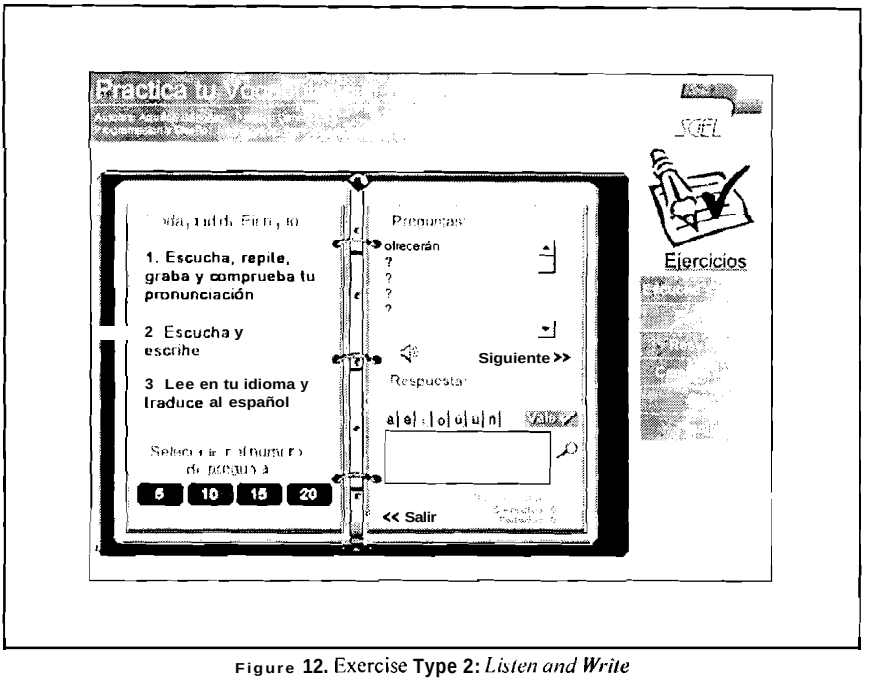


Figure 12. Exercise **Type 2:** *Listen and Write*

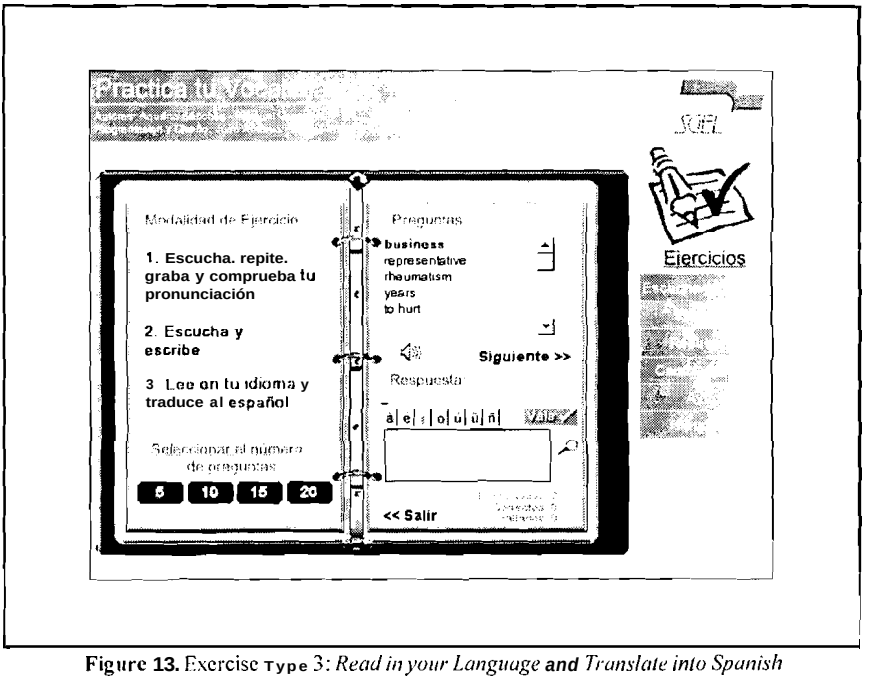


Figure 13. Exercise **Type** 3: *Read in your Language* **and** *Translate into Spanish*

On conipletion of each cxcrcise or at the end of a working session, students may consult his/her Iiits or failures by clicking on the *Resultados* tab on tlie right of the agenda.

III. INTEGKATING DDL ANU *LANGUAGE TECHNOLOGZES*

In rcccnt ycars, a ncw term has becn coined by the CALL community: *Human Language Technologies* (HLT). This tcrni embraces a wide range of research and developmcnt areas within the arca of Language Engineering or Language Teclinologics.

> The field of human laiiguagc tcclinology covers a broad range of activities with thc eventual goal ot cnabling pcoplc to coinrnunicate witli machines using natural coininunication skills. Rescarch and dcvclopiiiciit activities includc tlie coding, recognition. interprctation, translation. and geiieration of laiiguagc. ... Advaiiccs in human language tcclinology offertlie prorniseofnearly universal access to on-line information aiid scrviccs. Since almost evcryonc speaks and iindcrstaiids a language, tlie dcveloprnerit of spoken language systciris will allow tlie average pcrson to intcract with coniputers witliout special skills or training, using coininoii deviccs such as tlie tcleplrone. These systerns will coinbine spoken language uiidcistaiidiiigaiid gciicration to allow pcople to intcract witli cornputers using speech toobtain inforniation oii virtually any topic. to coiiduct business aiid to coininuiiicatc with each otlier niorc effectively. (Cole 1996)

III.1. Some HLT tools[5]

There are many HLT tools that have beconie commercial systems. Among those systems, probably tlie two arcas that have focused niost commercial and scientific motivation are Machine Translation (MI') and Specch 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. Intcrcsting 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 prcprocess texts and highlight ccrtain grammatical phenomena or patterns witliout the trouble of having to manually annotate a text.

In the following scctions, we shall introduce some HLT applications and try to highlight their interest Sor language tcacliers and learners, in general, and also for non-HLT initiated CALL practitioncrs.

III.1.1. Machine Translation[6]

Froiii tlic earliest days, MT has been bedevilled by grandiose claims and exaggerated expectatioiis. In prcsent day. Iiowever. tlie term MT is generally the standard for computerised systcnis responsible Sor tlic production of translations froni one natural language into another, with or without human assistance.

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

While iiiany of thc conin-iercially available MT packages may be useful for extracting the gist of a text tlicy should not be secn as a serious replacement for the human translator. Most iiiacliiiic translations are not that bad, tlicy are half-intelligible, letting you know whether a text is worth having translated properly aiid there are many situations where the ability of MT systems to produce reliable, if less than perfect, translations at high speed are valuable. Even where the quality is lowcr, 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 doniain, linguistic arcliitecture and coniponents.

