A COMPREHENSIVE DATABASE FOR SUCCESSFUL COMMUNICATION
UMA BASE DE DADOS ABRANGENTE PARA UMA COMUNICAÇÃO COM SUCESSO

ABSTRACT Considering globalization the Information and Communication Technology (ICT) becomes essential tool for providing information all over the world. The help of some online programmes is practically indispensable in certain services and areas where it is necessary to apply effective strategies to achieve specific goals in communication, overcome linguistic and cultural barriers, and avoid some misunderstandings that interfere with social interactions.

This article presents a framework of theoretical and practical aspects related to the use of electronic corpora; demonstrates how the application of monolingual and bilingual corpora involve in the contact with authentic language use that helps to improve the level of language proficiency and achieve successful communication.

RESUMO Considerando a globalização, as Tecnologias de Informação e Comunicação (TICs) tornam-se numa ferramenta essencial para prover informação a nível mundial. A ajuda de alguns programas online é praticamente indispensável em certos serviços e áreas onde é necessário aplicar estratégias eficazes para alcançar objectivos específicos em comunicação, ultrapassar barreiras linguísticas e culturais, e evitar alguns desentendimentos que interferem com as interacções sociais.

Este artigo apresenta um quadro de aspectos teóricos e práticos relacionados com a utilização de *electronic corpora*, demonstra como a aplicação bilingue ou monolingue em contacto com a linguagem verdadeira, ajuda a melhorar o nível de competência e alcançar uma comunicação bem-sucedida.
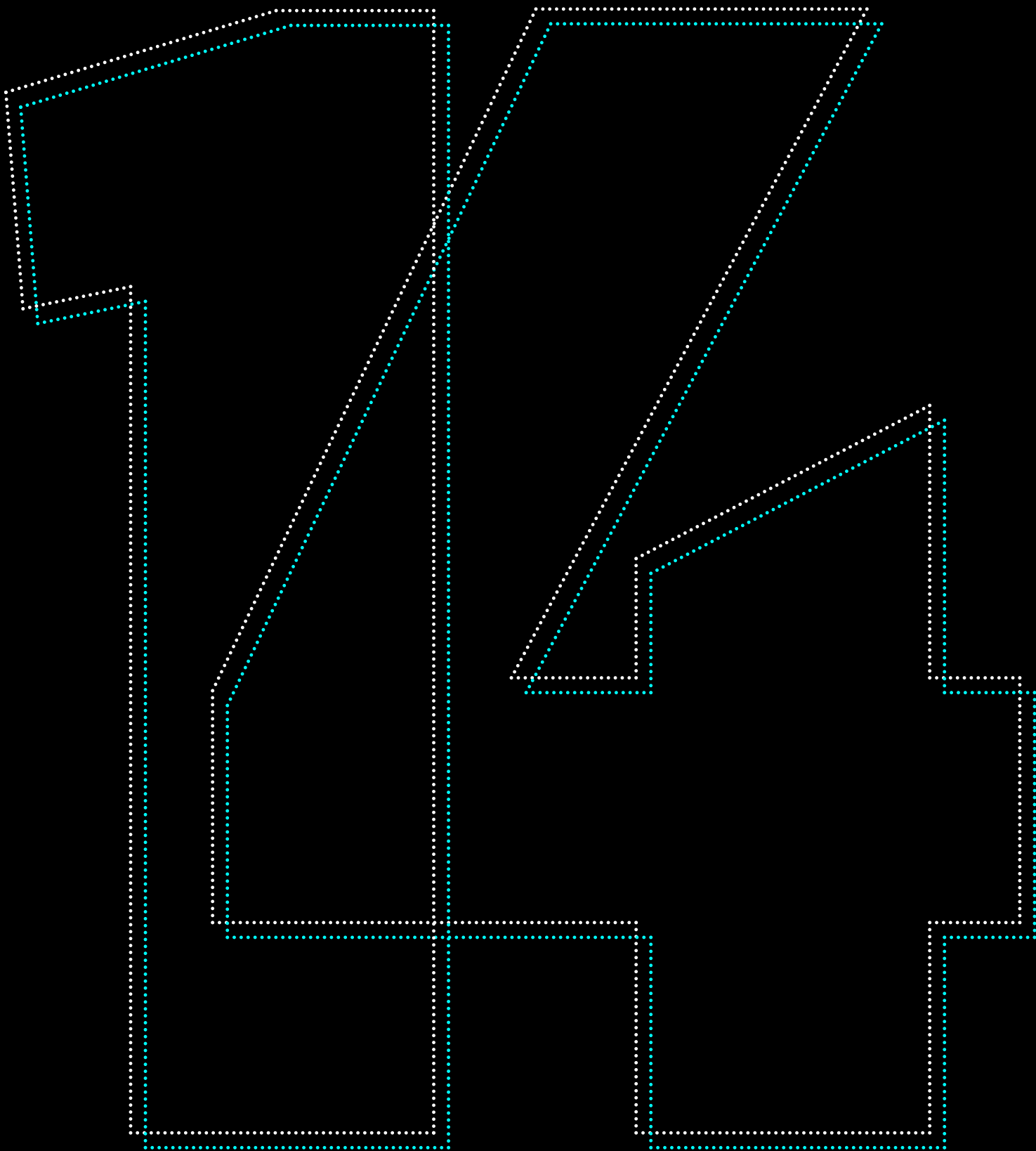
KEYWORDS
ICT, Communication, Database, Corpora

