What Do Statistics Tell Us?

Étienne Brunet

What Do Statistics Tell Us?

ETIENNE BRUNET

Should we regard statistical tables as the Tables of the Law, or as seance tables? If they do tell us something, should we believe them? How much credit should we accord to the medium, i.e. the statistician, who speaks in figures just as Hugo's medium spoke in verse at the seances held in Guernsey? In modern societies, religious wars and ideological conflicts have been replaced by a battle of the figures. Often the same statistics are employed on both sides, rather like the mercenary soldiers of former wars, to such an extent that if Plato were writing Gorgias nowadays, he would make rhetoric and sophistry amenable to figures, and turn Callicles into a statistician.

Numbers appear to have a divinely ordained superiority over words because they give the impression of conveying absolute numbers. But this apparent incontrovertibility, however impressive, often conceals relative and contingent procedures that have nothing essential about them. Just as nature rarely reproduces the same object, or time the same event, or speech the same words, or words the same meaning, so the facts out of which one produces a statistical result are not strictly identical. They only become so abstractly, by neutralizing the particular conditions that gave rise to them. They can only lend themselves to comparison if it is assumed that conditions never vary, and if the unit of measurement remains constant, which is rarely the case. Thus we are forced to hedge, to modify the absolute by making adjustments (for example, economic data may be 'seasonally adjusted', or opinion poll results may be corrected to take into account both those questioned and the pollsters themselves).

1. Statistical Linguistics

1.1 *A Unit of Measurement*

These difficulties, which are inherent in all human sciences, place a burden on statistical linguistics. What this discipline lacks is primarily a unit of measurement. The delimitation of a word itself is uncertain, even if one adopts the most simple graphical definition of the word as the printed space between two blank spaces or two separating markers. The problem is that many punctuation marks do not simply have a separating function; this is the case with the full-stop, the apostrophe, and above all the hyphen, which itself

denies the function of separation. Typographical separators are no more certain when we use them to define larger units of text, like sentences or paragraphs. What status should we give to expressive punctuation marks (interrogation, exclamation, suspension) when their meaning varies according to the context? To say nothing of intermediate signs (semi-colon and colon).

Furthermore, these markers, already ambiguous in themselves, are of questionable authority, because they often originate with the editor and not the author. Another problem is lemmatization, which attempts to standardize the multiplicity of forms in inflected languages, but which itself has never been standardized. For want of a proper yardstick, the statistical linguist examines texts in much the same way as an old-fashioned surveyor measures fields by counting his steps.

1.2 *Terms of Reference*

This discipline has for a long time lacked terms of reference. Statistical studies are essentially comparative, like experimental sciences. Measurement itself implies comparison, at least with a yardstick. But what can be compared? Can we compare two different languages, two widely separated periods, two opposing genres, two foreign writers, or two texts which differ in one or more of these respects?

This tricky situation was avoided by the first exponents of the discipline, precisely because the field was so new, and no previous studies suggested points of orientation or comparison. The pioneers often simply noted their observations, without generalizations or conclusions.

Many more elements are available today, but this superfluity of choice goes hand in hand with a paralysing logical impasse: before risking a comparison with someone else's results, it is essential to ask not only whether they are honest or reliable, but also whether they are compatible with the ones it is hoped to compare them with, and whether the principles of selection and processing are the same. Given measurements based on equal good faith but using different options, statistics can lead to conclusions which are rash or presumptuous.

1.3 *The Aim of Linguistic Statistical Studies*

The real aim of linguistic or lexical statistical studies has not been clearly defined; the very hesitation over the adjective ('linguistic' or 'lexical') illustrates the difficulty in delimiting the discipline. It is in fact easier to enumerate what is lexical, but words conceal hidden morphological, syntactic, rhythmic, and thematic features. Far from being an elementary unit, a word is a complex reality whose infra-lexical components (suffixes, prefixes, verbal inflec-

tions, phonemes and morphemes of all kinds) enable us to move on to wider, supra-lexical components.

Furthermore, a word is not an independent unit. The most important feature of a text is the way words fit together. By breaking the chain, it might be said that statistical study destroys its object, as happens in nuclear physics when the light projected on to elements disturbs the very movements one wishes to study.

1.4 *Problems*

Despite these logical difficulties we still look to statistics to solve the thorniest problems, such as those of the dating or attribution of texts. We have ingenuously asked the computer to provide scientific answers, assuming that the machine is endowed with an infallible memory, unwavering attention and irreproachable impartiality. The expectations first raised by these machines were like those raised by Carbon 14 as a method of dating prehistoric sites, or fingerprints for solving a murder. But where do we find the author's fingerprints? Where, in a tangle of words, can we discern the invisible but unmistakable and indelible signature of the writer?

The reasoning underlying the discipline generally ranges between two tendencies. First of all there is the classical method (for example, problems of attribution and dating), which consists in first advancing a hypothesis and then putting it to the test by considering the facts. Here the danger lies in failure, or, more exactly, the denial of failure. A risk of statistical obstinacy arises when testing does not provide the expected result, and the researcher does not want his efforts to go to waste.