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

Regarding MT arcliitecture, tlie first MT systenis are generally referred to as Iiaving a dirccl traiislation approach. The niain idea beliind this architecture is that source language scntences caii bc transformed into target language sentences by shallow analysing the source text, replacing source words with tlieir target language equivalents as specified in a bilingual dictionary. aiid then rouglily re-arranging tlieir order to suit the rules of the target language. The sccond basic type is tlic interlingua approach. This type assumes the possibility of converting texts to aiid from *meaning* representations coninion 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 tlie target language. The third type of MT systcnis, the transfer approach, involves three stages: (1) converting source texts into intcrniediate representations in which ambiguities have beeii resolved irrespectively of any otlier language, (2) converting these into equivalent representations of the targct language. and (3) generating the target texts (translations).

Soiiic other MI' systcnis rely less on the approaches mentioned above. Example-based iiiacliiiic traiislation. 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 evcn iiiorc radical approacli to MT is the statistical approach (Brown et al. 1993) which rcquircs tlie use of large bilingual corpora which serve as input for a statistical

translation modcl.

Regarding language teaching, MT systems can be easily and efficiently integrated into the learning process. Some potential applications are[7]:

- Translating full texts or paragraplis. Student can tratislate and then read the texts in their own language. extracting the gist without teacher intervention (Figure *14*). In most MT systenis, users can translate sentences automatically or interactively. Automatic translation proceeds autonomously. without the intervention of tlie user. whereas in interactive translation. the user can intervene in the translation process and choose the best word whencver more than one translation is possible.
- Translating sentence-by-sentencc 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 studetits to explore for equivalents between the source and target language, look to crroncous translations/false friends and assist in their own translations. (*Table 1*).
- Studying or writing in a foreign languagc.
- Looking up words (dicíionary) and their inflections (Spanish grammar) (Figure *15*).

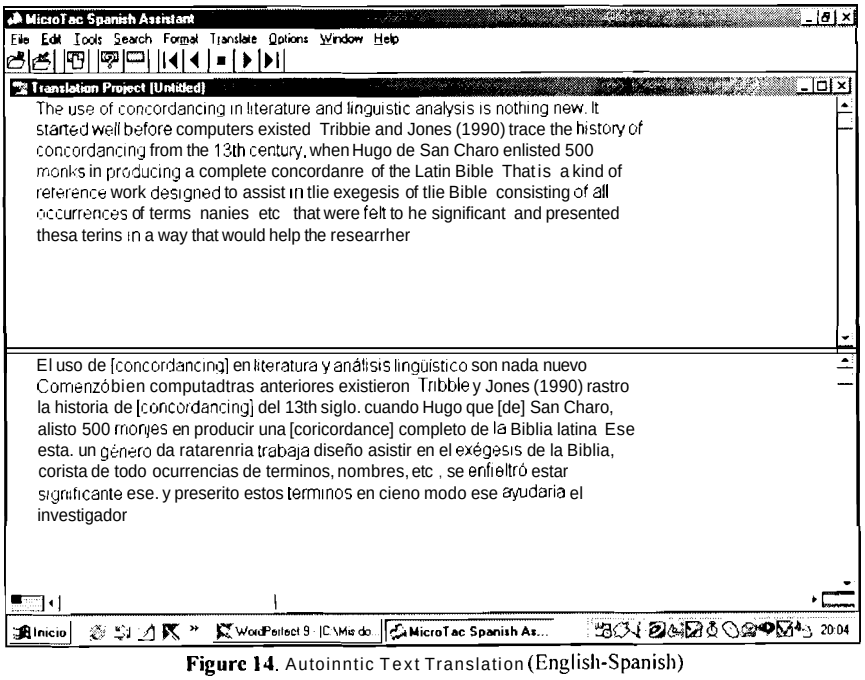


Figure 14. Autoinntic Text Translation (English-Spanish)

1.	Tlic use of coiicordanciiig in literature and linguistic analysis is nothing riew. El uso de [coiicordancing] en literatura y análisis lingüístico son nada nuevo.
2.	It staitcd well bcfore coinputers cxisted. Comenzó bien cornputadoras anteriores existieron.
3.	Tribble aiid Joiies (1990) trace tlie history ofconcordancing froni thc 13th century, when Hugo dc San Charo enlistcd 500 monks in producing a complete coricordance of the Latin Biblc. Tribble y Joiies (1990) rastro la liistoria de [concordancing] del 13th siglo, cuando Hugo que [dc] San Cliaro alistó 500 inonjcs en producir una [concordance] completo dc la Biblia latina.
4.	That is, a kiiid of refcrcnce work designed to assist iii tlic exegesis of tlie Biblc, coiisistiiig of all occurreiiccs oftcriiis. naiiies. etc., tliat were felt to be significant, and presentcd these teriiis in a way tliat would help tlie researclicr. Ése está, un género de referencia trabaja diseñó asistir en el exégesis de la Biblia. coiista dc todo ocurrciicias de térrriirios, nombres, etc.. se entieltró estar significante ése, y presciitó estos términos en cierto modo ése ayudaría el irivestigador.

Table 1. Line-by-line Printed Translation

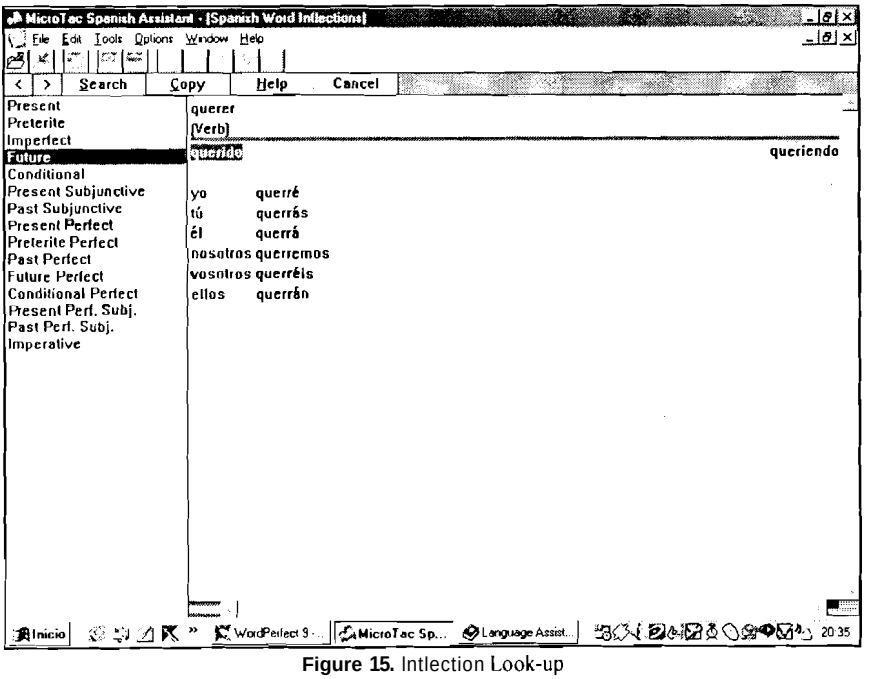


Figure 15. Intlection Look-up

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[8] 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*[9], 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@*[10]. 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*).

