Corpus Manager
A Tool for Multilingual Corpus Analysis

George Kouklakis,[1] George Mikros,[2]
George Markopoulos[1] and Ilias Koutsis[1]

1. Introduction

The vast amount of electronic texts in the web facilitates the creation of megacorpora at a very short time. However, as the corpus size increases, the complexity of its management is becoming a serious problem limiting its functionality. Furthermore, a great deal of corpus linguistics research is based on the quantitative comparison of small corpus samples, which are drawn from a bigger general language corpus based on specific criteria such as authorship, topic, genre, register etc (Biber 1993). For these reasons a number of tools have already been developed and aim to organize and handle texts in corpora (e.g. Christ 1994, Holmes-Higgin et al. 1994). However, most of the developed systems have a significant learning curve and limited flexibility regarding the metadata which can be used as subcorpus selection criteria.

2. Related work

There are a vast number of tools which perform basic text analysis. Some characteristic ones are the following:

2.1 Wordsmith tools

Wordsmith tools (Scott 1996) are an integrated suite of programs that process corpora in many different ways. The suite is mainly consisted of three tools:

- The Wordlist tool which lets the user see a list of all the words in a text, set out in alphabetical or frequency order, as well as a number of text statistics related to the analyzed corpus.
- The concordancer, Concordance, which gives the user a chance to see any word or phrase in context — as well as a number of collocation statistical data such as horizons, collocated frequency etc.
- The KeyWords which helps the user find the key words in a text.

In its latest version (ver. 4) WordSmith Tools are capable of extracting collocations, both general and keyword-based, and implement four association

[1] Department of Linguistics, University of Athens
e-mail: g_kouklakis@yahoo.gr, gmarkop@phil.uoa.gr, ilias_k@yahoo.com
[2] Department of Italian and Spanish Language and Literature, University of Athens
e-mail: gmikros@isll.uoa.gr

measures to compute them, namely the MI, Z-score, MI3 (i.e. MI cubed) and Log-likelihood. They can also handle multiple files and support XML.

2.2 Antconc

AntConc (Anthony 2004) is a free program which can handle txt files (.txt) and html files (.html) saved on the user's hard disk. AntConc allow users to search for concordances and then sort the concordance lines in several forms (e.g., alphabetically by the node word and by the left side and right side of the node word).

Antconc allows users to look for word clusters and permits the search for different sizes and types of clusters. It has a KeyWord feature that allows users to choose the list of words to which they want to compare their texts.

Antconc provides word lists sorted by frequency, alphabetically. It also provides frequency lists either by pre-establishing the minimum and the maximum number of appearances in the corpus or by searching for a specific word. AntConc saves all the output in text files only.

2.3 Concordance

By using Concordance a researcher can make full concordances to texts of any size or by picking a selection of words from text. The program can also makes concordances to html texts saved in the hard disk of the user's PC. Its user can:

- view a full wordlist, a concordance and the original text simultaneously,
- browse through the original text and click on any word to see every occurrence of that word in its context,
- edit and re-arrange a wordlist by drag and drop,
- search, select, and sort words,
- identify which section of a text each citation comes from,
- display the words together with their contexts which can vary by length or sense-unit.

2.4 Monoconc

Monoconc is mainly a concordance tool. It can handle multiple text files in different languages and their contents can be viewed from within the concordancer. The major functions of the tool are:

- Concordance: word (or part of word) or phrase concordance search. An "Append search" option allows the results of a new search to be added to the results of a previous concordance search. A larger (multi-line) context can be displayed for a selected concordance line.
- Sort: Concordance results can be sorted 1L (First Left), 1R, 2L, 2R, as well as by search word and by text order; it also allows primary and secondary sorts.

- Wordlist: a word list function provides a frequency count for the words in the corpus (with display in alphabetical or frequency order). In addition, the frequency of collocates of the search word (2L, 1L, 1R and 2R) is calculated. Specific words can be excluded from these frequency counts using a user-made stop-list.
- Display/Output: A "hide keyword" option is available and the search results can be toggled between a KWIC format and sentence mode. The concordance results can be saved to a file and/or printed. The name of the source file can be saved along with each concordance line.