The opposite method, unassuming but naive, consists in having no preconceived idea—therefore no hypothesis—and in observing every kind of disparity in the calculation. In writing, nothing is improvised, all is determined. There are oppositions, separations and discrepancies everywhere, but the same effect can have multiple causes. If I observe a variation in the form *aimait*, it could be due to the tense, the mood, the person, the number, or the sonorities of the word, or to its semantic content. Finding a deviation is hardly an advance; the cause needs to be determined, which is like searching for something one has already found. This situation is no more favourable *than the classical method, which tends to find and to prove what it is looking* for.

1.5 *Methodology*

What statistical studies lack as well is a consensus on methodology. At first statisticians were content simply to produce raw data, and when weighting appeared to be necessary, the data was transformed into simple percentages. But this rather unambitious method did not satisfy them, since they aspired

to enter the realm of probability. Is this aspiration a legitimate one? It would only be so if the choice of words in a text was determined by sheer chance (the urn model). Now even the promoters of the probability method have to admit that a writer does not pick his words out of an urn, and—except by way of surrealist experimentation—he does not choose them by sticking a pin in a dictionary. There is generally a great disparity between observed facts and the probability model, to the extent that the exception becomes the rule. The modest scale of the initial results hid this distortion for a time. When the numbers are low, it becomes more difficult to reject a null hypothesis, and there are fewer so-called significant results which cross the threshold level, which gives the illusion of a purely arbitrary choice. But as soon as larger numbers are involved, the landmarks in the world being described become manifest, and words acquire a significance that owes nothing to chance.

This statement, which reassures the literary researcher that texts have an irreducible meaning, worries the statistician and encourages him to seek another model. Unfortunately no alternative to the classical method has yet been found. We have seen a great number of all kinds of indexes, different quotients, and ingenious formulae, all of which are designed to measure the lexical richness, the proportion of grammatical categories, or the stylistic colouration of a text. But all these attempts lack the generality and coherence, as well as the theoretical perfection, of the theory of chance.

1.6 *The Reception of Linguistic Statistical Studies*

It is nothing new to find statistical studies on trial. Many aspects of this trial are to be found in the archives of the ALLC itself; the case was opened notably at conferences held in Pisa and Louvain, and, in Canada, in Waterloo and Victoria. But statistical linguistics has little to fear from this particular trial (for there is a certain amount of complicity between both sides of the argument). The reception given by the literary and linguistic community to statistical reports is more worrying. Outside the narrow circle of specialists and their own personal jealous or self-interested approval, it must be said that literary observers regard statistical debates sometimes with amusement, more often with indifference, and in certain cases with anger and indignation. At the opening of the ALLC conference of 1985, Charles Muller gave an amusing but by no means indulgent analysis of the reception of statistical studies by the public. The conclusions reached by this pioneer of the discipline were hardly optimistic: 'Confiance excessive d'un côté, méfiance abusive de l'autre . . . La cause est bien compromise!' (Muller 1986).

1.7 *Results*

In reality this reservation on the part of the public often takes the form of evasion, since statistical results are more often ignored than condemned.

Graecum est non legitur say the literary wits when faced with a mathematical formula, however childishly simple. They have even more reason to show distaste for those endless lists whose head and tail one can never see at the same time, for those bristling tables of figures, and for those piles of indexes and concordances which lie idle and uncalled-for. It must be admitted that these trawler-nets of 'results' are frequently off-putting. Often their author hardly plays any part himself; he amasses material which he submits in its raw state, or barely transformed. It is true that in the past it took a lot of effort to obtain results, and what ought to have been just the initial stage became the sole purpose of the research. But can we blame linguistic statistics for results that are not interpreted by the person who drew them up, lists that are not analysed, and concordances that are not exploited? Is it not simply a service that provides documents?

2. FRANTEXT

At the present time the collection of data is no longer so wearisome, at least as far as French is concerned, because in Chicago there is a distribution centre for the FRANTEXT database. This is a rich treasure-house, containing 160 million words, as well-organized as it is abundant. The creator of this database, Jacques Dendien, is better qualified to explain its function and describe its merits; suffice it to say that from all parts of the world it is possible to summon up by cable link any word, expression, or association from 3,000 complete texts available in the database, and it only takes seconds to accomplish, at any time of day or night. Now that the researcher has access to this store, it is to be hoped that there will be less waste, more coherence and commitment, and a better quality of results.

This quality of results depends first of all on strict formatting, that is to say the possibility of collecting together sets of elements sharing certain common properties, so that one can compare what is comparable. The most important factor is the homogeneity of the data and the consistency with which it is processed. The FRANTEXT database gives that guarantee, as all the texts in it have been entered and processed according to norms which have not changed one iota in twenty years. The coherence of the body of words which is selected from the database depends on the researcher, who can choose according to author, genre, or period.

2.1 *Monographs*

The most natural and frequent selection is based on the author, and its aim is to create a monograph, such as those we have produced on Giraudoux, Proust, Zola, and Hugo. In such cases the tendency is to restrict oneself to the

particular world of the writer and to consider the body of his works as the 'norm' for the texts making it up. These texts are then ordered according to the writer's development, unless chronology is affected by the predominant influence of literary genre. Different points of view can be taken in order to explore in turn the features of lexical structure segmentation, syntax or semantics. At the same time the tools used can vary from the most simple (z-score graphs) to the most complex (factor analyses).

