2 Corpora

2.1 What is a corpus?

A corpus (pl. corpora) is a large collection of texts electronically stored on a computer. These texts contain authentic language used in real situations and can represent both the language used in speech and in writing.
A corpus can be used for a number of reasons. Some of them are listed here:

- to check patterns of the language and its lexico-grammatical features;
- to check the use of words;
- to compare the use of words in different varieties of the same language (for example either in the language of economics or in the language of medicine and so on …);
- to compare and contrast translation equivalents across different languages;
- to draw examples for the preparation of teaching material;
- to obtain a list of the phraseology and the terminology of a language and its varieties;
- etc. etc.… .

Technology has made the procedure of corpus assembling easier and easier both because of the large number of texts available on t he internet and of the availability of more powerful computers which can store huge amounts of bytes without slowing down the proper working of a computer.
General English corpora are made up of millions of words in order to be representative of the whole English language. Two examples are:

- The British National Corpus (BNC), a 100 million word corpus of modern British English texts, both written and spoken, made available in 1995;
- The Bank of English (BoE), a 450 million word monitor[8] corpus under continuous development at the university of Birmingham since 1980[9].

Corpora may be of different types. These types depend on the type of language we are going to analyse. As seen in the previous chapter, language is dynamic and acquires different features depending on the social event in which it is used. A list of the main categories of corpora is provided below (adapted from Bowker and Pearson 2001:11-12 and from Teubert 1996:245-246; 2000):

General reference corpus: a general reference corpus is designed to be representative of a given language as a whole and can therefore be used to obtain insights on that language. It is usually constituted of a series of text types (both spoken and written) and focuses on the language used by ordinary people in everyday situations (newspaper, fiction, radio and television broadcasts, etc.);
Special purpose corpus: a corpus which focuses on a particular aspect of language, that is to say on a particular subject field, text type or language variety. A special purpose corpus may be a corpus constituted of tourist websites, or of articles from sports newspapers. The insights we obtain from this type of corpus are only valid for the type of language contained in it.
Monolingual corpus: a monolingual corpus contains texts in only one language.
Multilingual corpus: a multilingual corpus contains texts in two or more languages and can be *comparable* or *parallel*. A multilingual corpus is comparable when it is constituted by two or more sets of texts which have similar composition. Similar composition means that all the texts contained in the corpus have the same communicative function, they all deal with the same topic, they are all of the same type of text. The texts contained in a comparable corpus are all

[8] A monitor corpus is a type of corpus which is constantly updated to monitor changes in a given language.

[9] For details of the BNC visit http:// info.ox.ac.uk/bnc Information on the BoE is available at the Collind COBUILD website, http://titania.cobuild.collins.co.uk/boe_info.html

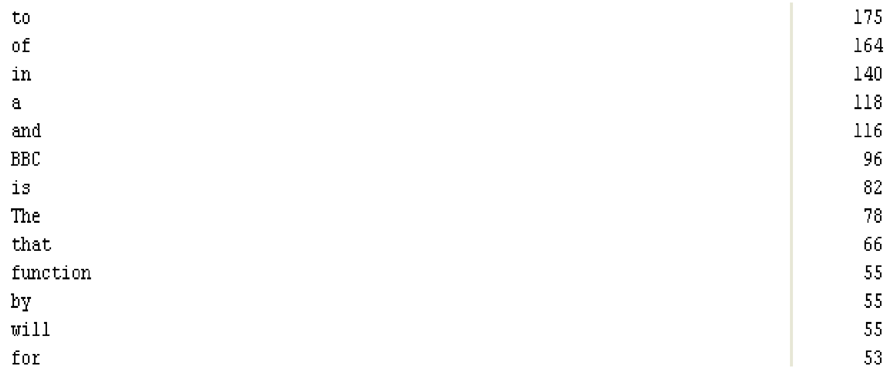
original texts and no translations are included. For example, a comparable corpus may be constituted of British newspaper articles on the economic recession and Italian newspaper articles on the same topic; similarly, it can be composed of three subcorpora collecting the speeches delivered by Barack Obama, Gordon Brown and Mario Monti. Analysing the original texts contained in these corpora we can make observations on the features (lexical, syntactic, ...) of these languages and we can compare them in order to detect differences and similarities.