ID	POS	Abbreviation
0	Noun	N
1	Verb (lexical)	V
2	Verb (aux)	Aux
3	Verb (modal)	Mod
4	Adjective	Adj
5	Adverb	Adv
6	Preposition	Pre
7	Particle	Par
8	Conjunction	Con
9	Interjection	Int
10	Determiner	Det
11	Pronoun	Pro
12	Punctuation	Pun
13	Other	Oth

Table 2. Tag-set of *eTiKeT(a)*

SessionCode	FileCode	CorrectGuess	WrongGuess	LeftContext	RightContext
2	2	2	2	3	3
3	3	3	1	3	3
4	4	5	0	3	3
5	5	6	1	3	3
6	6	6	1	3	3
7	7	2187	1407	3	3
8	10	63	50	3	3

Table 3a. Information oii Sessioii Perforiiiance aiid POS-Disaiiibiguation Context selected

TagLastWord	Date	FinishedSession	JustWords	Language
0	2001-04-19 14:51:09.36	YES	NO	0
0	2001-04-19 14:52:35.21	YES	NO	0
0	2001-04-19 14:54:48.02	YES	NO	0
0	2001-04-19 14:57:17.85	YES	NO	0
0	2001-04-19 14:58:53.97	YES	NO	0
2	2001-05-10 15:25:29.61	YES	NO	0
6	2002-04-25 10:20:48.49	YES	NO	0

Table 3b. Information oii Tags , Date, Sessioii. Text or Single Word Tagging aiid Language[1]

Word	POS	Frequency	SessionCode	CorrectGuess	Language
a	10	62	0	YES	0
about	6	17	7	YES	0
above	6	1	0	YES	0
accept	1	1	7	NO	0
acceptable	4	1	7	NO	0
achieve	1	1	7	NO	0
achieved	1	1	7	NO	0
achieves	1	1	7	NO	0
achieving	1	1	7	NO	0
across	6	1	7	YES	0
actual	4	3	7	YES	0
additional	4	2	7	NO	0
address	0	1	7	NO	0
address	1	2	7	NO	0
addressed	1	1	7	NO	0
administration	0	2	7	NO	0
admiral	0	1	7	NO	0
advance	1	2	7	NO	0
affect	1	1	7	NO	0
after	5	1	7	YES	0

Table 4. Data Basc Extract (Types. Tag. Frequency. Session. First-Tiine Guessing of the Type and Language)

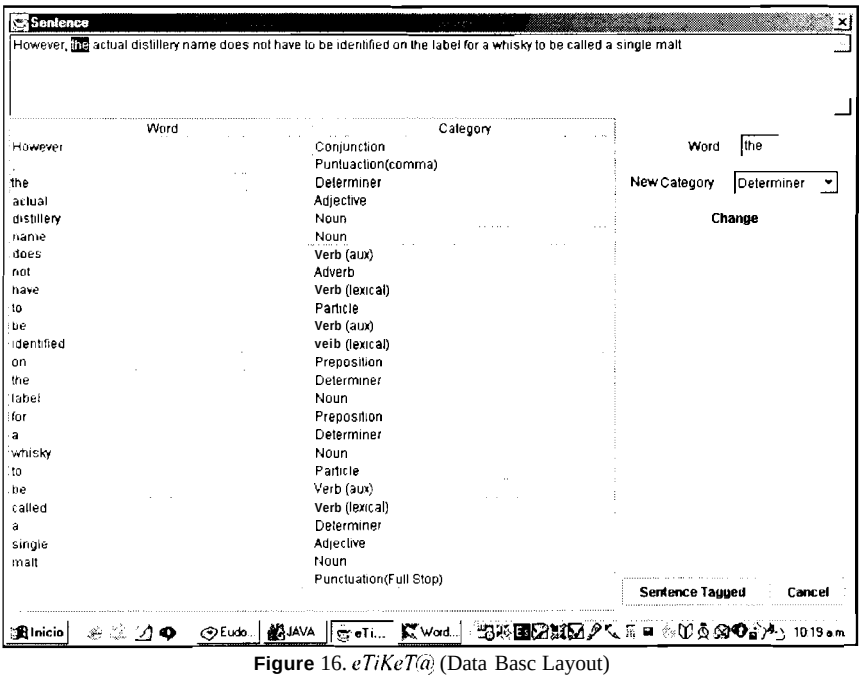


Figure 16. *eTiKeT@* (Data Basc Layout)

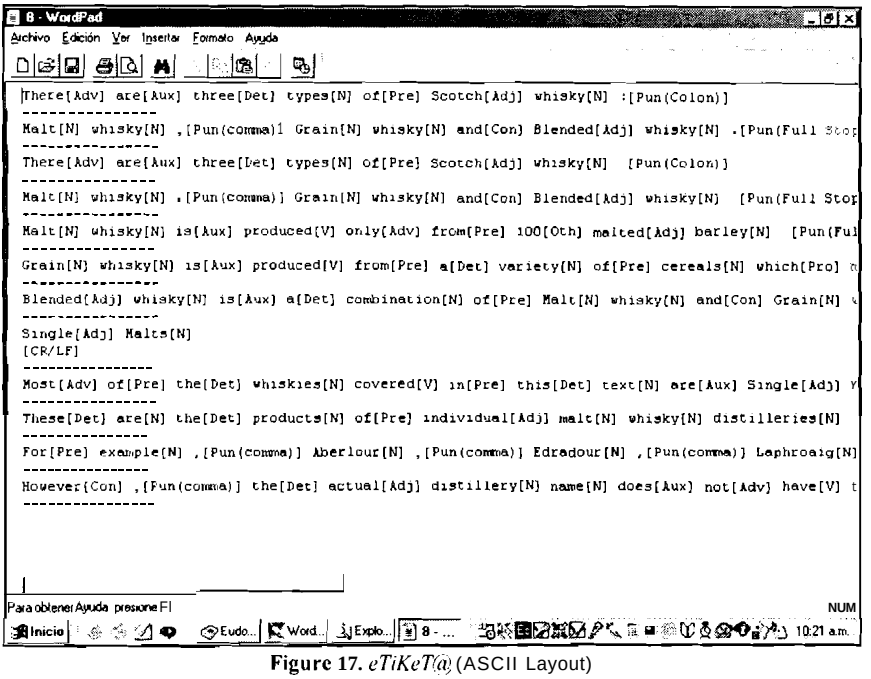


Figure 17. *eTiKeT@* (ASCII Layout)