In order to give some idea of the results obtained in this way, we will consider several of our works, varying the authors, methods, and the objects of the research. (We have chosen to give these examples because the works themselves, published by Slatkine-Champion, Paris and Geneva, are unlikely to be available to the reader, in view of their high cost.)

Figure 1 deals with the time-honoured question of lexical structure. Ever since Zipf's law, much has been written on the complex mathematical ratio between the number of word-types and the number of word-tokens, which governs the distribution of frequency-classes in a text. Many debates and suggestions about this relationship have arisen, so that for a long time it appeared to be the modern equivalent of the philosopher's stone or the golden section. Literary interest hardly ever goes beyond the appreciation of lexical richness and variety. In the case of Giraudoux, this ratio is favourable to the novels (marked in grey in the Figure), and unfavourable to the theatre; in the plays themselves, modern comedy fares better than ancient and tragic subjects.

Figure 2 illustrates the same stratification of literary genres, although it analyses a completely different feature: the average length of the word. Words are longer in the novel than in the plays, and shorter in the ancient plays than

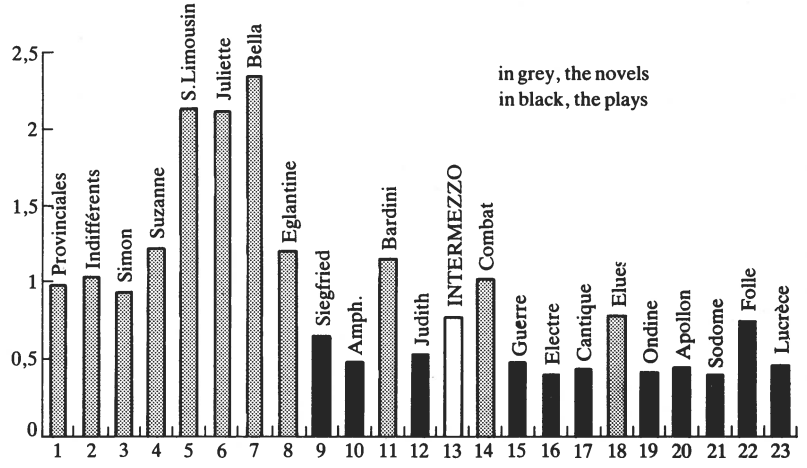

Fig. 1. Lexical Structure in the Works of Giraudoux.

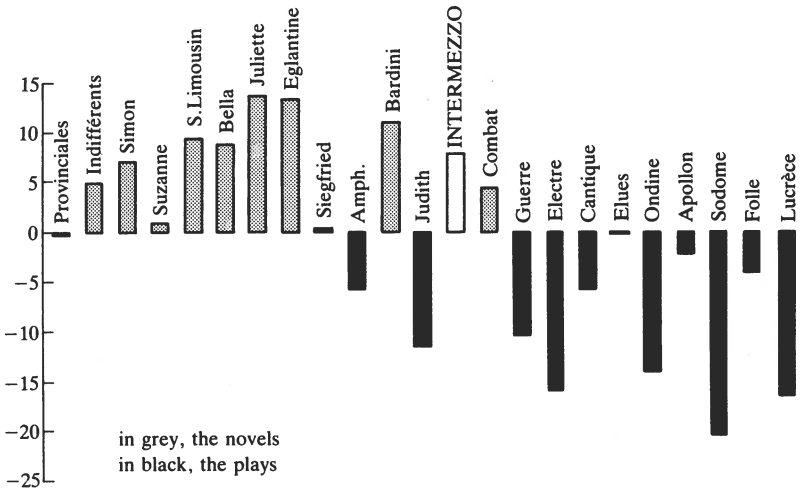

Fig. 2. Average Word Length in the Works of Giraudoux.

in the modern plays, as if these two facts were linked. The dryness of tragedy is incompatible both with an excessive number of words, and with words that are themselves too long.

The body of Hugo's works, like that of Giraudoux, is characterized by the opposition of genres. This is evident in the factor analysis shown in Figure 3, which takes into account the lexical connection (i. e. the distance between the vocabularies of two texts). We will not put the reader off by describing the infinite calculations necessary to obtain this distance by using an overall measurement. It was necessary to consider all the words in the two texts studied, and the frequency (real or theoretical) of each one in each text; and this complex calculation had to be done for each pair, 231 times for 22 texts. The result of the analysis is however perfectly clear: at the bottom are grouped all the poetic collections, while the novels, letters and plays are placed in the upper half, without merging too much into each other (except for the theatre, because of its internal division into prose and verse).

This law governing genres never ceases to surprise us when we look at the author of the *Préface de Cromwell*, where he said that he wished to break down restricting barriers. We also wanted to examine whether there was any development in this writer, who was both precocious and late-flowering, and whose output extends over sixty years.

The same measurement of lexical connection allows a glimpse of the chronology of Hugo's works, as shown in Figure 4. Here, texts are matched by genres, the influence of time producing the diagonal line in the graph.