A multilingual corpus is defined parallel when it contains original texts in one language and their translations in another language. For example, a parallel corpus may be constituted by original texts of fairy tales in English and their translations into Italian (French, German, ...). Parallel corpora can provide examples of how equivalence has been established by translators and what translation strategies have been adopted at different stages. They can be identified as a big repository containing previous translators' choices (Teubert 2000).

In order to carry out a corpus-based analysis, first a corpus to analyse is needed. There are different corpora that can be used (many of them are available on the internet). Students taking their first steps in corpus analysis should use small corpora (about 80,000/100,000 words or even less depending on the purpose of the analysis) which contain texts dealing with a specific topic. Johansson (1991: 305-6) suggests that in spite of the undoubted advantages of large corpora, there is still something to be said "for smaller, carefully constructed sample corpora which can be analysed exhaustively in a variety of ways". Of course, the results of an analysis based on a specific topic will only be valid for that specific language domain and not for the language as a whole.

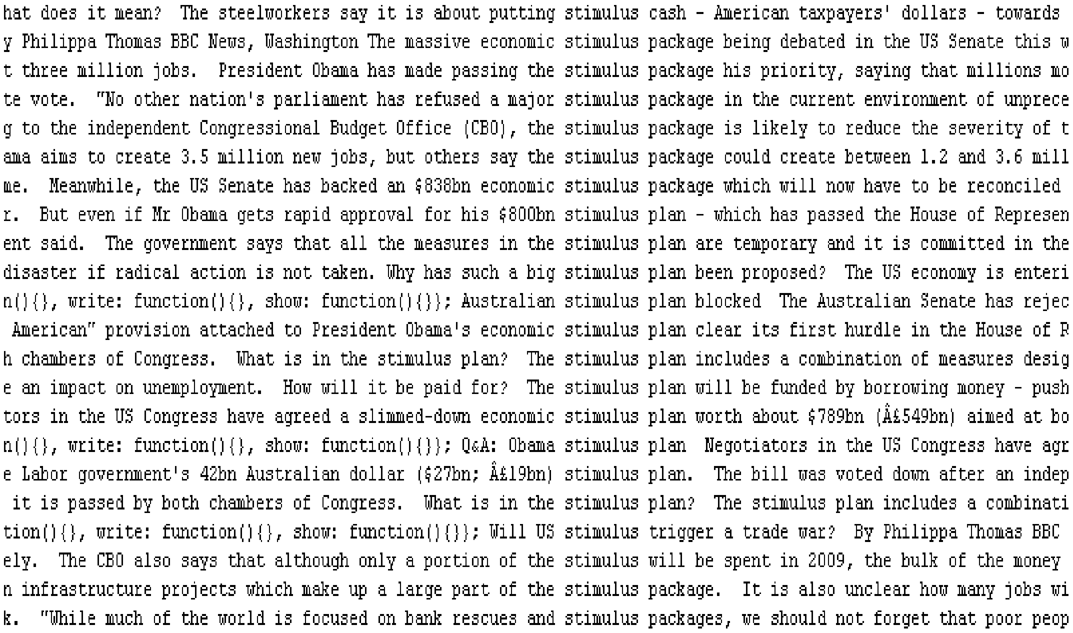
Among the most practical applications of corpora, particularly of special purpose corpora, there is the elaboration of wordlists and patterns of the language of a given topic. We may be requested, for example, to write or to talk about the "economic recession" but we may not be familiar with the terminology and the phraseology associated with it. We can, therefore, search the net to identify and download a number of newspaper articles which deal with global recession. A software will help us to analyse these texts and create useful *wordlists* and *concordances*.

