Computer processing and quantitative text analysis: HYPERBASE, an interactive software for large corpora

Étienne Brunet, Xuan Luong

Computer processing and quantitative text analysis: HYPERBASE, an interactive software for large corpora

Etienne Brunet and Xuan Luong

ABSTRACT. The Institut National de la Langue Française is at the head of the largest linguistic data-base available in the world. This paper is devoted to the description of software that makes possible the decentralised exploitation of the data-base without the need to use any data-transmission network. This software, called *hyperbase,* is written in the *hypertalk* object-language. It implements the methods of *hypertext* and emphasizes two main orientations, respectively towards documentation and towards statistics.

RÉSUMÉ. L'Institut National de la Langue Française dispose de la plus importante base données linguistiques qui existe au monde. On s'emploie à décrire ici un logiciel qui en permet l'exploitation décentralisée, sans recours à la télématique. Ce logiciel (HYBERBASE), réalisé dans un langage-objet (HYPERTALK), met en oeuvre les méthodes de l'hypertexte et privilégie deux orientations dont l'une est documentaire et l'autre statistique.

Do all roads lead to the CD-ROM[1] ? At any rate, a long academic tradition leads there. For over twenty years scholars concerned with the application to documentary research of both the resources of the commputer and the methods of statistics have multiplied indexing or concordancing programmes and frequency dictionaries. Within the *Institut national de la langue française* itself, several such achievements have come about, using various programming languages, *PL1, Cobol, Fortran, Pascal, C.* The principle is simple: once the text has been recorded page after page, a direct-access file is created in which each record contains one line of text - or a paragraph or any previously defined unit of discourse. At the same time, each word is identified and stored with its references (text code, page number, key to line or segment, etc.) in a so-called *reverse* file that has been suitably sorted and sifted. Afterwards, various

[1] - Is it necessary to translate the acronym, which refers to the optical reading compact disc (Compact Disk Read Only Memory) as opposed to the Worm (Write Once Read Many times) medium ? Both products share laser technology but only the latter makes writing possible. When the device is associated with magnetic recording (optical-magnetic technique), their advantages are combined, making available both laser's large capacity (a billion characters) and the magnetic medium's own recording and deleting facilities.

possibilities are open, depending on whether they go through the lemmatization[1] stage, whether they take into account the combinatory possibilities of words, of their locations, of their contents, or whether they address questions of syntax or semantics.

These now common products have been very useful. Yet, they have suffered from serious handicaps such as the various forms of the data, the too narrow specialization of the software and the inevitable inadequacy of the former to the latter. They lacked at the same time versatility, power, speed, reliability and standardization.

- I-

Those qualities required by the community of scholars are present in the *Frantext* data-base, that the *Institut National de la langue française* made available two years ago to French and foreign scholars in linguistics or literature. Some specific software, created by Jacques Dendien under the name *Stella,* enables one to know the context of any word or phrase in any text -or set of texts- in the base. Let us dwell here on some advantages that are to be found nowhere else together in applications of this type :

1 - The use is conversational, whereas the packages employed sofar, such as *Cocoa, Jeudemo, Lexicloud*[2]*,* were of the batch type. In the present software the information is distributed in a few seconds.

2 - All that is required is a terminal in a university library, without costly equipment or any technical knowledge.

3 - Yet the scholar has access to very large masses of data, since the Frantext base holds 150 million words, i.e. nearly 3000 complete texts of French literature from 1700 to the present day.

4 - But the scholar needs in no way to feel overwhelmed by such a mass of data for he can choose as he wishes the particular corpus in which he is interested by specifying various criteria such as the date, the genre, the author, the title, or even by imposing context-related constraints. For instance, he can select the works of Voltaire or the whole of XIXth Century novels or the texts containing the word "Revolution".

5 - Just as the scholar chooses his corpus, he also chooses the areas of his research , which can bear on individual forms, lists of words, cooccurrences, phrases. A very large range of combinations that call upon Boolean operators, makes it possible to outline the question with all the desirable precision.