2.5 Textstat

TextSTAT is a free program which can handle txt files (.txt), Word files (.doc) and html files (.html). It contains a Web spider that captures the text directly from the Internet. The users can type the Web address and choose the number of pages they want the Web spider to include in the corpus.

TextSTAT apart from providing concordance lines based on the uploaded texts, it also contains a feature named "query editor," which permits the localization of collocates.

TextSTAT provides word lists sorted by frequency, alphabetically. It also provides frequency lists either by pre-establishing the minimum and the maximum number of appearances in the corpus or by searching for a specific word.

The program can save the word lists in CSV (comma-separated values format) or Excel files and the concordances on text or Word .doc type files.

All of the above tools presented give the user a lot of functionality as far as basic text analysis features are concerned (concordance, extraction of wordlists/keywords lists etc). However, they do not address two major issues:

- Metadata extraction and management.
- Extraction of quantitative data from the analyzed corpus and merge with its associated metadata information.

3. Tool presentation

We are entering a new era in the World Web and (to be more specific) in the design and the creation of the Semantic Web with the use of metadata and metatextual information more generally. In the Semantic Web there is an increasing need for the use and process of big collections of texts which should be constituted by commented texts providing rich metadata information.

Corpus manager is different in its aim from the tools which were presented earlier. It was designed to give the ability to the user not only to execute basic text analysis in texts or collections of texts but also to perform statistical analyses and combine them with metadata information about the processed text. This functionality enables the user to create automatically vector representation of documents in a number of linguistic attributes which can readily be submitted to machine learning of statistical analysis software in order to analyze further the data.

Corpus Manager is a tool which provides the ability to the user of processing corpora that are accompanied by a file of metadata (file of a description for each file

in each corpus). As minimal information each file of metadata should have the place of the file in the user's hard disk. For those corpora that they do not have a metadata file, Corpus Manager provides the user with a relative tool so that he/she can produce an elementary metadata file. This tool enables the user to produce not only this simple metadata file (which includes only the place of the text file in the user's hard disk where the corpus is stored) but also an information "rich" metadata file, if that information can be extracted from the text files which the corpus consists of.

By the use of Corpus Manager the user can regroup the files in a corpus according to metadata information. This is particularly important because the user can perform statistical analyses based on a number of metadata attributes of each text file such as topic, genre, year, author etc.

Apart from performing statistical analyses in a corpus or parts of it, Corpus Manager can be used for the automatic creation of a sub-corpus from an initial corpus. The user selects specific criteria on the metadata of the text files of the initial corpus and the program creates a sub-corpus based on the user's selection. These subcorpora can be "physical" (i.e. they are created and saved in the hard disk of the user) or "virtual" (i.e. they are only created during the process but they are not saved anywhere).

3.1. Creation of a sub-corpus

As it is already noted, the user of the program can create a corpus of text files which will be a subset of a bigger corpus (it is saved in the hard disk of the user) (see figure 1). The new corpus not only will contain the text files from the initial corpus, but it will also have its own metadata. The new metadata file will have the same structure as the metadata file of the initial corpus. The new corpus will be stored in a folder in the hard disk of the user (the name of the folder is entered by the user).

This corpus will be available to the user for further analyses and process with the other units of Corpus Manager.

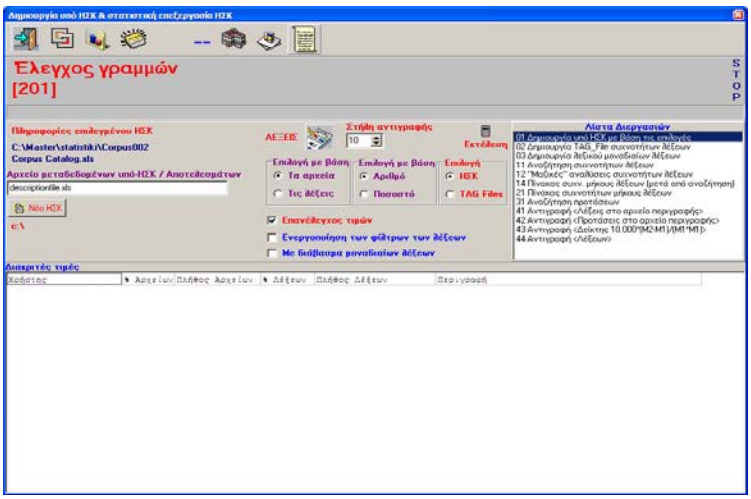


Figure 1: Sub-corpus creation

In order to create a sub-corpus the user must select criteria (see figure 2) which will be applied in order to form the new corpus. These criteria can be:

- The exact number of words of the text files of the initial corpus.
- A percentage of words of the text files of the initial corpus.
- The exact number of text files of the initial corpus.
- A percentage of text files of the initial corpus.