Wordlists are a list of all the words contained in the texts chosen for analysis. These words are listed in frequency order or in alphabetical order by the software, as visible in the figure below:



to	175
of	164
in	140
a	118
and	116
BBC	96
is	82
The	78
that	66
function	55
by	55
will	55
for	53

Concordances allow the researcher to identify and analyse the linguistic co-text of a word. The format data will acquire is called KWIC which stands for Key Word In Context. The *node word* is aligned in the centre and is preceded and followed by its co-text. The words frequently preceding and following the node word in a span of five words on the left and five words on the right are called *collocates* and form with the node words repeated string of words called *patterns*.



```
hat does it mean?  The steelworkers say it is about putting stimulus cash - American taxpayers' dollars - towards
y Philippa Thomas BBC News, Washington The massive economic stimulus package being debated in the US Senate this w
t three million jobs.  President Obama has made passing the stimulus package his priority, saying that millions mo
te vote.  "No other nation's parliament has refused a major stimulus package in the current environment of unprece
g to the independent Congressional Budget Office (CBO), the stimulus package is likely to reduce the severity of t
ama aims to create 3.5 million new jobs, but others say the stimulus package could create between 1.2 and 3.6 mill
me.  Meanwhile, the US Senate has backed an $838bn economic stimulus package which will now have to be reconciled
r.  But even if Mr Obama gets rapid approval for his $800bn stimulus plan - which has passed the House of Represen
ent said.  The government says that all the measures in the stimulus plan are temporary and it is committed in the
disaster if radical action is not taken. Why has such a big stimulus plan been proposed?  The US economy is enteri
n(){}, write: function(){}, show: function(){}}; Australian stimulus plan blocked  The Australian Senate has rejec
 American" provision attached to President Obama's economic stimulus plan clear its first hurdle in the House of R
h chambers of Congress.  What is in the stimulus plan?  The stimulus plan includes a combination of measures desig
e an impact on unemployment.  How will it be paid for?  The stimulus plan will be funded by borrowing money - push
tors in the US Congress have agreed a slimmed-down economic stimulus plan worth about $789bn (Â£549bn) aimed at bo
n(){}, write: function(){}, show: function(){}}; Q&A: Obama stimulus plan  Negotiators in the US Congress have agr
e Labor government's 42bn Australian dollar ($27bn; Â£19bn) stimulus plan.  The bill was voted down after an indep
 it is passed by both chambers of Congress.  What is in the stimulus plan?  The stimulus plan includes a combinati
tion(){}, write: function(){}, show: function(){}}; Will US stimulus trigger a trade war?  By Philippa Thomas BBC
ely.  The CBO also says that although only a portion of the stimulus will be spent in 2009, the bulk of the money
n infrastructure projects which make up a large part of the stimulus package.  It is also unclear how many jobs wi
k.  "While much of the world is focused on bank rescues and stimulus packages, we should not forget that poor peop
```

As visible in the figure reported above, if we choose to analyse the word *stimulus* which is one of the most frequent in the language of "economic recession"[10], its collocates are immediately visible and we are provided with a p attern such as "*pass/refuse/debate + economic stimulus package*", which we can add to our database on global recession.

Collocates and patterns of language will be dealt extensively in chapters 3 and 4.

In order to obtain *wordlists* and *concordances,* we need a s oftware that can be used for corpus analysis. Some of them are listed below[11]:

- **WordSmith Tools** by Mike Scott (A set of tools for text analysis. Includes a *wordlister*, a *concordancer*, a *keyword analyzer* and more. Distributed by Oxford University Press)
- **TextSTAT** - Free concordance software for Windows and Linux
- **Microconcord** (DOS version of WS Tools *concordancer*)
- **ConcApp** concordancing programs (Freeware)
- **Concordance** (Wordlists, concordances. For publishing concordances on the WEB. By R.J.C. Watt)
- **WordExpert** (A concordancer for technical translators by myteam-Software)
- **Monoconc** (*Concordancer* by M. Barlow, distributed by Athelstan)
- **Paraconc** (Concordancer for parallel texts, by M. Barlow)
- **Multiconcord** (Concordancer for parallel texts, by David Wools)
- **PWA**- The Plug Word Aligner A collection of tools for the automatic alignment of word correspondences in bilingual parallel texts
- **LEXA** *Corpus Processing Software* distributed by ICAME.
- **DBT** (*Corpus Processing Software* developed by Eugenio Picchi at the Istituto di Linguistica Computazionale del CNR in Pisa).