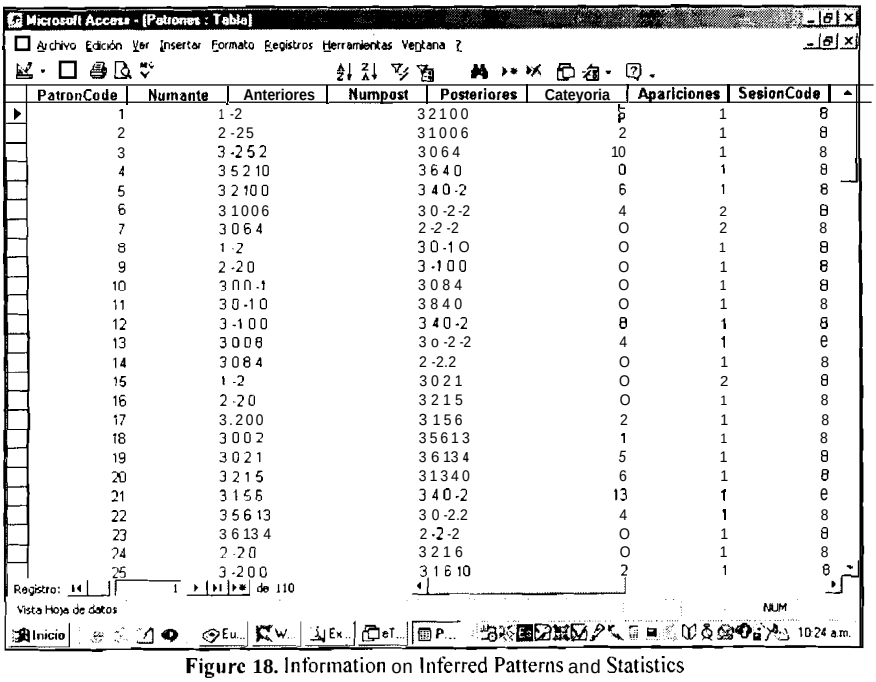


Figure 18. Information on Inferred Patterns and Statistics

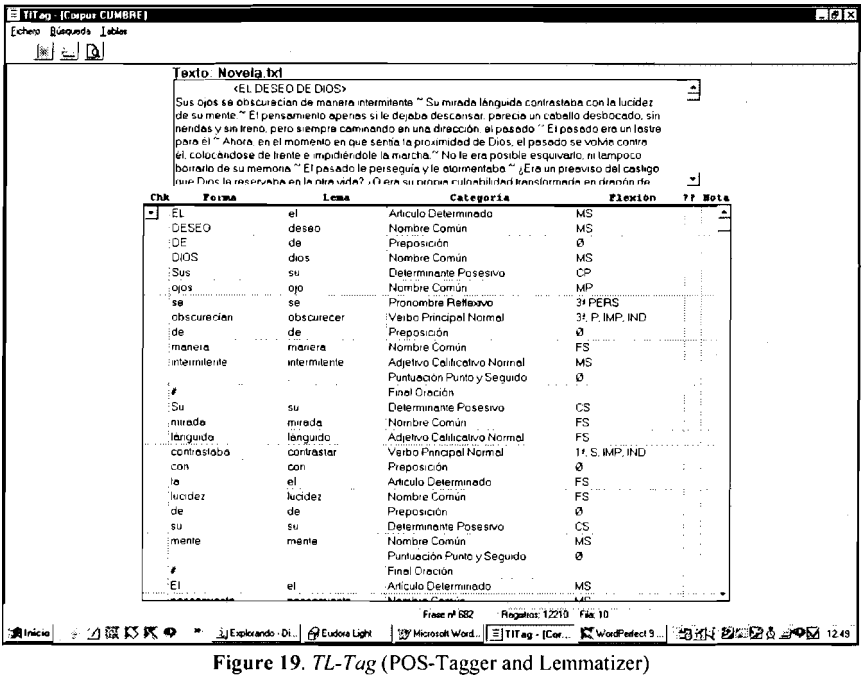


Figure 19. *TL-Tag* (POS-Tagger and Lemmatizer)

III.1.3. Lemmatisation

The distinction between words (tokens), the word forms (types) and base forms (lemmas) is important. Consider the following word sequence: *plays, playing, played, play, plays, play, 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) ɕase-bɔ-oase method to deal with irregularities, by means of rules; for example, *better* and *best* are listed and counted under the headword *good*; (2) affix stripping method; if a word form is not listed under any headword (oase-bɔ-oase method), then a number of affix stripping rules are apllied; for example, the plural suffix *-s* is taken off the word form *cars*, outputting the base form *car*. Finally, if a word form does not appear as part of the affix 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*[12] (*Figure 19*).

A useful CALL application based on the lemmatisation process is *Verbos Españoles Conjugados*[13] (*VEC*). *VEC* has been designed to assist students in the correct use and spelling of Spanish verbs. It can be used as a stand alone program or run parallel as a grammatical help tool; the student just needs to write any Spanish verb form and *VEC* feeds back with full information on tense, mood, person and number (*Figures 20* and *21*).

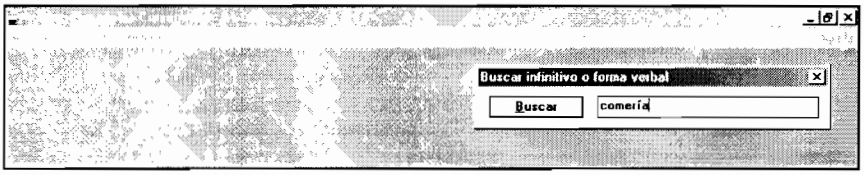


Figure 20. Student's Query

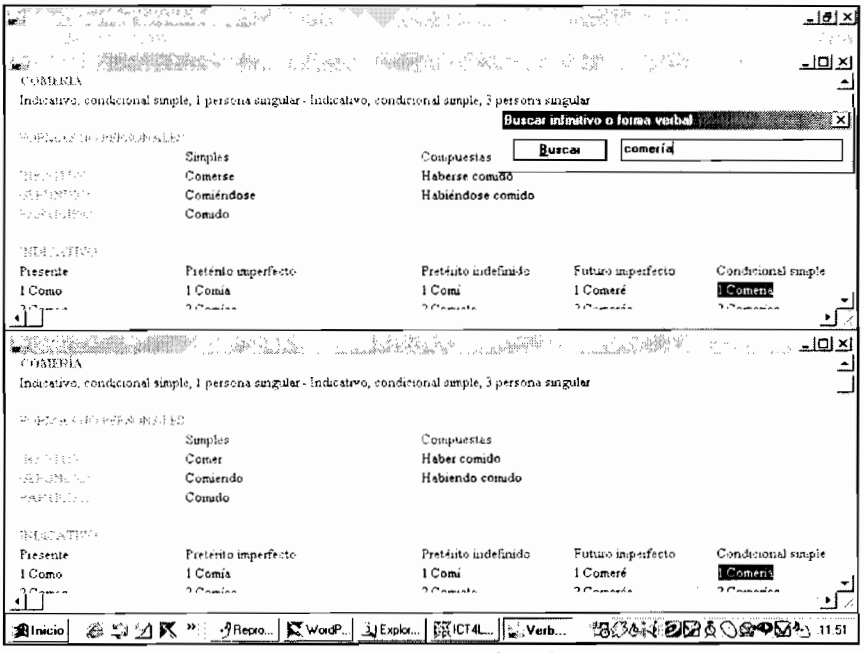


Figure 21. Full Verb Information

A much more interesting and challenging application is *Tagcorder*[14]. *Tagcorder* has been implemented to allow complex searches within the CUMBRE *Corpus* (Sánchez et al. 1995) and takcs full advantagc 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 oftypes. POS-tags and lemmas; *Figures* 25 and 26). The program itself is very interactive and flcxiblc in its search procedure and extremely fast as it works with pre-indexed text.