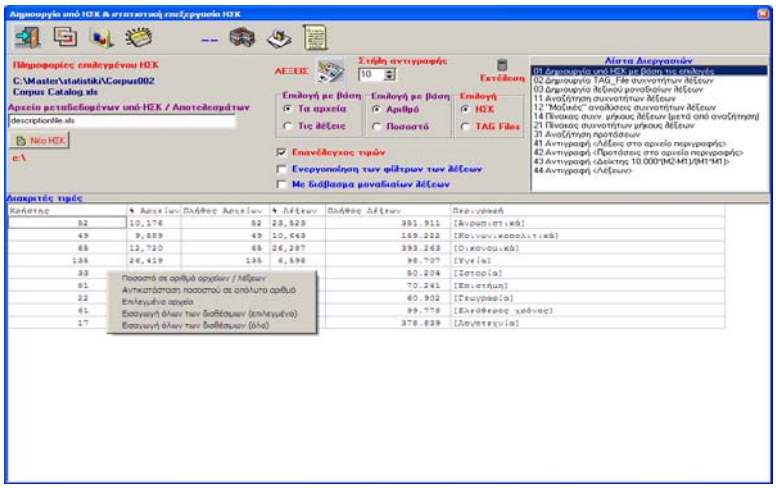


Figure 2: Criteria selection for the new sub-corpus.

The files included in the sub-corpus are randomly selected from the initial corpus. Whenever the user wants, he/she can see the names of the files selected and he/she can update the sub-corpus (select other files from the initial corpus or remove any of the already selected files).

For example the user can select the new corpus to be consisted of files from the initial corpus that have 1000 words overall or 25% of the total number of words in the initial corpus or 150 files from the initial corpus or 50% of the total number of files in the initial corpus.

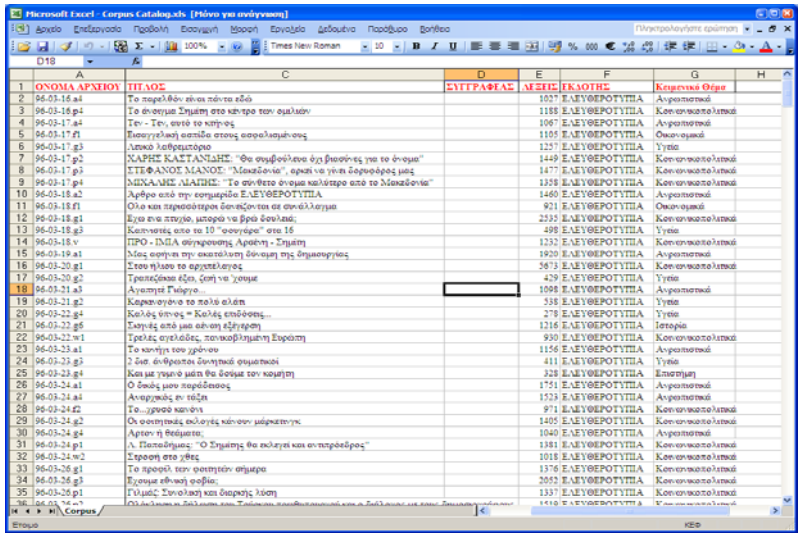
These criteria can be set either independently of the information stored in the metadata file of the initial corpus or by selecting one or more categories in the metadata file.

For example the user can decide that its new sub-corpus will consist of files that will belong to the topic category “History” (25%), to the category “Politics” (30%) and to the category “Environment” (45%). If the user gives numbers that don’t sum up to 100% (e.g. History 2%, Politics 1%, and Environment 1%) the program will convert them suitably (e.g. History 50%, Politics 25%, Environment 25%).