[1] One describes as *lemmatization* the bringing together of forms that belong to the same dictionary entry. Thus, the forms *va, allons, irai,* once lemmatized, are brought under the head-word *aller.* Lemmatization is less urgent in moderately inflected languages such as English.

[2] *Jeudemo* is a University of Montréal product (F. Ouellete creation*). Cocoa* is also known as *OCP* (Oxford Concordance Programme), a Susan Hockey creation. as to *Lexicloud,* it is a package developed at the Saint Cloud Ecole Normale Supérieure by the *URL 3* (A. Salem and P. Lafon). The product mentioned last has original properties, as it accounts for the combinations of words and repetition and association phenomena.

6 - Last but not least, the scholar can specify his own preferences as to the nature of the results - contexts, indexes, frequencies - and their final layout. He can at any moment print them, cave them or interrupt their display.

II

The *Frantext* is the perfect example of what telematics can contribute to literary and linguistic matters. It has no equivalent in the world in this field.

Yet, the evolution of documentary techniques imposes on one to go beyond the present achievement, however remarkable. The compact-disc market, a long time hesitant, is now growing apace in the United States and in Japan, and offers an interesting alternative to telematics. The move is spreading to Europe and at the present time there are fifty documentary projects using this technology which are in progress in France, an example being that recently presented by Emmanuel Le Roy Ladurie for the *French National Library* [1].

The *Institut National de la langue française* has long been interested in the possibilities offered by optical memories in terms of storage capacity and ease of diffusion. It is now preparing a portable version of the totality of its database, which will allow transfer to new systems and new media, including the optical one.

In order to pave the way for this vast enterprise, a prototype has been experimented and is now running, performing the chief functions of *FRANTEXT* . *It* is the subject of the present paper.

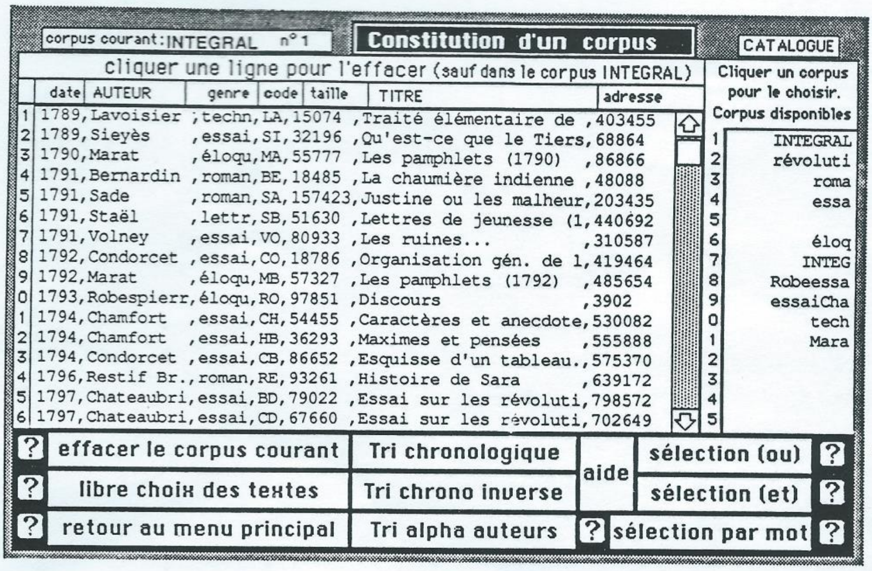
1 - A search algorithm has been developed and implemented, taking into account CD-ROM dependent constraints, particularly the relatively slow disc access, which for any exploration restricts to two the types of movements of the reading-head. Accesses themselves have been optimized (on the basis of the supposed frequency of requests).

2 - The required speed has not entailed the sacrifices common in documentary products, such as their neglect of function-words. Those words on the contrary are nearly priceless in a linguistic type of research aiming at exhaustivity. The reverse file contains all the words and also includes narrative markers, particularly punctuation marks.