[10] The corpus of Economic recession is constituted of BBC newspaper articles downloaded from www.bbc.co.uk from February to March 2011.

[11] The source of this list is Federico Zanettin's website at http://www.federicozanettin.net/sslmit/cl.htm

— **TactWEB** (*Corpus Processing Software* developed by John Bradley and Lidio Presutti, University of Toronto.

Some software is available for free and can be downloaded directly from the internet.

In the following section, a corpus of newspaper articles on the economic recession will be used to describe the features of TextSTAT, a concordance software freely downloadable from the internet[12], and WordSmith Tools, one of the most widely used software for corpus analysis..

2.2 TextSTAT and WordSmith Tools

TextSTAT

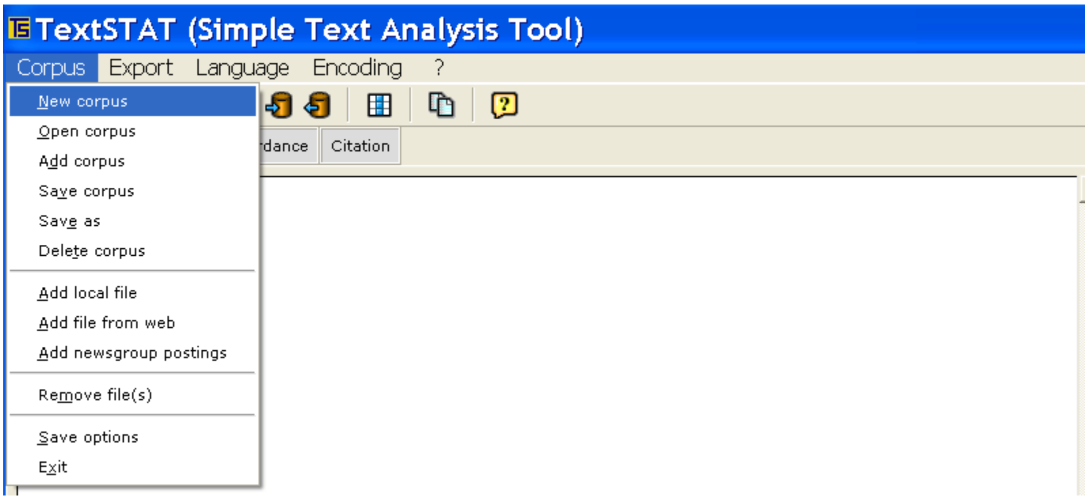
TextSTAT[13] is a very simple and basic programme and only produces wordlists and concordances. However, it is ideal for those students who take their first steps into the world of corpus linguistics.

With TextSTAT it is possible to learn how often a certain word occurs or in what contexts it is used.

To use TextSTAT, first a corpus should be created by clicking on "Corpus" and then on "New corpus". (It is possible to open an existing corpus which has already been created in previous applications).