3.2 Conditions for the creation of a sub-corpus

Before the creation of the sub-corpus, the user can choose to display the names of the files which have been selected from the initial corpus. Therefore, if he/she wishes to, he/she has the opportunity to replace or even add other files (manual file selection).

When the new sub-corpus is created, it is accompanied by a metadata file (see figure 3) and it is henceforth ready for any process. In other words it has the same characteristics as the corpus from which it was created.



	A	C	D	E	F	G	H
1	ΟΝΟΜΑ ΑΡΧΕΙΟΥ	ΤΙΤΛΟΣ	ΣΥΓΓΡΑΦΕΑΣ	ΛΕΞΕΙΣ	ΕΚΔΟΤΗΣ	Κειμενικό Θέμα	
2	96-03-16.a4	Το παρελθόν είναι πάντα εδώ		1027	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
3	96-03-16.p4	Το άτομο με Σημάτη στο κέντρο των συμλιών		1188	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
4	96-03-17.a4	Τεν - Τεν, αυτό το κτήνος		1067	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
5	96-03-17.f1	Εισαγγελική αστίδα στους ασφαλισμένους		1105	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Οικονομικά	
6	96-03-17.g3	Λιπικό λαβυρινθόδρομο		1257	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Υγεία	
7	96-03-17.p2	ΧΑΡΗΣ ΚΑΣΤΑΝΙΔΗΣ: "Θα συμβούλευα όχι βιασύνες για το όνομα"		1449	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
8	96-03-17.p3	ΣΤΕΦΑΝΟΣ ΜΑΝΟΣ: "Μακεδονία", αρκεί να γίνει δορυφόρος μας		1477	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
9	96-03-17.p4	ΜΙΧΑΛΗΣ ΛΙΑΠΗΣ: "Το σύνθετο όνομα καλύτερο από το Μακεδονία"		1358	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
10	96-03-18.a2	Άρθρο από την εσημερίδα ΕΛΕΥΘΕΡΟΤΥΠΙΑ		1460	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
11	96-03-18.f1	Όλο και περισσότερο δανειζόνται σε συνάλλαγμα		921	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Οικονομικά	
12	96-03-18.g1	Έχω ένα πτυχίο, μπορώ να βρώ δουλειά;		2535	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
13	96-03-18.g3	Καπνιστές από τα 10 "σουγιάρα" στα 16		498	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Υγεία	
14	96-03-18.v	ΠΡΟ - ΠΑΑ σύγκρουσης Αρσένη - Σημίτη		1232	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
15	96-03-19.a1	Μας αφήνει την ακατάλυση δύνομη της δημιουργίας		1920	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
16	96-03-20.g1	Στου ηλιου το αρχτέλαγος		3673	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
17	96-03-20.g2	Τραπεζάκια έξω, ζωή να 'χουμε		429	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Υγεία	
18	96-03-21.a3	Αγαπητέ Γιώργο...		1098	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
19	96-03-21.g2	Καρκινογόνο το πολύ αλάτι		538	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Υγεία	
20	96-03-22.g4	Καλές γίνες, Καλές επιδόσεις...		278	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Υγεία	
21	96-03-22.g6	Σκηνές από μια αέναη εξέγερση		1216	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ιστορία	
22	96-03-22.w1	Τρελές αγελάδες, πανικοβλημένη Ευρώπη		930	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
23	96-03-23.a1	Το κανήγι του χρόνου		1156	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
24	96-03-23.g3	2 δισ. άνθρωποι δυνητικά σωματακοί		411	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Υγεία	
25	96-03-23.g4	Και με γυμνό μάτι θα δούμε τον κομήτη		328	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Επιστήμη	
26	96-03-24.a1	Ο δικός μου παράδεισος		1751	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
27	96-03-24.a4	Αναχωρώ εν τάξει		1523	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
28	96-03-24.f2	Το...χρυσό κανόνι		971	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
29	96-03-24.g2	Οι ανοιηταικές εκλογές κάνουν μάρκετινγκ		1405	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
30	96-03-24.g4	Άρτευ ή θεάματα;		1040	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Ανθρωπιστικά	
31	96-03-24.p1	Λ. Παπαδήμος: "Ο Σημίτης θα εκλεγεί και αντιπρόεδρος"		1381	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
32	96-03-24.w2	Στροφή στα χθες		1018	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
33	96-03-26.g1	Το προφίλ των φοιτητών σήμερα		1376	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
34	96-03-26.g3	Έχουμε εθνική σοφία;		2055	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	
35	96-03-26.p1	Γιλμάζ: Συνολική και διαρκής λύση		1337	ΕΛΕΥΘΕΡΟΤΥΠΙΑ	Κοινωνικοπολιτικά	

Figure 3: Example of metadata file.