PALAVRAS-CHAVE
TICs, Comunicação, Base de dados, Corpora

NATALIYA GODINHO SOARES VIEIRA PhD researcher of the Centre for the History of Culture, Faculty of Social Sciences and Humanities (FCSH), NOVA University of Lisbon (UNL). Portugal

1.
INTRODUCTION

The term 'corpus' (plural 'corpora') had its own history before the invention of computers. In Latin it signifies (literally) 'body' (Oxford Dictionaries, 2011: information is available on line). In antiquity this term was used in some different collocations with figurative sense, for example, 'corpus delicti', the body of the offence, the essence of the crime; 'corpus juris canonis', the body of the canon law; 'corpus juris civilis', the body of the Civil law. The abovementioned Latin notions illustrate the use of the concept in the narrow sense, whereas in the broad sense 'corpus' is

1) "a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject" (Oxford Dictionaries, 2011: information is available on line). The examples are the following: the Hippocratic Corpus, the corpus of Dickens's works, the entire corpus of Shakespeare's works, etc;

2) "a collection of written or spoken language that is used for studying the language" (for example, a corpus of spoken English, etc.) (Longman Dictionary of Contemporary English, 2011: information is available on line).

The availability of a great number of electronic corpora led to more common understanding of 'corpus' as "a collection of texts, written or spoken, usually stored in a computer database" (McCarthy, 2004: 1) or "a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research" (Oxford Dictionaries, 2011: information is available on line). Analyzing these definitions we can conclude that any corpus (in the broad sense) is created for certain purposes, in most cases, for quick access to some kind of necessary information.

Nowadays different kinds of electronic corpora are available online. Electronic monolingual or bilingual corpora, in general, may contain texts in various subjects from all sorts of sources (scientific and educational literature, popular science, advertising, private correspondence, diaries etc.). In a moment internet users can find the samples almost of any kind of the existent written or spoken discourse: from a modern political debates to the instructions how to care about plants, from fairy-tails to academic works. Basically all of the advantages came from one point – electronic language corpora allow anyone to search for any word or combination of words in a particular grammatical form or certain grammatical forms. Furthermore there are some specific corpora or subcorpora that allow the search some data from a specific area (for example, the Business Letter Corpus, Corpus of the Contemporary Russian Press, etc.).