[12] http://www.niederlandistik.fu-berlin.de/textstat/software-en.html

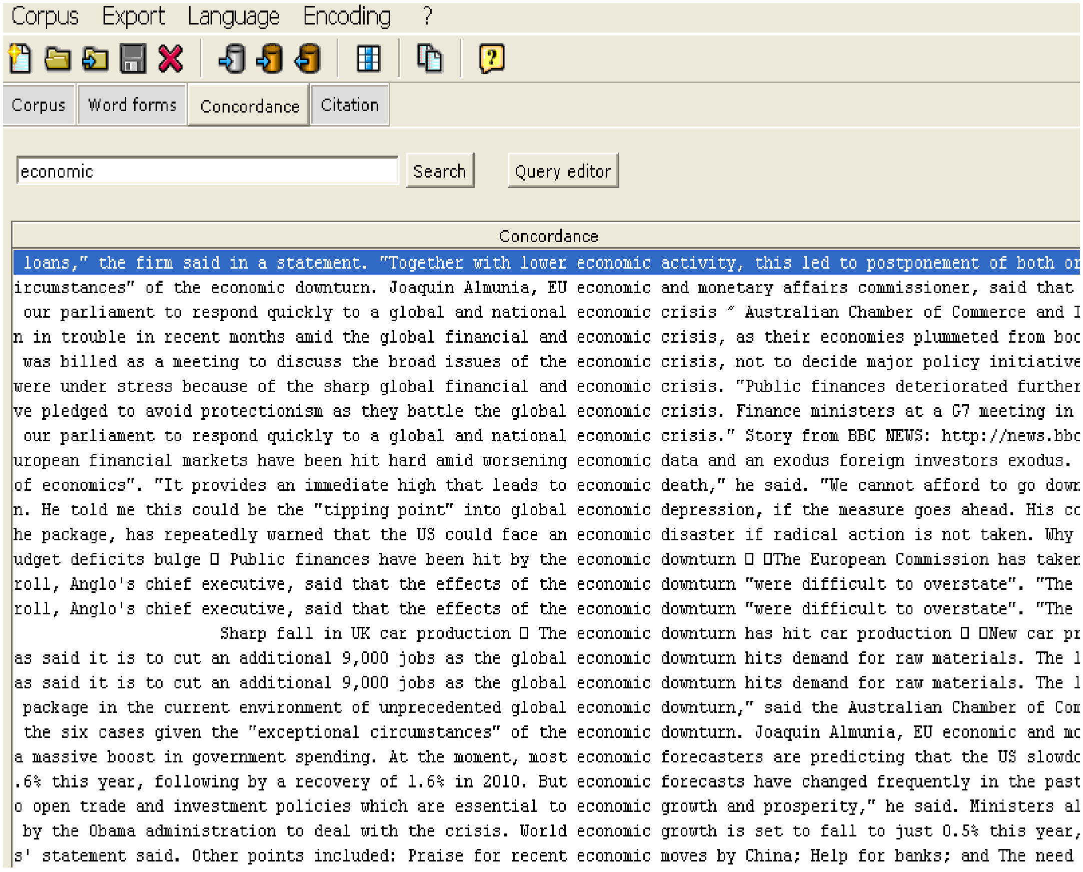
[13] A detailed description of how to use TextSTAT is provided by its creator Mattias Huening at http://www.niederlandistik.fu-berlin.de/textstat/TextSTAT-Doku-EN.html



After creating a new corpus, it is possible to add the files in your corpus by clicking on the second cylinder in the upper toolbar.

When your texts have been uploaded, click on the tab sheet “Word forms” and then on “ Frequency list”. A list of all the words contained in the corpus in frequency order can thus be obtained. The number of words constituting our corpus is indicated in the lower bar.

By clicking twice on e ach word of the wordlist, it is possible to have the concordance of the word chosen, which contains the node word and its co-text as it occurs in all the texts selected for analysis. The table below provides the concordance of the word *economic*. As can be easily seen, the words immediately following and preceding the node word can be sorted alphabetically in order to make the retrieval of recurrent patterns easier.