3.3 Quantitative analysis of corpus data

3.3.1 Quantitative text analysis.

The user is provided with the ability of producing various quantitative analyses that concern words, sentences, sets of characters etc. More specifically, the user has the ability to calculate:

- The frequencies of all different words which are contained in the corpus files.
- The word length spectrum that is the frequency of 1 character words to 14 character words in the corpus.
- The average word length of each text measured in characters.
- The sentence length of each text in the corpus measured in words.
- Vocabulary richness measures such as Yule's K, and Type/Token ratio (Tweedie and Baayen 1998).

The above analyses can apply to all or in some of the files in the sub corpus. If the user wants to apply the above analyses in a specific number of files in its sub-corpus, he has the ability to select the files in two ways:

- Automatically using random selection.
- Manually establishing specific criteria which are of the same type that the user utilizes for the sub-corpus creation (see figure 4).

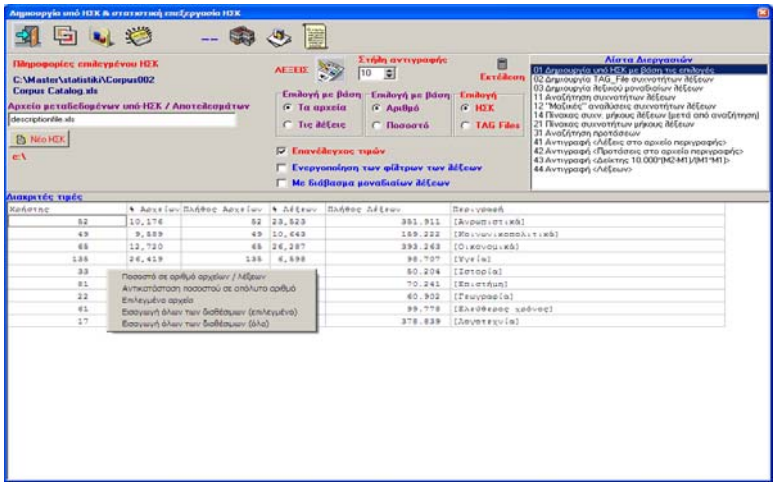
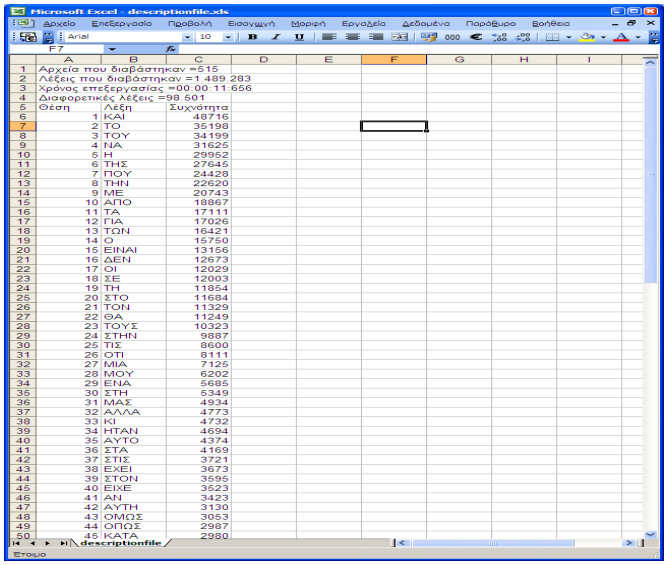


Figure 4: Criteria selection for quantitative text analysis.