3 - To speed and versatility, the HYPERBASE software adds flexibility and precision of search. The user is given a choice of both data and types of processing. He alone determines the sub-set of texts he wishes to consult -by means of a CHOICE OF CORPUS programme which multiplies selection criteria such as genre, author, date, title, contents and applies to them Boolean operators. See figure 1. He also freely draws up the list of the words he is interested in - CHOICE OF WORDS programme -

[1] *PC Informatique*, september,1 , 1988, n° 46, p. 18.

Figure 1. Choice of corpus



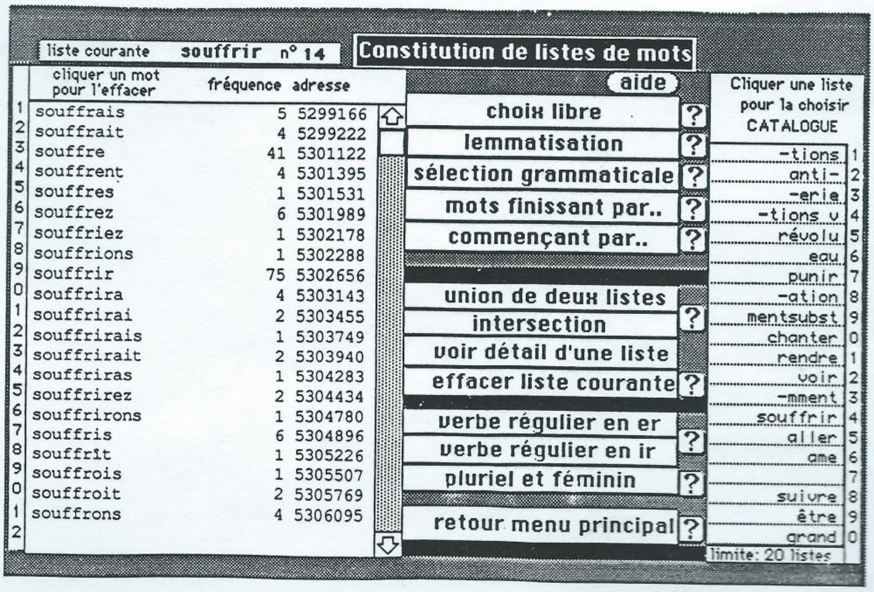
cliquer une ligne pour l'effacer (sauf dans le corpus INTEGRAL)

```
  date AUTEUR      genre code taille  TITRE                    adresse
1 1789,Lavoisier ;techn,LA,15074 ,Traité élémentaire de ,403455
2 1789,Sieyès     ,essai,SI,32196 ,Qu'est-ce que le Tiers,68864
3 1790,Marat      ,éloqu,MA,55777 ,Les pamphlets (1790)  ,86866
4 1791,Bernardin ,roman,BE,18485 ,La chaumière indienne ,48088
5 1791,Sade       ,roman,SA,157423,Justine ou les malheur,203435
6 1791,Staël      ,lettr,SB,51630 ,Lettres de jeunesse (1,440692
7 1791,Volney     ,essai,VO,80933 ,Les ruines...         ,310587
8 1792,Condorcet ,essai,CO,18786 ,Organisation gén. de 1,419464
9 1792,Marat      ,éloqu,MB,57327 ,Les pamphlets (1792)  ,485654
0 1793,Robespierr,éloqu,RO,97851 ,Discours              ,3902
1 1794,Chamfort   ,essai,CH,54455 ,Caractères et anecdote,530082
2 1794,Chamfort   ,essai,HB,36293 ,Maximes et pensées    ,555888
3 1794,Condorcet ,essai,CB,86652 ,Esquisse d'un tableau.,575370
4 1796,Restif Br.,roman,RE,93261 ,Histoire de Sara      ,639172
5 1797,Chateaubri,essai,BD,79022 ,Essai sur les révoluti,798572
6 1797,Chateaubri,essai,CD,67660 ,Essai sur les révoluti,702649
```