In principle, chronology reigns supreme in a body of texts where genre oppositions play no part, as is the case in *A La recherche du temps perdu* and

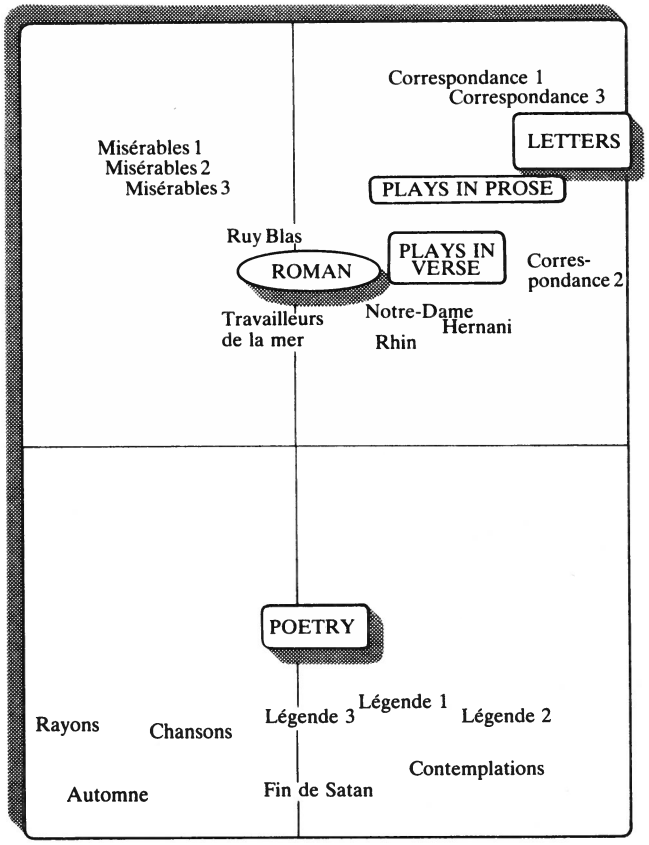

Fig. 3. Factor Analysis of Lexical Connection in the Works of Hugo.

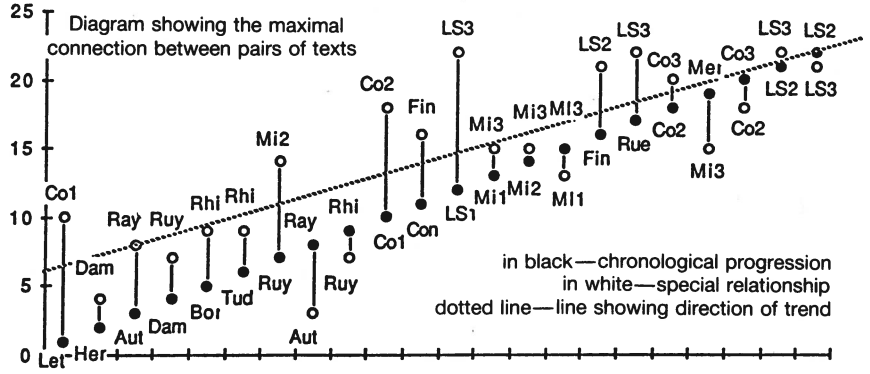

Fig. 4. The Influence of Time on Hugo's Vocabulary.

Les Rougon-Macquart. We will give two examples, taken in fact from Proust and Zola. The first examines the lexical content of the seven books of *A La recherche*, and more precisely two particular semantic fields: the theme of nature and the supremely Proustian theme of time. Figures 5 and 6 have been placed side by side because the graphs are complementary: while nature becomes less prominent during the course of the work, time becomes more and more obsessive, with a more sombre tonality that emphasizes years (*années*) rather than days (*jours*), a lifetime (*vie*) rather than moments, and memories (*souvenirs*) rather than youth (*jeunesse*).

The second illustration, in Figure 7, relates to punctuation in *Les Rougon-Macquart*, which is reliable because it is faithful to Zola's wishes. There are 132, 114 full-stops and 340, 479 commas, and their distribution over the twenty years of the cycle follows two opposing movements. The graph depicting the full-stop shows a surplus in the first half and a deficit in the second, while the comma follows a continuous progression. Zola's sentences, which are structured by his punctuation, expand as he writes more novels.

2.2 *Studies of More than One Author*

Monographs can be linked together if they have been drawn from the same bank of data. Thus Figure 8 gathers together six writers whose data was

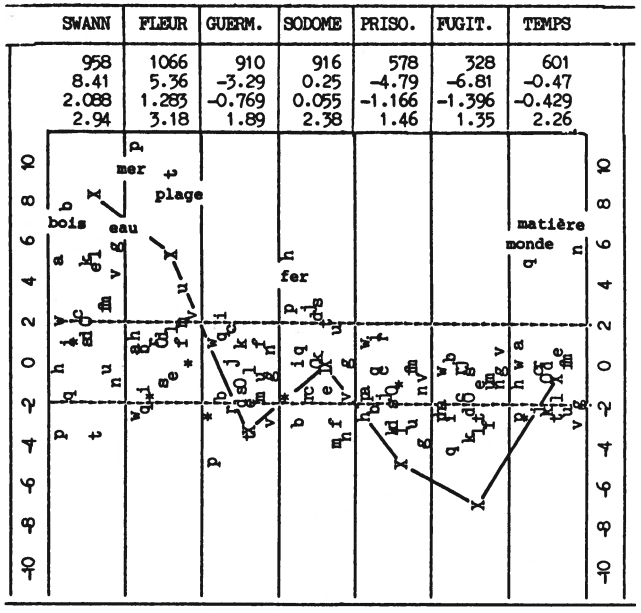

Fig. 5. Nature in the Works of Proust.