An example of the results obtained is shown in Figure 5:



	A	B	C
1	Αρχεία που διαβάστηκαν =515		
2	Λέξεις που διαβάστηκαν =1.489.283		
3	Χρόνος επεξεργασίας =00:00:11.656		
4	Διαφορετικές λέξεις =98.501		
5	Θέση	Λέξη	Συχνότητα
6	1	ΚΑΙ	48716
7	2	ΤΟ	35198
8	3	ΤΟΥ	34199
9	4	ΝΑ	31625
10	5	Η	29952
11	6	ΤΗΣ	27645
12	7	ΠΟΥ	24428
13	8	ΤΗΝ	22620
14	9	ΜΕ	20743
15	10	ΑΠΟ	18867
16	11	ΤΑ	17111
17	12	ΓΙΑ	17026
18	13	ΤΩΝ	16421
19	14	Ο	15750
20	15	ΕΙΝΑΙ	13156
21	16	ΔΕΝ	12673
22	17	ΟΙ	12029
23	18	ΣΕ	12003
24	19	ΤΗ	11854
25	20	ΣΤΟ	11684
26	21	ΤΟΝ	11329
27	22	ΘΑ	11249
28	23	ΤΟΥΣ	10323
29	24	ΣΤΗΝ	9887
30	25	ΤΙΣ	8600
31	26	ΟΤΙ	8111
32	27	ΜΙΑ	7125
33	28	ΜΟΥ	6202
34	29	ΕΝΑ	5685
35	30	ΣΤΗ	5349
36	31	ΜΑΣ	4934
37	32	ΑΛΛΑ	4773
38	33	ΚΙ	4732
39	34	ΗΤΑΝ	4694
40	35	ΑΥΤΟ	4374
41	36	ΣΤΑ	4169
42	37	ΣΤΙΣ	3721
43	38	ΕΧΕΙ	3673
44	39	ΣΤΟΝ	3595
45	40	ΕΙΧΕ	3523
46	41	ΑΝ	3423
47	42	ΑΥΤΗ	3130
48	43	ΟΜΩΣ	3053
49	44	ΟΠΩΣ	2987
50	45	ΚΑΤΑ	2980

Figure 5: Example of result file in Excel

3.3.2 Multiple automatic analyses

Corpus Manager, implements a module which can perform automatically multiple comparisons in frequency wordlists. The aim of this tool is to compare the lexical frequency in different subcorpora which vary systematically in specific metadata values. An example could be the investigation of the topic influence in word frequency. Corpus manager can create a number of subcorpora in different topics in which the percentage of words in each topic is systematically manipulated in relation to the others. In such an example we can compare a number of frequency wordlists that represent subcorpora with topics which vary in relation to the words that each topic contribute to the sub-corpus (e.g. First Wordlist: Sport 33,3%, Politics 33,3%, Culture 33,3%, Second Wordlist: Sport 20%, Politics 40%, Culture 40%, Third Wordlist: Sport 40%, Politics 20%, Culture 40%, etc.).

The user should create a file of parameters-commands which will activate the above automatic analyses. Theoretically the user has the ability to create as many analyses as he wishes. Practically, however, he/she is limited by the Excel file format (that is the appearance of only 256 columns).

All the analyses are executed one by one and the results are exported to a tab delimited file in .xls format (Microsoft Excel). All the word frequency lists resulted from the different subcorpora specified by the user are presented in the results file and each frequency wordlist is aligned in word level in order to facilitate word frequency comparisons.

3.4. Auxiliary tools of Corpus Process Unit

In order to accelerate the process of word searching and frequency counting, the following two techniques have been used in the Corpus Process Unit:

i. The first technique that we named "Types Dictionary in the File (TDF)" creates a metadata structure for each text of the corpus. This metadata structure is stored by the program as a textfile in a folder named "tagfile". Each metadata file contains:

a. The number of all the words in the file.
b. The number of the different types of words in the file.
c. The list of all types of the words in the file and their frequency in it.

If the TDF technique has not been used, Corpus Manager searches for the frequencies of words using the following procedures:

1. Reads the text and searches for the existence of tags which should be removed.
2. Identify the words of the text, one by one.
3. Identify the different words of the text.
4. Calculate the frequency of each word.

With the use of the technique TDF, only the third step is executed. The first two steps together with the fourth one are not needed because the particular information already exists in the TDS files stored in "tagfile" folder.