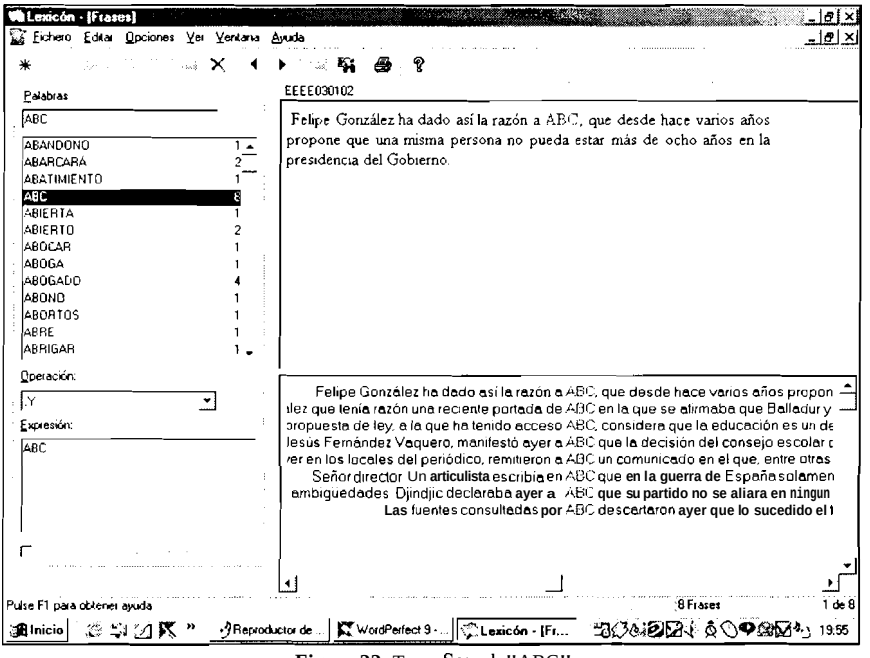


Figure 22. Type Search "ABC"

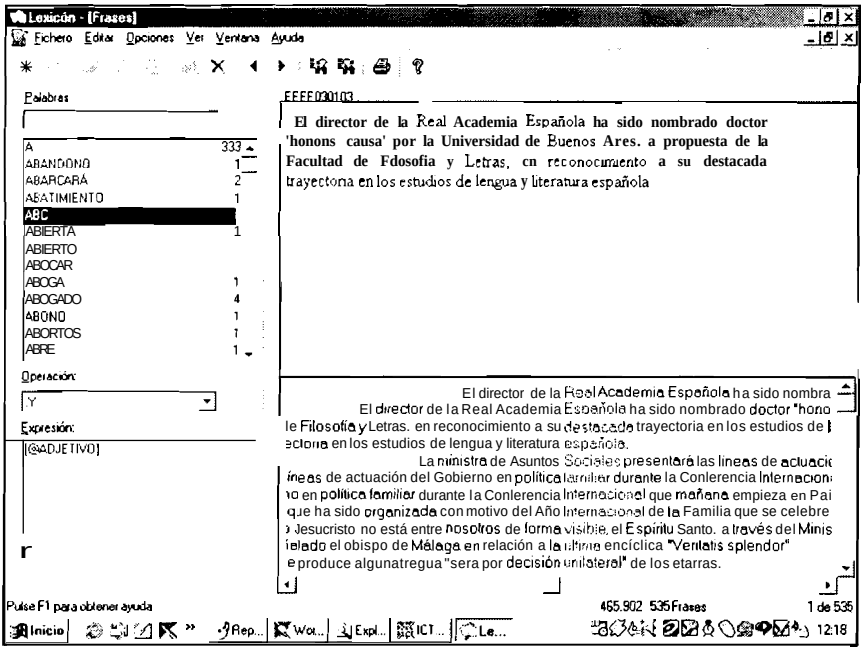


Figure 23. POS-Search "ADJECTIVE"

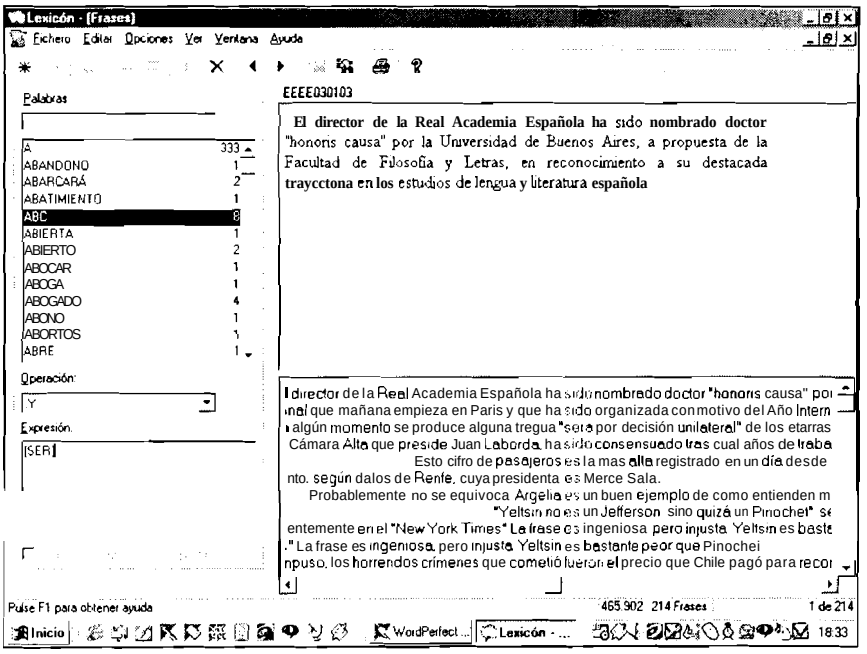


Figure 24. Lernina Search "SER"

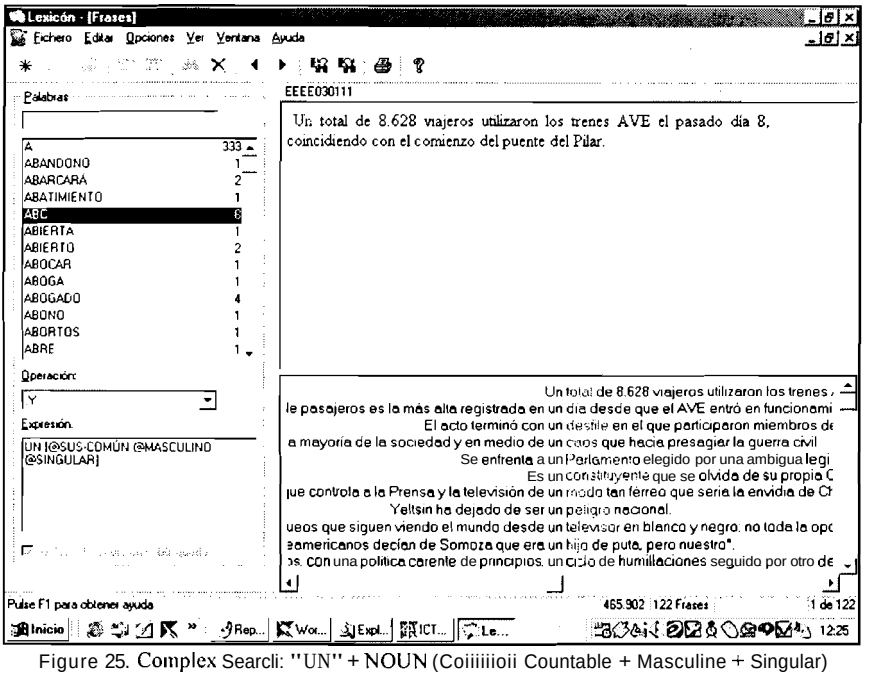


Figure 25. Complex Searcli: "UN" + NOUN (Coiiiiiioii Countable + Masculine + Singular)