Cliquer un corpus pour le choisir. Corpus disponibles

```
1  INTEGRAL
2  révoluti
3      roma
4      essa
5
6      éloq
7     INTEG
8  Robeessa
9  essaiCha
0      tech
1      Mara
2
3
4
5
```

Figure 2. Choice of words



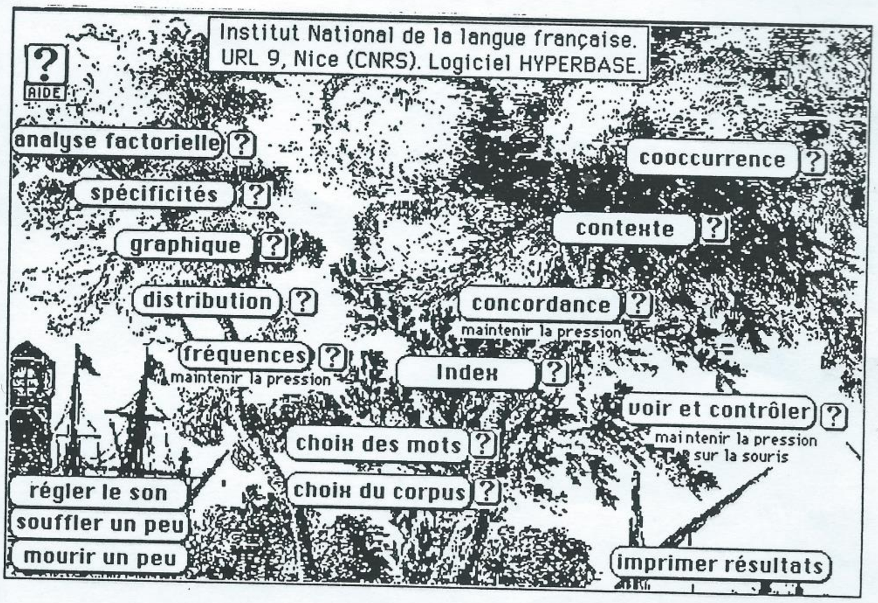
```
   cliquer un mot
   pour l'effacer      fréquence adresse
1  souffrais                  5 5299166
2  souffrait                  4 5299222
3  souffre                   41 5301122
4  souffrent                  4 5301395
5  souffres                   1 5301531
6  souffrez                   6 5301989
7  souffriez                  1 5302178
8  souffrions                 1 5302288
9  souffrir                  75 5302656
0  souffrira                  4 5303143
1  souffrirai                 2 5303455
2  souffrirais                1 5303749
3  souffrirait                2 5303940
4  souffriras                 1 5304283
5  souffrirez                 2 5304434
6  souffrirons                1 5304780
7  souffris                   6 5304896
8  souffrît                   1 5305226
9  souffrois                  1 5305507
0  souffroit                  2 5305769
1  souffrons                  4 5306095
2
```

Cliquer une liste pour la choisir CATALOGUE

```
    -tions 1
     anti- 2
     -erie 3
  -tions v 4
    révolu 5
       eau 6
     punir 7
    -ation 8
mentsubst  9
   chanter 0
    rendre 1
      voir 2
    -mment 3
  souffrir 4
     aller 5
       ame 6
           7
    suivre 8
      être 9
     grand 0
limite: 20 listes
```

offering automatic lists based on the selected suffix, prefix or conjugation and allowing additions, deletions, overlaps. See figure 2. Finally, he chooses from the menu the tasks to perform, which can lead to an index - INDEX programme -, to a concordance - CONCORDANCE programme -, to a list of phrase-contexts -CONTEXT programme -, or to cooccurrence searches - COOCCURRENCE programme -. He is totally free at any of those stages to supply a single word, a string of words, i.e. a phrase, or a list of words. See figure 3.