Thus, the speed in the search of frequencies of words is particularly improved by using this technique. If a user wants to use it, he must make the TDF files for the corpus the first time he will use it. The production process of the TDF files requires almost the same time as the one required for a simple search of the unique types in the corpus. Once the TDF files are produced, they are available to each user of the Corpus Manager.

ii. The second technique, which we named 'Types Dictionary in the Corpus (TDC)', aims at improving even more the processing time during the search of word frequencies. More specifically the process steps, each time the user asks for a list of word frequencies in a corpus, are the following:

1. The process unit of the Corpus Manager reads all the corpus files one by one.
2. From each file the Corpus Manager collects all the words.

3. The Corpus Manager classifies each word in a list in the computer memory (in case this word is not already located in the list). In case the word is located in the list, the unit increases its frequency by one.

In order to improve the total processing time, each processing step has to be improved. In fact the processing times of steps 1and2 do not accept further improvement than the one achieved by the TDF technique. During speed experiments we noticed that most of the time was consumed in the registration of each new word in the list (step 3) and not in the increase of the frequency of a word already existing in the list (the time necessary for the increase of the frequency is negligible). Moreover, we noticed that the time required for sorting each new word was much more than the time required for both the first two steps and more particularly when the number of the different words in the list was substantially increased.

In order to achieve a substantial decrease in the processing time during the word frequency searching, we should decrease the sorting time of the new words in the list in step 3. Yet, this is not possible, as the sorting algorithm is standard.

Having reached this point we decided to create in each corpus a 'file of types' (types dictionary). The first time the user uses the corpus, he will follow all three steps in order to produce a file with the corpus types. After that, each time a tool unit executes a word frequency search in the corpus, it utilizes this file and simply calculates their frequencies in the files of the corpus.

If the user wants to produce a file of types, he uses the procedure "Creation of Types Dictionary in the Corpus (TDC)". The production procedure of the types file requires exactly the same time as the time required for the process of calculating the frequencies of the types in the corpus, but it has the advantage of being executed only once and the results are directly utilizable by all the users of the software. The use of the TDC improved drastically the processing time. More specifically, the word frequency searching requires about the 1/30 of the initial time (the time in which neither the first nor the second techniques are used) in a corpus of 1.500.000 words.

It should be noticed that the two techniques (TDF and TDC) can be combined for even higher processing speeds. Furthermore, the software provides the user with the possibility to join files of types coming from different corpora, as well as to utilize in another corpus the types file coming from one corpus.

The following graph displays the times required in order to find the types frequency in a corpus. The corpus under examination consisted of 515 files with a total of 1.663.067 words. The computer used was a HP having a PII processor at 450 MHz.