In work with online corpora it's important to identify the basic terminology such as *a word form, lemma and concordance.* "In corpus linguistics, 'word' may be used to cover the concepts of both *word form and lemma*". **Lemma** signifies "the 'dictionary form' of a word, an abstract form which encompasses all the possible forms in which the word is found in language. For example, GO is lemma relating to the word-forms *go, goes, went, gone*". **Word-form** means "each possible sequence of characters which represent a lemma. *Sleep, sleeps, slept, sleeping* are all word-forms related to lemma SLEEP". **Concordance** denotes "a list of all of the occurrences of a word or phrase in a corpus, presented in their immediate context. Often called a KWIC (Key Word In Context) concordance" (Anderson and Corbett, 2009: 46, 194, 196, 200).

There are more specific terms which are widely used in corpora database, and often they have the explanations in the introduction of corpora interfaces.

2.
ELECTRONIC CORPORA AND THEIR FUNCTIONS

The first electronic corpus named The Standard Corpus of Present-Day Edited American English or the Brown Corpus was released in 1964. It contained of over 1 million words from edited English prose printed in the United States during 1961 (Brown Corpus Manual, 1964). The Brown Corpus presented the examples from political, financial and cultural reportages, periodicals, tracts, government documents, industry reports, novels, short stories, essays, etc. That time, due to lack of development of the computer technologies, the computerized data of the Brown Corpus served, mainly, to linguistic researches; otherwise anyone could have the opportunity to observe various contexts of authentic American English of 60s.

Considering that English is a language of primary works in Corpus Linguistics (Francis, Kucera, 1964; Sinclair, 1987; Biber, Conrad, Reppen, 1998) the source of most of monolingual and bilingual corpora is a written or verbal English discourse. In spite of that fact, electronic corpora continue to be developed in other languages (Table 1).

Every electronic corpus has its own nature and design. The choice of a corpus depends on the type of information that is searched. For instance, it's evidently that there are more useful phrases for an academic discourse in some specialized academic corpora (for example, the British Academic Spoken English Corpus, the British Academic Written English Corpus, etc.) than in some other corpora where the samples are taken from fiction or other not academic sources.

TABLE 1. THE LIST OF SOME AVAILABLE ONLINE CORPORA.

THE NAME OF A CORPUS	THE SIZE AND SOURCE OF A CORPUS	WEB SITE
THE BRITISH ACADEMIC SPOKEN ENGLISH CORPUS (BASE)	"The *BASE Corpus* consists of 160 lectures and 40 seminars recorded in a variety of departments (video-recorded at the University of Warwick and audio-recorded at the University of Reading). It contains 1,644,942 tokens in total (lectures and seminars)".	http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/ http://ca.sketchengine.co.uk/bonito/run.cgi/first_form?corpname=preloaded/base;
THE BRITISH ACADEMIC WRITTEN ENGLISH CORPUS (BAWE)	"The BAWE corpus contains 2761 pieces of proficient assessed student writing, ranging in length from about 500 words to about 5000 words… Holdings are fairly evenly distributed across four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) and across four levels of study (undergraduate and taught masters level)".	http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/ http://ca.sketchengine.co.uk/bonito/run.cgi/first_form?corpname=preloaded/bawe2