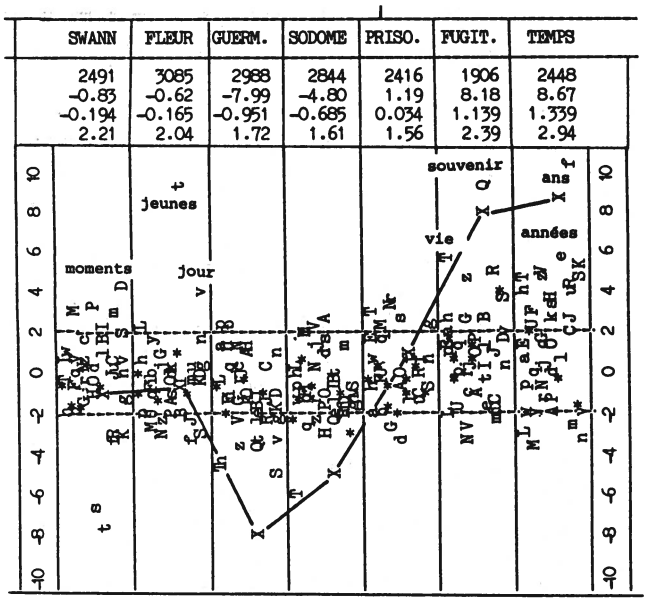

Fig. 6. Time in the Works of Proust.

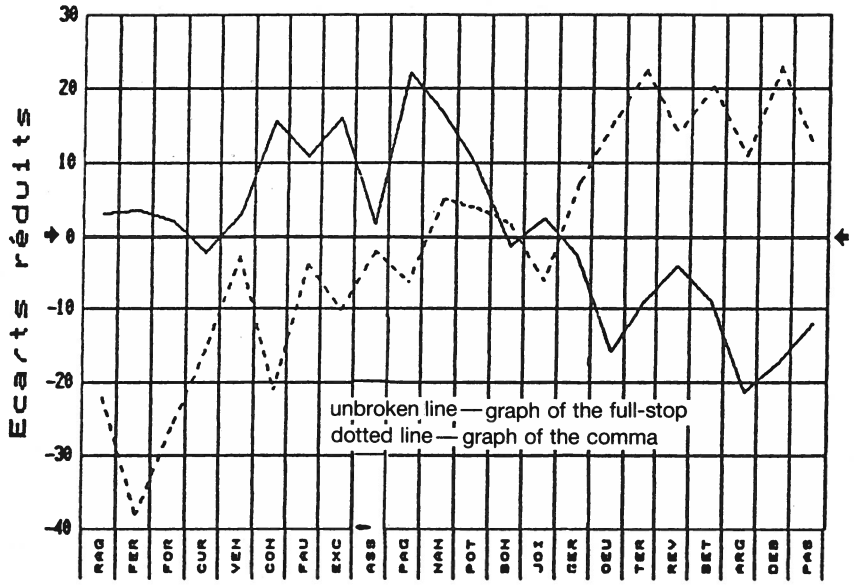

Fig. 7. The Development of Zola's Punctuation in *Les Rougon-Macquart*.

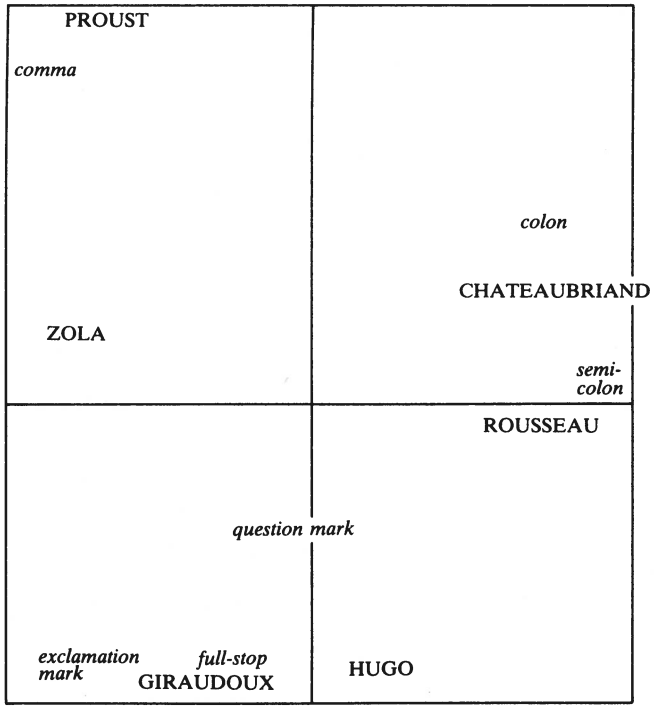

Fig. 8. Factor Analysis of the Punctuation of Six Authors.

jointly processed. It can be seen that over two centuries of literature, the system of punctuation has clearly undergone changes. Rousseau and Chateaubriand are alone in cultivating the intermediate signs (colon and semi-colon), while Hugo and Giraudoux use many full-stops (as well as expressive signs), and Proust employs many commas. As is to be expected, Proust beats the record for the length of his sentences (31 words on average), ahead of Rousseau (27) and Chateaubriand (22). Hugo and Zola prefer shorter sentences (14 words on average).