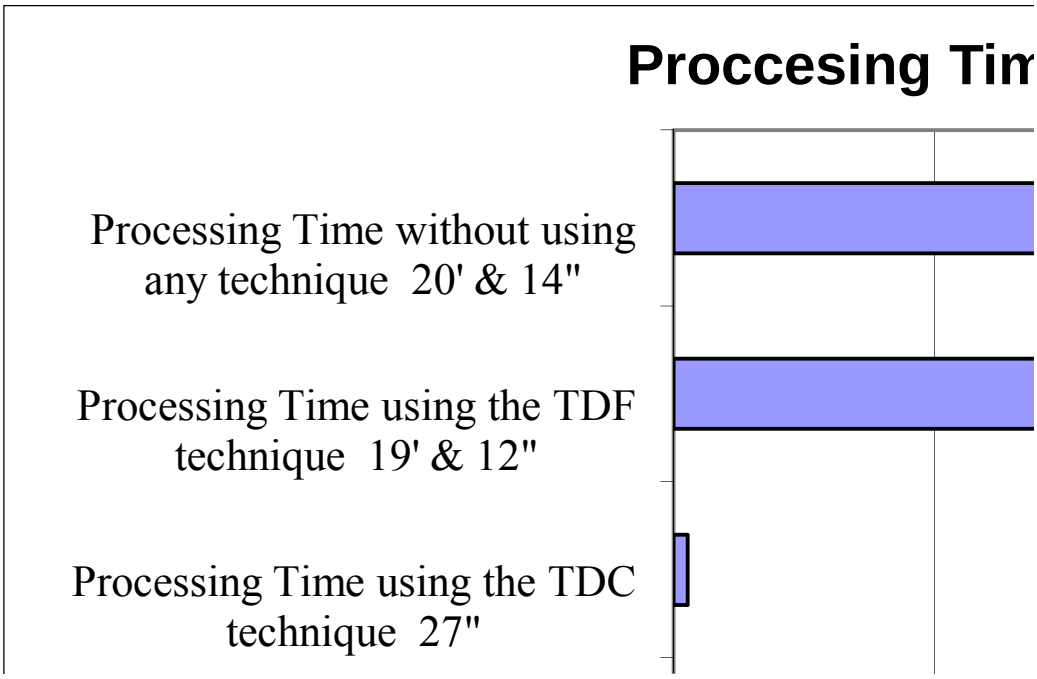


Figure 6: Processing times of different search techniques

In the Figure 6 above, it should be noted that the time needed for the creation of the TDC file is 20΄:14΄΄, while the time needed for the TDF files is 2΄:52΄΄.

3.5 Word frequency counting

The search of specific words or word groups is particularly useful in corpus linguistics research. Corpus Manager can search and count the frequency of:

- words
- phrases
- word groups or phrase groups

The user is able to form complex queries using the following search criteria:

- The word position within the sentence (left, right, justified, etc.).
- The word's case (uppercase, lowercase, etc.).

3.6 Initialization of the Corpus process unit

The user of the Corpus Manager can customize a number of processes. More specifically, the user can:

- Determine the word separators from the ASCII table. The tool provides the possibility of using various word separators on the right and the left side of the words.

- Determine which file types will be identified by the software as metadata corpus files. These types will appear in the interface during the search of the corpus.
- Determine how many and which words will be stored in the results file after the corpus process. This choice can be done either on the basis of their highest frequency or on the basis of their lowest frequency in the corpus (e.g. the first 100 or the last 10), or on the basis of their position in a sorted list (e.g. the 7^{th}, 15^{th} and 50^{th}) or on the basis of a combination of the afore mentioned (e.g. the 1^{st}, 11^{th}, 21^{st}, and 51^{st}-100^{th}).
- Finally, determine if every file described in the metadata corpus file is stored in the hard disc. The control of the existence of the files contained in the corpus is a very important and has been added in order to eliminate errors made during the process. Unfortunately, it burdens significantly the process time.

In addition to the above mentioned, the tool uses two parametric command files which are formatted in a predetermined way.

In the first one, the line delete command file, the user determines which lines will be deleted and will not be analyzed into words, sentences or any other linguistic unit.

In the second one, the text portion delete file, the user determines which characters set will be deleted from the processed file. The user creates the command file by simply writing these words in this file. Moreover, the user can define two sets of characters: the beginning set of characters and the end set of characters. If the two sets of characters belong to the file to be processed, then the text between them as well as the sets of characters themselves will be deleted. In this way, the user can delete a big part of the useless part of the text. An example of character sets of this type (i.e. 'from'…… 'to') are: '<', and '>'. In this way the software deletes all tags.

4. Future objectives for "Corpus Manager"

Further development of the Corpus Manager aims to add more functionality such as:

1. Incorporation of concordance display of the searched words (the word is presented in the centre while parts of the sentence are displayed on the right and on left of it).
2. Addition of new statistical analyses.
3. N-grams search (e.g. the 100 more used 3-grams).
4. Provide graphical user interface in order to view specific statistical analyses (e.g. word dispersion).

References

Anthony, L. (2004) AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning pp. 7–13.

Biber, D. (1993). Using register diversified corpora for general language studies. Computational Linguistics, 2: 219–41.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system". In Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research (July 7–10 1994). Budapest, Hungary, pp 23–32.

Holmes-Higgin, P.R., Ahmad, K. and Abidi, S.S.R. (1994). A Description of Texts in a Corpus: 'Virtual' and 'Real' Corpora. In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg and P. Vossen (eds.), EURALEX'94: (Proc. of the 6th EURALEX International Congress on Lexicography), Amsterdam, The Netherlands, pp. 390–402.

Scott, M. (1996) Wordsmith Tools, Oxford: Oxford University Press. ISBN 0-19-458984-6.

Tweedie, F. and Baayen, H. (1998) How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32: 323–52.