ENGLISH LANGUAGE INTERVIEW CORPUS AS A SECOND-LANGUAGE APPLICATION (ELISA)	"The corpus is a collection of video-based interviews with native speakers of different varieties of English (e.g. US, England, Scotland, Ireland, Australia) and from different walks of life. The corpus currently contains 25 interviews of 5 to 15 minutes, the transcripts amount to about 60,000 words in total".	http://www.uni-tuebingen.de/elisa/html/elisa_index.html
BRITISH NATIONAL CORPUS (BNC)	"100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written".	http://www.natcorp.ox.ac.uk/
CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)	"...more than 425 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts".	http://corpus.byu.edu/coca/
BUSINESS LETTER CORPUS (BLC)	"1 million specified words or phrases from business letters".	http://www.someya-net.com/concordancer/
MICHIGAN CORPUS OF ACADEMIC SPOKEN ENGLISH (MICASE)	"...nearly 1.8 million words of transcribed speech (almost 200 hours of recordings) from the University of Michigan".	http://micase.elicorpora.info/
TIME MAGAZINE CORPUS OF AMERICAN ENGLISH	"...more than 100 million words of text of American English from 1923 to the present, as found in TIME magazine".	http://corpus.byu.edu/time/
RUSSIAN NATIONAL CORPUS (RNC) (it has Russian and English surfaces)	"...over 150 million words... includes primarily original prose representing standard Russian (from the middle of the 18th century) but also, albeit in smaller volumes, translated works (parallel with the original texts) and poetry, as well as texts, representing the non-standard forms of modern Russian: spoken (recordings of oral speech, spontaneous and public) and dialectal".	http://www.ruscorpora.ru/en/index.html
SINICA CORPUS OF MODERN CHINESE	"...the size of this corpus is 5 million words... This corpus is a representative sample of modern Chinese language".	http://db1x.sinica.edu.tw/kiwi/mkiwi/
CORPUS DEL ESPAÑOL (SPANISH CORPUS)	"...over 100,000,000 words from more than 20,000 Spanish texts of the XIII – XX centuries...".[1]	http://www.corpusdelespanol.org/
COMPARA (PARALLEL CORPUS OF ENGLISH AND PORTUGUESE)	"...Portuguese-English parallel corpus..., totalling around three million words... COMPARA is made up of 75 text pairs. These texts are published literary source-text and translation extracts from Angola, Brazil, Mozambique, Portugal, South Africa, the United Kingdom and the United States".	http://www.linguateca.pt/COMPARA/psimples.php
THE POLYU LANGUAGE BANK (a bank of corpora of English, Chinese, French, German, Italian, Japanese, Spanish)	...is a large archive of language corpora made up of a wide range of written and spoken texts totalling over 12 million words".	http://langbank.engl.polyu.edu.hk/indexl.html

3.
INTERPER–SONAL DISCOURSE WITH ONLINE CORPORA

People build relationships through communication. The formation of a verbal or written discourse depends on numerous factors. Information exchange requires the properly chosen lexical units – it helps to avoid misun-

derstandings and achieve the final objectives of communication. The use of corpora opens the possibility to carry out the necessary concordances to check for errors, determine the frequency of use of some word forms or to find their meanings, etc.

For example, many non-native speakers struggle with prepositions, because, in the most cases, different languages have dissimilar linguistic forms. Corpora may help to narrow the possibility for errors. In a few seconds corpora present the variety of concordances with the necessary combinations.

Let's try to search the English phrasal verbs 'stay at + noun' and 'stay in + noun' in the Corpus of Contemporary American English (COCA). According to the Corpus instructions (see the question marks) we have to type *stay at [n*]* in a search string, activate a KWIC display, sort by frequency, relevance or alphabetically, etc. (a choice of the options depends on purposes of the search).

TABLE 2. COCA CONCORDANCE OF *'STAY AT + NOUN'*, SORTED BY THE RELEVANCE.

TABLE 3. COCA CONCORDANCE OF *'STAY IN + NOUN'*, SORTED BY THE RELEVANCE.

CLICK FOR MORE CONTEXT						SAVE LIST	CHOOSE LIST	CREATE NEW LIST	[?]			
1	1990	SPOK	ABC_Nightline	A	B	C	. But it immediately passed a resolution authorizing troops to	stay in Lebanon				
2	1995	MAG	GolfMag	A	B	C	tempt Woods to break his promise to himself that he will	stay in school				
3	1992	MAG	SportsIII	A	B	C	She went to some high school football games , jogged to	stay in shape				
4	2007	MAG	Prevention	A	B	C	" It 's a fun activity that helps all of us	stay in shape				
5	1993	MAG	PsychToday	A	B	C	and dropped him a note ; " I 'd like to	stay in touch				
6	1994	SPOK	PBS_Newshour	A	B	C	. Why did you accept an agreement that allows them to	stay in Haiti				
7	2009	ACAD	GeorgiaHisQ	A	B	C	politicians , all of whom Haygood said needed whiskey money to	stay in business				
8	2009	NEWS	Houston	A	B	C	poor listening skills . So , it is respectfully suggested Joe	stay in D.C.				
9	1992	SPOK	CBS_Sixty	A	B	C	at the Comedy and Magic Club in Hermosa Beach just to	stay in practice				
10	2000	MAG	SportingNews	A	B	C	think that 's one of the reasons we 're able to	stay in games				
11	2005	SPOK	CNN_SunNight	A	B	C	be able to have some money . but they ca n't	stay in hotels				
12	2007	MAG	NatGeog	A	B	C	fallen for her . In any case , she agreed to	stay in Jamestown				
13	2010	SPOK	ABC_ThisWeek	A	B	C	is worth it because - it is now worth it to	stay in Afghanistan				
14	2000	NEWS	Atlanta	A	B	C	the lodge itself , which is just common areas . They	stay in log				
15	2006	NEWS	WashingtonPost	A	B	C	is personal is we believe there should be requirements that kids	stay in school				