economic

```
 loans," the firm said in a statement. "Together with lower economic activity, this led to postponement of both or
ircumstances" of the economic downturn. Joaquin Almunia, EU economic and monetary affairs commissioner, said that
 our parliament to respond quickly to a global and national economic crisis ″ Australian Chamber of Commerce and I
n in trouble in recent months amid the global financial and economic crisis, as their economies plummeted from boc
 was billed as a meeting to discuss the broad issues of the economic crisis, not to decide major policy initiative
were under stress because of the sharp global financial and economic crisis. "Public finances deteriorated further
ve pledged to avoid protectionism as they battle the global economic crisis. Finance ministers at a G7 meeting in
 our parliament to respond quickly to a global and national economic crisis." Story from BBC NEWS: http://news.bbc
uropean financial markets have been hit hard amid worsening economic data and an exodus foreign investors exodus.
of economics". "It provides an immediate high that leads to economic death," he said. "We cannot afford to go dowr
n. He told me this could be the "tipping point" into global economic depression, if the measure goes ahead. His cc
he package, has repeatedly warned that the US could face an economic disaster if radical action is not taken. Why
udget deficits bulge □ Public finances have been hit by the economic downturn □ □The European Commission has taker
roll, Anglo's chief executive, said that the effects of the economic downturn "were difficult to overstate". "The
roll, Anglo's chief executive, said that the effects of the economic downturn "were difficult to overstate". "The
                     Sharp fall in UK car production □ The economic downturn has hit car production □ □New car pr
as said it is to cut an additional 9,000 jobs as the global economic downturn hits demand for raw materials. The l
as said it is to cut an additional 9,000 jobs as the global economic downturn hits demand for raw materials. The l
 package in the current environment of unprecedented global economic downturn," said the Australian Chamber of Com
 the six cases given the "exceptional circumstances" of the economic downturn. Joaquin Almunia, EU economic and mc
a massive boost in government spending. At the moment, most economic forecasters are predicting that the US slowdc
.6% this year, following by a recovery of 1.6% in 2010. But economic forecasts have changed frequently in the past
o open trade and investment policies which are essential to economic growth and prosperity," he said. Ministers al
 by the Obama administration to deal with the crisis. World economic growth is set to fall to just 0.5% this year,
s' statement said. Other points included: Praise for recent economic moves by China; Help for banks; and The need
```

Furthermore, by clicking twice on each concordance line it is possible to see the original document which contains the string selected. This action corresponds to the tab sheet "citation".

WordSmith Tools[14]

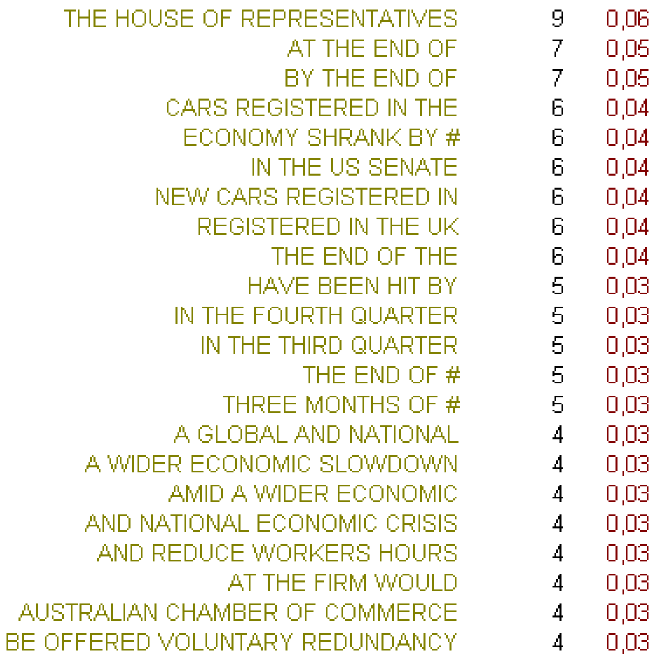
WordSmith Tools has many features in common with TextSTAT but it is a more advanced tool for corpus analysis. The main tools it provides are *Concord, KeyWords* and *WordList* but other more specific and complex tools are also available.

Concord provides the concordance of the node word selected as well as a list of its collocates and of the recurring clusters or phrases (Scott 2001).

[14] A step-by-step guide to the use of version 5.0 of this software is provided at http://www.lexically.net/wordsmith/step_by_step/index.html.