While monographs on authors remain limited in number, comparisons can be made between each of them and the entire body of nineteenth- and twentieth-century texts. This provides us with a common backdrop against which the specific features of each writer stand out. If necessary one can extract from the entire corpus a sub-group of texts better adapted to the model required, if the dates and genres are selected appropriately. The example of grammatical categories in the works of Hugo (see Figure 9) shows how this can work. The conclusions are almost invariable, whether comparison is made with the entire corpus, or with that drawn from the time of Hugo, or whether it is confined only to the poetry or only to the literary prose of the

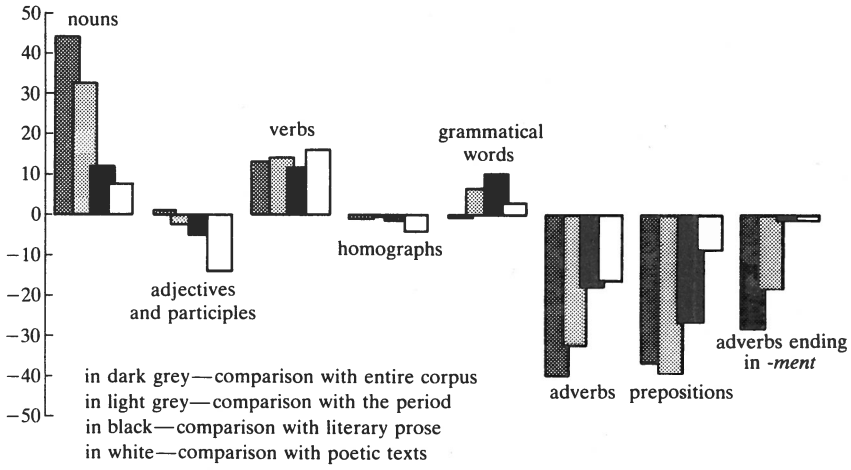

Fig. 9. Grammatical Categories in the Works of Hugo. External Comparison.

time: among the parts of speech, Hugo chooses the most substantial, that is to say nouns, and to a lesser extent verbs, and he has little regard for prepositions and adverbs. Proust's or Zola's choices are not the same.

It is above all in thematic studies that this external comparison can be useful. Calculating and sorting the z-scores (or Chi-squared values) for all a writer's words, produces two batches of words which are equally interesting: those which the author likes more than other words, and more than other writers do, and those he avoids. In this way, Proust's and Zola's lists are the inverse of each other. True to his naturalist creed, Zola concentrates on places, bodies and things, while Proust concentrates on feelings and psychological realities. In Hugo, statistical tests throw light on light itself, and, even more so, on darkness. The word *ombre* appears at the top of the list of nouns, while *sombre* appears at the top of the list of adjectives, and the rhyme *sombre–ombre* is the most frequent. The reader's instincts in this matter accord with the statistics, but this is not always the case: the lists can provide surprises. Thus the vocabulary of sensation and feeling is strangely deficient in the works of Chateaubriand, author of *René* and *Atala*, who shows a classical restraint in using words like *passion, affection, impression, attention, sensation, émotion, plaisir, expression.*

2.3 *Monographs on Words*

Just as there are monographs on authors, there can also be monographs on words, or on larger families of words and themes. Instead of examining all the words of a writer, this time one selects one single word (or a restricted group

of words) throughout all the writers in the database. This approach is the one most often followed by linguists and historians, while monographs on authors are better suited to the needs of literary scholars. For example, in Figure 10, we show a graph depicting the word *imagination*, the frequency of which has decreased since 1789. Does that mean that imagination is drying up nowadays? It would be rash to make that assumption in the age of the image. In actual fact the word *imaginaire* has made up for the deficiencies of *imagination* (the two contrary movements are shown in Figure 10), and the verb *imaginer* has also increased in frequency, like the word *image* itself. In both cases the progression is underlined by a very positive correlation coefficient ($r = +0.93$ and $r = +0.70$ respectively).

The interest in such cross-studies is even greater when one runs through the corpus in search not only of an individual word but a set of words. The case of the word *imagination* proves that conclusions are more nuanced when one looks up other members of the lexical family. It is possible to widen the circle and admit into the set the words *esprit*, *âme*, *coeur*, and several others. In such cases we come across the phenomenon of substitution: fashionable terms like *conscience* or *psychologie* take over from words which no longer please (like the word *âme*), or else substitutions take place among grammatical categories, the adjective taking the place of the noun in an expression like 'mental faculties', no doubt because we no longer dare to represent them as substances.