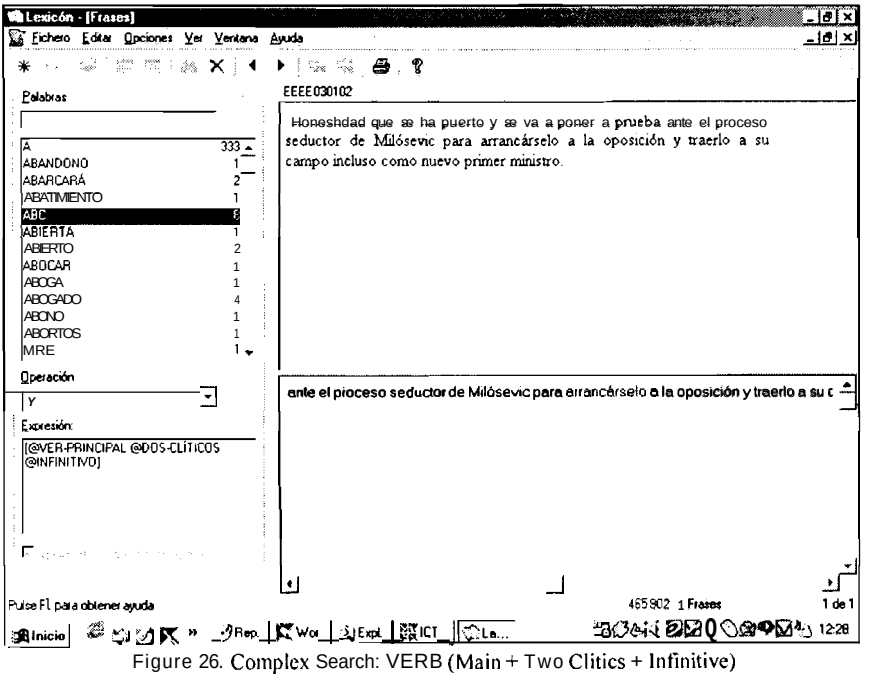


Figure 26. Complex Search: VERB (Main + Two Clitics + Infinitive)

III.1.4. Parsing

Parsing involvcs tlie procedure of bringing basic morphosyntactic categories into high-level syntactic relationships with onc anotlicr. This is probably the most commonly encountered form of corpus annotation aftcr POS tagging. Parsed corpora are sometimes known as *treebanks*.

Tlicrc arc rules governing tlie way in which words can be put together to form syntactically well-formed or grammatical sentcnces: the study of syntax aims to discover them and to describe and analyse language in terms of these rulcs. Consider the sentence ***A dog chased that girl***, where we Iind the samc pattcrn of constituents before and after the verb. that is *determiner + noun*. These two words also appear to be belong togcther more closely than say the noun *dog* and tlie vcrb *chased*. Anotlier way of illustrating that these words belong together is to givc tlie *girl* and the ***doga*** 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*.

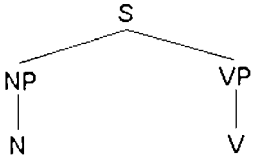
It seems clear that natural languages or human languagcs have a role of *constituent structure*. A sentence is not just a mere string of words. The words are groupcd into phrases, each of which consists of a short phrasc. Many of tlie important properties of languages are organised around constituent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

$S \rightarrow NP\ VP$

$NP \rightarrow N$

$VP \rightarrow V$

where S stands for scntcncc. *NP* for noun phrasc, *VP* for verb phrase. *N* for noun and Vfor verb. So tlie PS-rulcs abovc states that a scntence consists of a noun phrase followed by a verb phrase. In turn, the *NP* ofan *N* and the *VP* ol'a single V. The tree structure derived or generated by that rulc would bc



Parsing algorithms can proceed top-down or bottoni-up. In some cases. top-down and bottom-up algorithms can bc combined[15].

Tlic Visual Intei-active Syntax Learning (VISL) website[16] is particularly interesting and useful for language learners. It contains an on-line parser and a variety of other tools concerned with English graniniar. including ganies and quizzes. The parser itself is an excellent and very transparent application tliat allows learners to analyse and experinient on sentences and study their structure *(Figure* 27).

Interesting in this respect is also tlie parsing of students erroncous input. Integrated parscrs into CALL softwarc can be prepared to dcal with linguistic errors in the input. So the grammar tliat copes with correct sentcnces is complemented with a grammar of incorrect sentences. The advantage of this error grammar approach is that the feedback to students' ouput can be vcry specific and is normally fairly reliable as it can be attached to a very specific rule. However, tlie major drawback of this approach is tliat individual learner errors have to be anticipatcd in thc sense tliat eacli error nceds to be covered by an adequate rule.

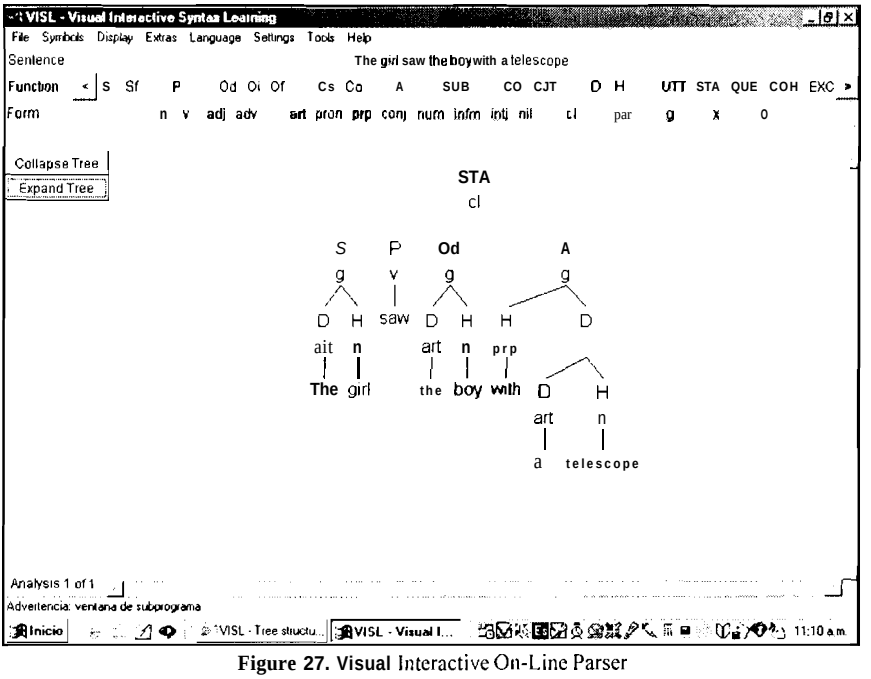


Figure 27. Visual Interactive On-Line Parser

III.1.5. Speech Technology[17]

CALL softwarc has normally bccn rcstricted 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. Furtl-ierniorc, 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 coniprcssion algorithms for this kind of data (i.e. MP3 files).