So the COCA selected 969 tokens with a phrasal verb *'stay at + noun'* (Table 2) and 4,175 tokens with a phrasal verb `*stay in + noun*` (Table 3). From the contexts it's simple to recognize the difference between the use of the prepositions *at* and *in*. As we see 'stay at + noun' means *take board or lodging (at a place), stop at* ... And 'stay in + noun' signifies *remain in position, be in* ... As it was indicated, corpora help to clarify the use of various lexical and grammatical structures.

It's naturally that non-native speakers have problems with a choice of the gender of nouns. Communicators may be trapped in their guesses, because some languages have three genders: masculine, feminine and neuter (eg., Russian, German, etc.) and other some languages have only two genders: masculine and feminine (eg., Portuguese, Italian, etc.). The following example is going to illustrate how corpora help in making the right choice of the gender of nouns. For this purpose we have to observe the semantic environment of nouns in different contexts, in other words, to browse the concordance in a corpus.

For instance, if we don't know exactly what gender has a Russian word 'ДОМ' (house), let's open the Russian National Corpus (RNC) and type 'ДОМ' in the main corpus, selecting a part of speech (noun), a case (nominative), a number (singular) and click on the option 'search'. On another page we have to choose a KWIC format. It helps to visualize the concordance.

TABLE 4. RNC CONCORDANCE OF '*дом*' (LEXICO-GRAMMATICAL SEARCH) IN THE KWIC FORMAT.

Всё-таки мой дом явно проклят За неполные 2 ←...→
Добротный деревянный дом , рядом с крыльцом — берёза, а ←...→
Это и есть дом , где Андрей Тарковский провёл довоенное ←...→
Детство...", — Тот самый дом Фотоавтора ←...→
нции метро "Третьяковская", Черниговский пер., дом 9/ 13, стр. 2. ←...→
Просто потому, что дом — отдельностоящий, а мышей я не ←...→
раньше жили на 8 Марта, д . 15, поэтому в 227 и ←...→
раньше жили на 8го марта д 2/ 10. ←...→
фотографией занялась — то весь был дом реактивами пропах, плёнки с потолка ←...→
обрести спокойствие и безмятежность: уютный дом , крепкая семья, финансовая стабильность ←...→
Твой дом был мне лымгнёздышком, куда хотелось ←...→
на себя бытовые заботы, чтобы дом опять превратился из временного приста нища ←...→
Заинтересован Белый дом в подключении к этой инициативе ←...→

If to analyze the individual segments 'adjective + noun' or 'pronoun + noun' (in the given contexts), it will be obviously that a Russian word 'ДОМ' (house) has a masculine gender. Similar observations help to avoid cross linguistic errors, especially, when the nouns haven't the same genders in different languages (for example, in Portuguese 'casa' (house) has a feminine gender).

4.
BUSINESS DISCOURSE WITH ONLINE CORPORA

Any type of communication in the business field is required the use of special vocabulary. Managers, secretaries, administrators, directors' assistants, etc. have to respond to a large number of business emails per day, prepare business posters and presentations, to be able to negotiate,

etc. In this case, sometimes, it's necessary to apply the knowledge of foreign languages, especially English.