Figure 3 : The main menu



Institut National de la langue française.
URL 9, Nice (CNRS). Logiciel HYPERBASE.

4 - The user can also conduct control operations and examine the texts, page after page -SEE A TEXT programme-, or consult the word file, in alphabetical order or in order of decreasing frequencies -SEE SOME WORDS programme- or pass instantly, by means of a simple "click" from the texts to the words or the words to the texts, according to the methods of hypertext. Should he wish to check the contents of the current list and corpus -which are variable and can be altered at will- a rapid overview will flip through the bibliographical slips of the texts in the corpus or the identity slips of the words in the current list.

The HYPERBASE software also leads towards quantitative methods. It can draw up contingency tables -FREQUENCY programme-, and produce graphs comparing the distributions of two words in the various texts, or the profiles of two texts with respect to a set of words - GRAPH programme-. There is also an interface for exploiting such tables by the methods of multidimensional analysis -FACTOR ANALYSIS programme-.

Finally, various modules are available that display the distribution of frequency-classes within each text and in the whole corpus -DISTRIBUTION and ZIPF LAW buttons- or that list the specific vocabulary of a text or the specific properties of a word -SPECIFICITY button-.

6 - Lastly, the **results** appear under two forms : on the **screen,** at the very moment when they are obtained, and in a **file** where they are saved before being checked and printed. Both forms are not necessarily identical, the screen alone allowing the display of the complete text. **The** results then undergo various sorting operations and yield, among other things, concordances that corne in various guises : chronological order, reverse chronological order, authors' alphabetical order, sort on the left-hand or right-hand context. An editor -complete with printing module- is supplied which facilitates corrections and layout. Let us note that the results themselves are not the end of the line and that they are amenable to the methods of hypertext : it is for instance enough to point to a line in the concordance to make the corresponding page of the text instantly appear.

7 - All the software has been written bearing in mind **user-friendliness.** Even if it is meant for a specialised public, it has not been given a stern countenance. On the contrary, one has attempted to exploit the **graphic** potentialities of the screen -for instance, the portraits of the writers or the fac-similes of the original editions are displayed-, and full use has been made of the operating facility of scrolling menus or of control buttons and the flexibility of use of the "mouse". Sound itself is welcome and **help** has been provided for each of the functions at the very moment when it is offered : it is enough to activate the question mark that comes with each button.

8 - This concern for user-friendliness and versatility explains the choice of an **object-language** for the whole of this software written in *HYPERTALK* and complemented by external functions and controls. As it is an interpreted language and a modular process, the software is open and admits an kinds of complements.

III-

1 - Some complements were devised at a later stage. One has endeavoured to supply each word with a grammatical tag -one is even thinking of semantic tagging- in order to make possible the study of grammatical categories, the disambiguation of homographic forms or thematic research. One has taken the option of automatic lemmatization, which opens the way to the treatment of lexemes as opposed to mere forms. One can join together all the forms connected with the same

[1] As a matter of fact, large portions of the processing have been compiled in order to gain both speed and reliability. One has not given up hope of compiling the whole, when the HYPERTALK language, yet too young, has been stabilized and an efficient compiler is available for it.

dictionary entry, either to draw up a list in the CHOICE OF WORDS menu, or to produce the concordance of a given lexeme, or even to find out what lexeme a form is connected to and what other forms are likewise connected to it -LEMMATIZATION button available for each form-. The new version of the programme also supplies for each word -at least for each lexeme- the frequencies observed in the entire TLF corpus and in the Revolution corpus. This therefore makes a very wide backdrop against which the specificities of the 1789 corpus stand out very sharply (z-scores are calculated as soon as permissible).