Presently, there is a widc range oí' speech software available. This includes (a) spoken input processing[18] or spccch analysis, where speech input is analysed and represented graphically or numerically; (2) spccch-i rccognition: the transformation of spoken input into written output; and (3) spccch synthesis, that is the conversion of text to speech"; this includes no just matching characters to sounds, but also intonaiion and the rhythm particular utterances have. Advances in spccch-i synthesis technology I-iave reached a high level of performance and robustness and some CALL applications have started considcring its integration[20].

In contrast, specch recognition is far more complex than speech synthcsis. Speech rccognition needs an cxtensivc analysis of spccch by means of a number of parameters, which arc vcry difficult to establish as they can be easily affected by background noise, speech speed (connected spccch). particular acccnts or idiosyncraiic individual's speech. All this leads to coiiiplicatc aiid interfere in the fixing and interprctation of the established paranictcrs.

Thcre arc somc commercial applications able to "understand" natural speech and can providc language students with realistic, highly effective, and motivating spcech practice. One of this application is *IBM® ViaVoice®*. This program runs on normal PCs and includes speaker indcpcndcni continuous specch recognition cngines and is able to deal with complete sentences spoken at a natural pace, not just isolated words. though it requircs a minor training period. To run the program, the uscr just needs to associate it with any wordprocessor, where the user's utterances will to bc iranscribed in (*Figure* 28) or run *Speech Pad*, a simplified standard wordprocessor that includes *ViaVoice*.

Many multimedia CALL courses already I-iave and still include some naive direct pronunciation practicc. That is, exercises which focus on pronunciation, fluency and word order, and with native spcakcr 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 ratc in %. The negative side of these exercises is that in some instances it is cven for naiivc speakers of the language very difficult, if not impossible, to achievc satisfactory success rates. Somc of these applications are neither very flexible nor acciiraic and sometimes students would need to repeat their utterances in several occasions before the program "understands" them corrcctly. In turn, this might lead to some small frustration.

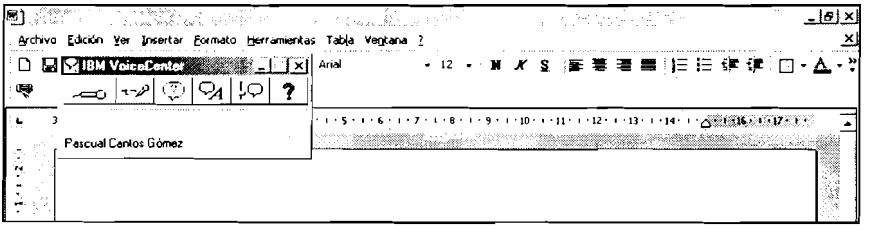


Figure 28. *IBM ViaVoice*

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, more interactive, meaning dominated, task-oriented activities (Kennedy

1998: 393).

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 conform 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[21] 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.

And since we have all these computational tools are 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

1. See among others, the Oxford Concordance Program (OUP 1988), Longman Mini Concordancer (Chandler and Tribble 1989), MicroConcord (Scott and Johns 1993), MonoConc (Barlow 1996), TACT (Bradley and Presutti 1989), or WordSmith (Scott 1999).
2. http://web.bham.ac.uk/johnstf/timconc.htm
3. *Context* can be downloaded free of charge for non-commercial purposes from Tim Johns DDL-web page: http://web.bham.ac.uk/johnstf/timconc.htm
4. Sánchez, A. and P. Cantos (3000) *Practica tu vocabulario*. Madrid: SCEL.
5. For a more complete, exhaustive survey, visit "Module 3.5. Human Language Technologies (HLT)" of the *Information and Communications Technology for Language Teacher* (ICT4LT) web page: http://www.ict4lt.org/en/en_mod3-5.htm
6. Hutchins and Somers (1992) and Arnold et al. (1994) provide excellent introductions to MT.
7. The MT system used here is *Spanish Assistant* (MicroTac Software). Other commercial MT software: *Systran* (http://babelfish.altavista.digital.com/) or *Power Translator* (http://www.lhsl.com/powertranslator/).
8. See Greene and Rubin (1971) for a detailed description of this tagger.
9. Described in detailed by Garside (1987) and Marshall (1987).
10. To get a free copy, for academic purposes only, e-mail Rafael Valencia (rafavalencia@ono.es), Rodrigo Martinez (rodrigo@dif.um.es) or Pascual Cantos (pcantos@um.es).
11. Where 0 = English and 1 = Spanish.
13. *TL-Tag* (TechnoLingua) is part of the *CUMBRE Corpus* Project and is not yet commercially available. For those interested in it contact any of the people involved in the Project: Enrique Pérez de Lema (delema@jazztree.com), Jose Simón (jsg38746@teleline.es), Aquilino Sánchez (asanchez@um.es) or Pascual Cantos (pcantos@um.es).
13. Diez, P. L. and J. Iborra (1909) *Verbos Españoles Conjugados*. Madrid: SGEL.
14. See citation 12.
15. Allen (1995), Covington (1994) and Gazdar and Mellish (1989) include excellent introductory sections on parsing algorithms.
16. http://visl.hum.ou.dk/
17. An excellent site is Integrating Speech Technology in (Language) Learning: http://www.instil.org.
18. Among the many interesting web sites, check: http://agoralang.com/signalyze.html.
19. A good example is *Winspeech*: http://www.peww.com.
20. The Polytechnic University of Hong Kong site includes a number of text-to-speech tools: http://vlc.polyu.edu.hk/TextToSpeech
21. We have deliberately not considered the integration of *Information Technologies* here. This would have inevitably expanded the potential of the "tool" proposed here.

REFERENCES

Allen, J. (1095). *Natural Language Understanding.* Redwood City, CA: Tlie Benjamin/Cumming.

Arnold, D. et al. (1994). *Machine Translation.* Oxford: NCC Blackwell.

Aston, G. (1995). Corpora in Language Pedagogy: Matching Tlieory aiid Practice. In G. Cook aiid B. Seidlhofer (Eds.), *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson.* Oxford: Oxford Uiiiversity Press, 257-270.

Astoii. G. (1996). Tlie British National Corpus as a Laiiguage Leariier Resource. In S. Botley, J. Glass, T. McEnery and A. Wilson (Eds.). *Proceedings of Teaching and Language Corpora 1996.* Lancaster: Lancaster Uiiiversity Press, 178-191.