A wider application of FRANTEXT is the ability of the LISTEMOTS command to summon up sets of words which can be as extensive as one wishes, and to define wide semantic areas, such as animals in literature, or the system of kinship, or the representation of the body, or space, or time. We shall choose a curious example which has rarely been the object of statistical research: the examination of proper names, or more precisely the Christian names of the Roman calendar (about 300 of them). Since the seventeenth and eighteenth centuries are now available in the database, without posing in this case problems of spelling or lemmatization, Figure 11 is able to represent four

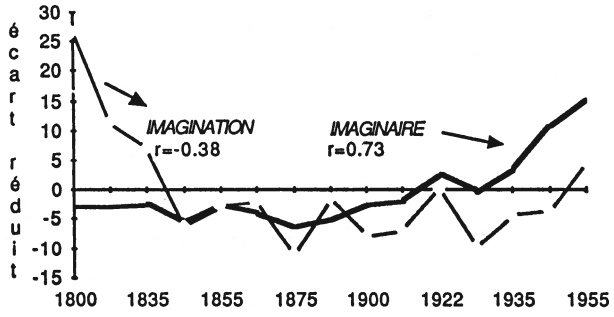

Fig. 10. The Words *imagination* and *imaginaire*.

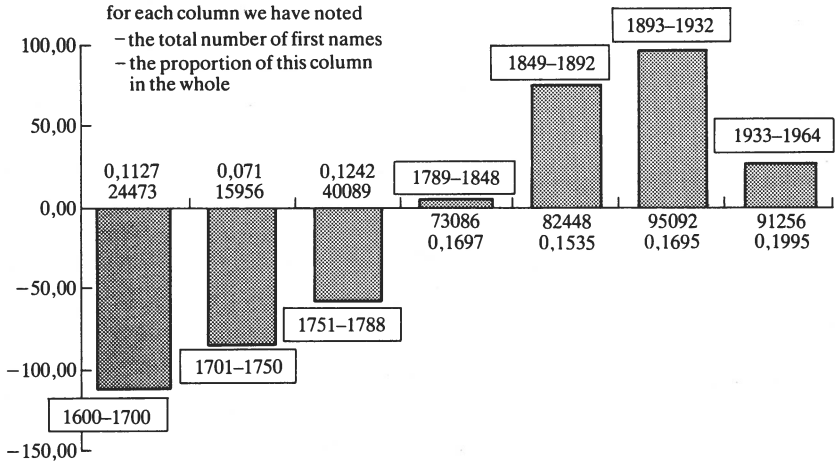

Fig. 11. First Names from 1600 to the Present Day.

centuries of literature. It is admitted that the raw material is composite, and that these names often designate places as well as real or fictional characters, ancient and modern. It remains none the less true that their numbers (nearly half a million occurrences) are distributed very clearly, showing a continuous progression from 1600 to 1900. Doubtless the modern era, since the Revolution and the period of romanticism, has given more prominence to the individual and to the specific nature of beings and places, of which the proper name is the symbol.

We considered the proper names en bloc, but this global approach can be refined. We could consider each line of the table (each proper name on the list), or each column (each of the seven periods making up the corpus), in order to establish individual profiles; or else we could study the entire table by means of factor analysis. One example will suffice. It is the first, chronologically speaking, since it concerns the pair *Adam* and *Eve*, as shown in Figure 12. (But these are not the first in order of frequency; that palm is reserved rather for the saints and martyrs, namely *Jean*, *Louis*, *Jacques*, *Charles*, *Marie*, *Pierre*, *Henri*, *Antoine*, *Paul*, *Philippe*). *Adam* and *Eve* were rehabilitated during the nineteenth century, when *Eve* in particular gained an advantage. In actual fact, *Eve* profits from a general movement which,during the nineteenth century, redressed the imbalance between the sexes (see Figure 13). The ratio between male and female first names, which was of the order of 4:1 in the seventeenth century, was corrected during the the eighteenth and nineteenth centuries to reach a proportion of 2:1. But this promotion of women seems insecure since the proportion returns to 3:1 in the most recent column, and out of the whole, *Marie* is the only female first name to feature among the top ten.

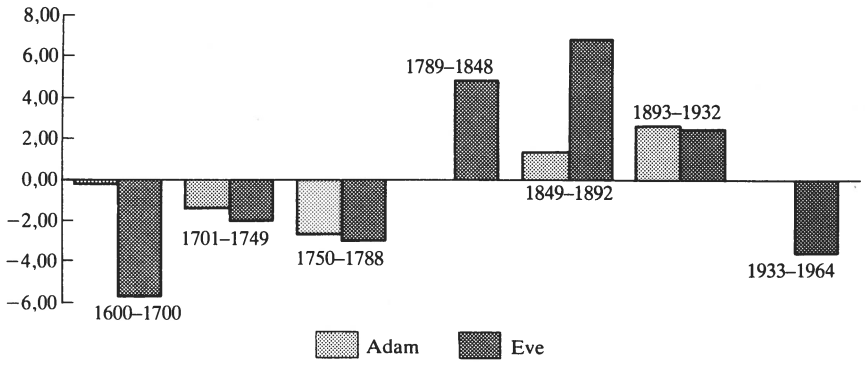

Fig. 12. The Pair *Adam* and *Eve*.