2 - In its present state, the prototype exploits a base of 24 complete texts of the revolutionary period, from 1789 to 1800. Robespierre rubs shoulders with Sade, Siéyès with Marat, Madame de Staël with Chateaubriand, and Chamfort with Condorcet. If the corpus gives pride of place to essayists -among whom Volney, Marmontel, Bonald-, it does not exclude novelists -for instance Restif de la Bretonne, Madame Cottin or Bernardin de Saint Pierre- or dramatists -Pixéricourt, Lemercier- or scientists -Lavoisier, Monge, Lagrange-. In all, a total of 1 300 000 words are available.

The HYPERBASE software and the associated data have been made accessible to the public of the Centre Georges Pompidou (Beaubourg) as part of the celebrations of the bicentenary of the French Revolution, with the backing of the Apple company who granted the free loan of a MAC IIX. The version on show -which cornes in colour- works with a 40 Mo hard disk. The base has also been stored on a WORM compact disc, in order to experiment laser technology. But the CD-ROM version is not planned for the immediate future, because the HYPERTALK language does not seem to be stabilized and because great improvements are expected in the coming months as far as confidentiality and speed[1] are concerned. The HYPERBASE software being under development contract (Apple company), we can hope to be among the first to reap the fruit of future improvements.

3 - As is usual with producers and distributors of data-bases, the interrogation programme is inseparable from the processed data.

a - But copyright and confidentiality problems have led us to envisage offering an empty software, devoid of any data. The user would enter his own texts. It remains to supply him with a chain of data-preparation

[1] We have known four versions of this language within one year. The example of SUPERCARD shows that the handling of multiple windows, of large screens, of colour, is not impossible and that the advantages of compiling can be expected in the near future.

programmes[2]. Such programmes exist since they have made possible the processing of the Revolution texts. But it remains to endow them with the required universality, speed, versatility and reliability[3] .

b - There are also plans for a version of the software that might not display the entire text, when the privileges of copyright oppose any direct publication. The original text would indeed be used to supply the answers, but once the answers given, the text would be, as it were, spirited away. This amounts to the opposite of the present HYPERBASE procedure which waits for the user's questions and looks for the answers in real time. The enormous possibilities of the CD-ROM make it possible to store in advance all foreseeable answers, for instance the concordance to all the words in the corpus and to restore them at the time of questioning, when the original text has been made to disappear. As a matter of fact this product has a fair likeness to what was done in the past using other media such as paper or microfiches. The advantage of the CD-ROM however would be to allow direct access without any manipulation whatsoever. By combining several modes of display of the results -based on preliminary sorts- one might offer various options. Also, by playing on the fields of reference, it would become easy to filter the results and to facilitate their selection.

In short, the two versions envisaged here lie on either side of the data-base watershed. The former is richer, perhaps more exciting because it makes the pleasure of discovery more vivid in the presence of the entire original text. The latter is well-mapped, clearly sign-posted. The routes have been carefully charted and travelling them is quicker. It is the field of structured data bases.

[2] Is it useful at this point to go into the detail of the various preparatory phases ? Let us simply say that after the checking operations the texts are segmented into paragraphs and pages and recorded before creating the reverse file and the access-keys -the so-called search algorithm-. Let us specify that these programmes are written in a conventional language, Pascal or C, and that at certain stages of the treatment bridges will be provided towards Hypercard, particularly for knowing the physical address given by Hypercard to each of the created cards.

[3] If a bare software is offered, it is probably not necessary to plan an optical medium, at least not the CD-ROM, whose manufacturing costs cannot be justified for a single copy -which is often the situation of the scholar in the humanities-. If the WORM system can perhaps be suitable, until the optical-magnetic technique is available, the hard disk is the natural medium for such applications. One becomes convinced of this when one looks at the market : the available products all Lean towards the hard disk. It is for instance the case of *Micro-OCP*, created by Susan Hockey, Oxford Computing Service, of *TACT* (Thematic Analysis of Computer-readable Texts), written by John Bradley and Lidio Presutti, Centre for Computing in the Humanities, University of Toronto and of *WordCruncher* (Electronic Text Corporation, Provo, Utah), even if this well-known software can now accept the CD-ROM.