KeyWords works with two wordlists originated by two different corpora, generally a co rpus and a b igger reference corpus. These wordlists can be compared in order to identify ‘key words’, that is to say those words which are more (or sometime less) frequent in the main corpus than in the reference corpus. These words are defined as keywords in that they are a key in the analysis of the language of a genre, of a domain or even of the works or speeches of a single author. Milizia and Spinzi (2008; see also Phillips 1989; Berber-Sardinha 2004; Scott 2006; Scott & Tribble 2006) have analysed the word *people* in the speeches of Tony Blair and George W. Bush. The word is heavily used by both politicians. They notice, however, that in the British list *the Iraqi people* is more frequently used than *the British people*, ranking 7th, whereas *the American people* ranks first in the American corpus. Scrolling down the list, it is clear that the main concerns of the two politicians are different.

WordList, as in TextSTAT, creates word lists, ordering them by frequency and alphabetically. However, *WordSmith* can generate wordlists that can be either one-word list or up-to-five-word lists, thus allowing the researcher to identify the most frequent patterns of a language in a few seconds. See figure below:



THE HOUSE OF REPRESENTATIVES	9	0,06
AT THE END OF	7	0,05
BY THE END OF	7	0,05
CARS REGISTERED IN THE	6	0,04
ECONOMY SHRANK BY #	6	0,04
IN THE US SENATE	6	0,04
NEW CARS REGISTERED IN	6	0,04
REGISTERED IN THE UK	6	0,04
THE END OF THE	6	0,04
HAVE BEEN HIT BY	5	0,03
IN THE FOURTH QUARTER	5	0,03
IN THE THIRD QUARTER	5	0,03
THE END OF #	5	0,03
THREE MONTHS OF #	5	0,03
A GLOBAL AND NATIONAL	4	0,03
A WIDER ECONOMIC SLOWDOWN	4	0,03
AMID A WIDER ECONOMIC	4	0,03
AND NATIONAL ECONOMIC CRISIS	4	0,03
AND REDUCE WORKERS HOURS	4	0,03
AT THE FIRM WOULD	4	0,03
AUSTRALIAN CHAMBER OF COMMERCE	4	0,03
BE OFFERED VOLUNTARY REDUNDANCY	4	0,03

2.3 Assembling corpora

The advantages of using corpora will be described in the rest of this book, particularly relating them to the fields of translation and language learning.

However, one of the main advantages of corpora lies in the authenticity of the texts contained in them which allow us to validate the theory and to have some insights into actual language that would not be possible using only our intuition of native speakers of a language.

For this reason, in order to be representative of a language and "to capture the regularities of a l anguage" (Tognini Bonelli, 2001:53), corpora should be assembled on the basis of a number of criteria. In the EAGLES recommendations on corpus typology (1996) linguistic criteria may be of two different types: 1. external, in that they concern the participants, the occasion, the social setting or the communicative function of the pieces of language; and 2. internal, in that they concern the recurrence of language patterns within the pieces of language. According to Sinclair (2005: 1) and Clear (1992) corpora should be designed and constructed exclusively on external criteria. Texts should be selected on the basis of their communicative function in the community in which they have been produced.

Tognini Bonelli (2001:54ff) analyses some authoritative definitions of corpus provided by Francis (1992:17), Sinclair (1991:171), Aarts (1991:45), the EAGLES project report (1996:2.1) and identifies three main issues which are extremely relevant to the process of text selection and corpus assembling. The issues are:

1. Authenticity of the texts included in the corpus.

All texts should represent language used in authentic language events. Invented examples or texts fabricated by the linguist cannot be the object of analysis in that they are not representative of real language (see Biber 1988; Aarts 1991; Sinclair

1991, 2005[15]; Johansson 1995; Carter and McCarthy 2003; Baroni and Bernardini 2006, to name but a few);

2. The representativeness of language included in the corpus.

Sinclair (2005: 2) says that "corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen". It means that texts should be chosen according to the purpose of the analysis.