Aston, G. (1997). Small aiid Large Corpora in Language Leariiiiig. In B. **Lewaiidowska-Toinaszczyk** aiid J. Melia (Eds.), *Proceedings of the First International Conference on Practical Applications in Language Corpora.* Lodz: Lodz Uiiiversity Press, 51-62.

Ball, C. (1995). Hwæt! A Corpus-based Hyperiiiedia Resource for Old English Lexical Acquisition. *Linguistic Society of America Annual Meeting,* New Orleans, Jaiiuary 7, 1995 (Soltware poster session).

Barlow, M. (1992). Using Concordance Software in Language Teaching aiid Research. *Second International Conference on Foreign Language Education and Technology.* Kasugai, Japan (Conference).

Barlow, M. (1995). Corpora for Theory and Practice. *International Journal of Corpus Linguistics,* 1(1), 1-38.

Bazerman, C. (1994). *Constructing Experience* Carboiidale: Southern Illinois Uiiiversity Press.

Brown P.F. el. al.(1993). Tlie Mathematics of Statistical Machine Translation: Paraineter Estimation. *Computational Linguistics,* 19(2), 263-311.

Burnard, L. (1995). *The British National Corpus Users Reference Guide.* Oxford: Oxlord Uiiiversity Computing Services.

Burnard, L. aiid T. McEnery (Eds.)(1999). *Rethinking Language Pedagogy from u Corpus Perspective.* Hamburg: Peter Lang.

Celce-Murcia, M. (1990). Data-based Language Analysis aiid TESL. In J. E. Alatis (Ed.), *Georgetown University Round Table on Language and Linguistics 1990.* Georgetowii: Georgetown Uiiiversity Press.

Cobb. T. (1997). Is there aiiy Measurable Leariiiiig from Haiids-oii Coiicordaiiciiig'?. *System,* 25(3), 301-315.

Cole, R. (1996). Foreword. In R. Cole (Ed.), *Survey of the State of the Art in Human Language Technology.* In http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Collins, H. (1999). Materials Design and Language Corpora: a Report in the Context of Distance Education. In L. Burnard and T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Hamburg: Peter Lang, 51-63.

Covington, M. (1994). *Natural Language Processing for Prolog Programmers*. Englewood Cliffs, NJ: Prentice Hall.

Flowerdew, J. (1993). Concordancing as a Tool in Course Design. *System* 21(3), 231-243

Flowerdew, J. (1996). Concordancing in Language Learning. In M. Pennington (Ed.) *The Power of CALL*. Houston: Athelstan, 97-113.

Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside et al. (Eds.), *The Computational Analysis of English. A Corpus-Based Approach*. London: Longman, 30-41.

Gavioli, L. (1997). Exploring Texts through the Concordancer: Guiding the Learner. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), *Teaching and Language Corpora*. London: Longman, 83-99.

Gazdar, G. and C. Mellish (1989). *Natural Language Processing in PROLOG. An Introduction to Computational Linguistics*. Wokingham: Addison-Wesley.

Greene, B. and G. M. Rubin (1971). *Automated Grammatical Tagging of English*. Providence Rhode Island: Brown University Press.

Higgins, J. (1988). *Language, Learners and Computers*. London: Longman.

Higgins, J. (1991a). Looking for Patterns. In T. Johns and P. King (Eds.), *Classroom Concordancing*. Birmingham: Birmingham University Press, 63-70.

Higgins, J. (1991b). Which Concordancer? A Comparative Review of MS-DOS Software. *System*, 19(1/2), 91-100.

Hutchins, J. and H. Somers (1992). *An Introduction to Machine Translation*. London: Harcourt Brace Jovanivich Publishers.

Johansson, S. (1980). The LOB Corpus of British English Texts: Presentation and Comments. *ALLC Journal*, 1, 25-36.

Johns, T. (1986). Microconcord: a Language-learner's Research Tool. *System*, 14(2), 151-162

Johns, T. (1988). Whence and Whither Classroom Concordancing?. In T. Bongaerts, P. de Haan, S. Lobbe, and H. Wekker (Eds.), *Computer Applications in Language Learning*. Dordrecht: Foris, 9-27.

Johns, T. (1991). Should you be Persuaded —Two Samples of Data-driven Learning Materials. In T. Johns and P. King (Eds.), *Classroom Concordancing*. Birmingham: Birmingham University Press, 1-13.

Johns, T. (1993). Data-driven Leariiiiig: an Update. *TELL&CALL*, 1993/2,. 4-10.

Johns, T and P. King (Eds.)(1991). *Classroom Concordancing.* Birmingham: Birmingham Uiiiversity Press.

Kacowsky, W. (1987). *Better English.* Salzburg: Jugend-Verlag.

Kaliski, T. (1992). Computer-assisted Language Leariiiiig. In P. Roach (Ed), *Computing in Linguistics and Phonetics: Introductory Readings.* Loiidoii: Academic Press, 97-106.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London: Longman.

Kettemann, B. (1995). On tlie Use of Concordancing in ELT. *TELL&CALL*, 199514, 4-15

Kučera, H. aiid N. Francis (1967). *Computational Analysis of Present Day American English.* Provideiice Rhode Island: Brown Uiiiversity Press.

Last, R. (1992). Computers and Language Leariiiiig: Past, Present — aiid tlie Future?. In C. Butler (Ed.). *Computers and Written Texts.* Oxford: Blackwells, 227-245.

Leech, G and C. N. Candlin (Eds.)(1986). *Computers in English Language Teaching and Research.* London: Longman.

Marshall, I. (1987). Tag Selection Using Probabilistic Methods. In R. Garside et al. (Eds.), *The Computational Analysis of English. A Corpus-Based Approach.* Loiidoii: Longman, 42-56.

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese aiid English by Aiialogy Principle. In A. Elithorn and R. Banerji (Eds.). *Artificial and Human Intelligence.* Ainsterdaiii: Elsevier Science Publisher, 173-180.

Sánchez, A. et al. (1995). *CUMBRE corpus lingüistico del español contemporáneo. Fundamentos, metodologia y aplicaciones.* Madrid: SGEL.

Sinclair, J. (1991). *Corpus, Concordance and Collocation.* Osford: Oxford Uiiiversity Press

Stevens, V. (1005). Concordancing with Language Learners: Wliy? When? What?. *CAELL Journal*, 6(2), 2-10.

Tribble, C. (1007). Improvising Corpora for ELT: Quick aiid Dirty Ways of Developiiig Corpora for Language Teaching. In B. Lewandowska-Tomaszczyk aiid J. Melia (Eds.), *Proceedings of the First International Conference on Practical Applications in Language Corpora.* Lodz: Lodz University Press. 106-117.

Tribble. C. and G. Jones (1090). *Concordances in the Classroom.* London: Longman.

Widdowson, H.R. (1983). New Starts and Different Kinds of Failure. In D. Larsen-Freeman et al. (Eds.), *Learning to Write: First Language/Second Language*, New York, Longman, 34-47.