The language of the business letters is specific. A writer or a speaker has to be concrete and care about readers and listeners to avoid misunderstandings. From time to time in the business sphere some verbal or written errors lead to serious consequences. For this reason it's important to have quick access to the necessary database that helps to correct errors. There are some corpora that provide the information for a business discourse.

Let's see some examples. Someone has to send a business apology letter to extend an apology in a professional business environment. He/she supposes to use some words related to the subject, but he/she doesn't know exactly how to formulate a complete sentence. In this case he/she has to practice with the Business Letter Corpus (BLC).

First of all, a business apology letter requires to ask sorry for something, probably, for 'the inconvenience'. Secondly, it's necessary to type a key-word 'inconvenience' in a search string of the BLC, thirdly, to chose a position of the word in the whole phrase (*search type:* equal to, start with, end with or contain) and a sort of the concordance lines (*a sort type:* right, left, unsorted). As a result 169 possible combinations (number of hits) with a key-word 'inconvenience' will appear (Table 5).

TABLE 5. BLC CONCORDANCE OF *INCONVENIENCE*, SORTED BY THE FIRST WORD TO THE RIGHT OF THE NODE.

Online BLC KWIC Concordancer Search Result

Search String: equal to "inconvenience"
Search Corpus: 01. Business Letter Corpus (BLC2000)
No. of Hit(s): 169

```
1      [BLC2:06:02072] I apologize for the inconvenience and aggravation we caused.
2      [BLC2:23:01501] To avoid further inconvenience and annoyance to customers, as well as
3    BLC2:36:01056] We regret very much this inconvenience and ask you to accept a slight delay in
4    2:27:00336] We apologise to you for any inconvenience and ask you to accept the enclosed ( PF
5    may show a film at any time at no great inconvenience and be assured that the sound and visua
6    urther delay and thus save yourself the inconvenience and considerable costs of a legal actic
7    everything we possibly can to eliminate inconvenience and discomfort for our friends.
8    00109] Please accept my apology for the inconvenience and distress this problem has caused yc
9    at you will accept my apologies for the inconvenience and frustration you experienced in tryi
10   1448] A car pool would help to ease the inconvenience and save money.
```

So, it's possible to receive one hundred sixty nine suitable variants that help in formulating of the beginning of a business apology letter.

There is one more example. He/she has to select some typically patterns for a business conversation. Certainly, it is impossible to preview what will be said, but observations are always relevant, especially for study of the correct use of some expressions that can be useful in the business sphere. If he/she knows the objectives and the main topic of a future business debate, it's enough to type some principal key-words in a specific business corpus. If to implement the same procedure as it's shown in the first example, it will be possible to receive the necessary patterns.

5.
TRANSLATION /INTERPRE- TATION WITH ONLINE CORPORA

Translators and interpreters have to be familiar with the vocabulary of the field that requires their service. The process of transmission of the information from one language to another is a sort of the interpretation of this information. That's why it's very important to know the correct equivalents in a target language to not distort the original meanings. Translators and interpreters regularly make an effort to discover some lexical items, especially when they face with terminology, for instance,

medical, economic, legal, linguistic, military, etc. The translation has to be confirmed by the authoritative sources. In this case, bilingual corpora give a wide picture of translation equivalents taken from different types of a discourse. Parallel concordances offer relevant and contextualized equivalents in two languages at the same time. Evidently, many translators and interpreters recognize the significance of bilingual corpora.

For example, we face with the task of translating the following context of a military discourse from Portuguese into English – *"...a necessidade mundial de um sistema de reconhecimento aéreo, vigilância e patrulha armada de baixo custo capaz de suportar uma ampla gama de operações"* (Aeronave Revolucionária anuncia o renascimento da indústria aeroespacial da África, 2011: information is available on line). What are the English equivalents for the Portuguese terms 'reconhecimento aéreo' and 'vigilância'? The bilingual corpus COMPARA answers this question.