The Brown Corpus of Standard American English was the first of the modern, computer readable, general corpora[16]. The corpus contains one million words of American English texts printed in 1961. The texts for the corpus are samples of 15 different text categories. Nowadays, it is considered too small to be a good standard reference for the English language, particularly if compared to the Bank of English[17] which, at the time of writing, amounts to 400 million words. It is a collection of samples of modern English language and contains both written and spoken samples from British, US, Australian and Canadian sources. Written texts come from newspapers, magazines, fiction and non-fiction books, brochures, leaflets, reports, letters, and so on. S poken texts are transcriptions of everyday casual conversation, radio broadcasts, meetings, interviews and discussions, etc. The corpus is constantly updated so that the corpus can be representative of the English which most people read, write, speak and hear every day of their lives. It is used by the Collins COBUILD Advanced Learner's English Dictionary as evidence of patterns of word combination, word frequencies, uses of particular words, and of meaning disambiguation.

Specialised corpora are made up of texts that belong to the same genre and deal with a specific topic. Sinclair (2001:xi) says that "if it is a general corpus, researchers expect to find in it information about the language as a whole, and if it

[15] Sinclair, J. (2005), *Corpus and Text - Basic Principles*, in Developing Linguistic Corpora: a Guide to Good Practice, Oxford, Oxbow Books, pp. 1-16.

[16] See http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

[17] See http://www.titania.bham.ac.uk/docs/about.htm

is a more specialised corpus, then the characteristics of the genre will be discoverable".

3. The sampling criteria used in the selection of texts.

According to Sinclair (2005:3) common sampling criteria used in the selection of texts for corpora include:

(a) the mode of the text; whether the language originates in speech or writing or in electronic mode;

(b) the type of text; for example if written, whether a book, a journal, a notice or a letter;

(c) the domain of the text; for example whether academic or popular;

(d) the language or languages or language varieties of the corpus;

(e) the location of the texts; for example (the English of) UK or Australia;

(f) the date of the texts.

All these criteria should be taken into account when assembling a corpus. If the purpose of analysis is the academic written language the corpus cannot include spoken texts or articles which have been published on a magazine rather than on a scientific journal. A corpus on academic written language should be constituted by scientific articles from different scientific domains (such as history, literature, language, biology, economics, mathematics, …). Furthermore, a decision should be made on whether to select only articles written by native speakers of English or to consider articles from all the scientific communities. Texts should have been published in a limited time span: selecting together scientific articles published in 1960 and scientific articles published in 2009 should be avoided; language changes over time and considering together texts in a too wide time span would make it impossible to have a clear picture of the features of modern academic written language. However, if our main aim is that of analysing this type of language as it has changed across the years, two corpora can be assembled: one made up of scientific articles published, for example, between 1960-1970 and another corpus of articles published between 1990-2000. The two wordlists obtained could be compared in order to identify the key words which represent the changes that the academic language has undergone across the years.

2.4 Conclusion

This chapter has described the features of corpora as used in Corpus Linguistics. Corpora are large collections of texts stored in electronic format and allow both linguists and students to obtain different types of insights into a language or a variety of a language. The size of corpora may vary according to the purpose of our analysis. Observations on the language as a whole and as used in everyday contexts by ordinary people can only be obtained from big corpora constituted of millions of words. Conversely, special corpora, which although small are specialised in their content, allow us to get insight into the typical phraseology, expressions and lexico-grammatical features of a variety of a language.

The internet has many features in common with corpora. It is constituted of millions of texts belonging to different text types. However, it also has a number of shortcomings. The main problem is, obviously, constituted by authorship. Most of the times we do not have any information on who has written a text stored on the internet, we do not know their nationality, age, gender, social class, education and so on. Furthermore, information such as the date of the text and the location of the text are usually unknown.

Nevertheless, if its limits are properly taken into account, the internet may be a valid support for students when they need to check the correctness or the frequency of usage of some expressions and collocations they use in language production.