At first it is necessary to choose the Portuguese-English translation. After then to type only one word *'reconhecimento'* in a simple search option and click to search. Now we are ready to observe the concordance (Table 6). There are twenty five occurrences.

In a moment, in the comprehensible contexts, it's possible to see the differences between 'recognition' and 'reconnaissance'. Obviously that in the abovementioned military discourse the more appropriate English translation equivalent for 'reconhecimento aéreo' is 'aerial reconnaissance'.

The same search is applicable for another Portuguese term 'vigilância'. According to COMPARA it has to be translated as 'vigilance' or 'surveillance' (see COMPARA). The careful translator/interpreter will deduce to use 'surveillance' in the context of military discourse.

To sum up, bilingual corpora confirm the hypothesis about the use of translation equivalents in different types of a discourse; moreover, eliminate the long search for the necessary information.

TABLE 6. COMPARA PARALLEL CONCORDANCE OF *RECONHECIMENTO*.

EBDL1T1(1362):	No entanto, senti uma ligeira tremura perante o **reconhecimento** da palavra *angústia* .	But I felt a little shiver of recognition at the word «dread».
EBDL1T2(1362):	Mas senti um tremor de **reconhecimento** na palavra «angústia».	But I felt a little shiver of recognition at the word «dread».
EBDL4(902):	No cais, em Limassol, mesmo antes da partida, havia um camião francês para três camiões ingleses e, como os veículos ingleses tinham sido pintados a aerossol, com a camuflagem amarela do deserto e os franceses não, essa disposição revelava claramente às operações de **reconhecimento** aéreo a presença dos camiões.	On the quayside at Limassol, prior to embarkation, one French truck was grouped with every three British trucks, and as the British vehicles had been sprayed with yellow desert camouflage and the French ones had not, this pattern gave away the presence of the trucks to aerial reconnaissance as clearly as could be.

6.
CONCLUSIONS

The development of the ICT helps to improve a variety of human skills and to learn the new ways to search information. Electronic corpora present a huge database not only for linguists and researchers but for everyone who is interested in improving the quality of a verbal or written communication. For successful use of corpora for the purposes of spoken interaction or written production it's enough to follow some basic steps: (1) to identify the goals of communication; (2) to discover a corpus that is more consistent with the main objectives; (3) to retrieve the necessary information from this corpus. So the use of electronic corpora gives the opportunity to resolve some communicative problems at the professional level and in a short period of time.

BIBLIOGRAPHY

Aeronave Revolucionária anuncia o renascimento da indústria aeroespacial da África (2011) *Press Release.* Paramount Group. Available at <http://www.paramountgroup.biz/pt/AHRLAC_press_release.html > (last access 7/11/2011).

Anderson, W., Corbett, J. (2009) *Exploring English with Online Corpora.* Basingstoke: Palgrave MacMillan.

Biber, D., Conrad, S., Reppen, R. (1998) *Corpus linguistics – Investigating language structure and use.* Cambridge: Cambridge University Press.

Brown Corpus Manual. Manual of information to accompany a standard Corpus of Present-Day Edited American English, for use with Digital Computers / Francis, W., Kucera, H. (eds.) (1964) Available at <http://khnt.aksis.uib.no/icame/manuals/brown/INDEX.HTM#bc1>> (last access 11/10/2011).

McCarthy, M. (2004) T*ouchstone: from corpus to course book.* Cambridge: Cambridge University Press. Retrieved from <http://www.cambridge.org/elt/teacher-support/pdf/McCarthy-Touchstone-Corpus.pdf> (last access 11/ 10/ 2011).

Sinclair , J. (eds.) *Collins (1987) Collins COBUILD English Language Dictionary* [Collins Birmingham University International Language Database]. London: Collins.

DICTIONARIES ONLINE:

Oxford Dictionaries. Available at <http://oxforddictionaries.com/> (last access 11/ 10/ 2011).

Longman Dictionary of Contemporary English. Available at <http://www.ldoceonline.com/> (last access 11/ 10/ 2011).