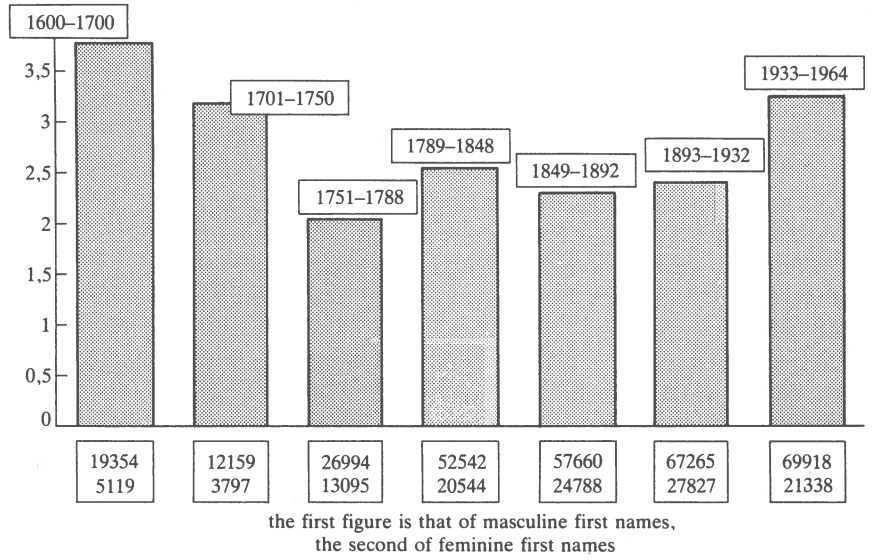

Fig. 13. The Ratio between Masculine and Feminine First Names.

2.4 *Future Developments*

It might be possible to go even further and study the FRANTEXT corpus in its entirety, looking at every word, every text, every author, every genre, every period. This ambition is really incompatible with the restraints of present technology, and furthermore it would come up against the absence of lemmatization in the available data. This general study can only be envisaged in off-line, batch-type processing, which we implemented in 1981, at a time when

the data only covered the last two centuries. It could be resumed to great advantage if the FRANTEXT data were available on an optical disk that would allow the transfer of texts, so that data could be processed locally. A large CD-ROM project along these lines is at present being developed under the supervision of Jacques Dendien.

3. CD-ROM

Times have changed with regard to documentary and statistical research. In the past, many indexes and concordances were drawn up that were only of use to their author, and some manufacturers presented their products just as they were, without any applications. They took a long time to produce, their distribution was restricted and their use was cumbersome. When information is found in a form not directly accessible to the computer, like paper or microfiche, it is difficult to use, particularly if one wishes to cut it or add to it; in a word, to edit it. Such applications are of practical necessity to statistical exploitation. Furthermore, options made at the time of development can be restricting (e. g. sorts, rejects, groupings, lemmatization, context-length). At least the power, speed, extent, and versatility of FRANTEXT represented real progress on that front.

Documentary and statistical computing is tending towards even more versatility, by breaking out of the confines of present technology. What the scientific community wants now is easy, immediate and free access, not just to such fixed by-products as indexes, concordances, and frequency directories, but to the text itself. The user may desire the help of interrogation software, but he wants to choose his own questions and texts, as well as the layout of his results. He also wishes to have the illusion of being the first to make a discovery, without having to follow in the footsteps of a predecessor or to cite the pages of an existing publication.

Without doubt such expectations are met by CD-ROM. We have worked along these lines for a year, imagining what such a CD-ROM could be. On the occasion of the Bicentenary, the Georges Pompidou Centre asked the National Institute of the French Language to provide a data-bank on the 1789 Revolution, which could be consulted by the public. A prototype was therefore developed, which made use of some of the essential functions of FRANTEXT. It was deliberately oriented towards statistical application as well as documentary research. Figure 14 illustrates the presentation of the main menu, showing two branches: on the right the user finds documentary research, and on the left he finds statistical processing.

Before choosing one of these two directions (which are not mutually exclusive), the user is asked to define his choices. He himself draws up the subset of texts that he wishes to consult (through a CHOICE OF CORPUS program, which has many criteria of selection—genre, author, date, title and

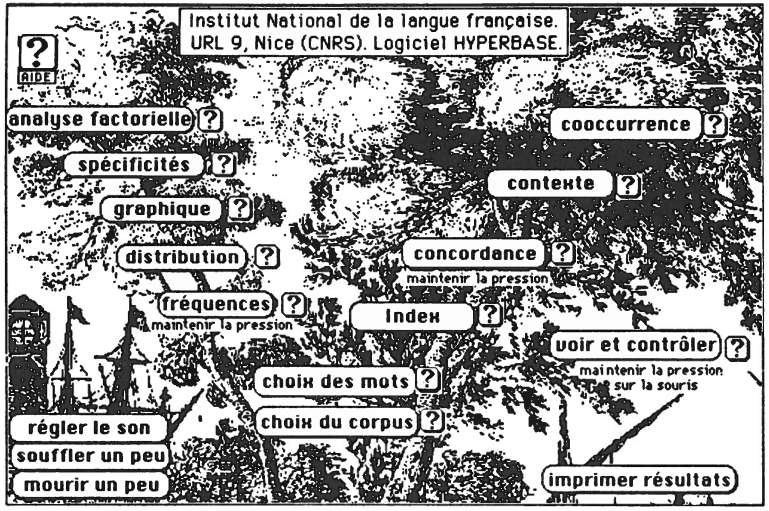

Fig. 14. HYPERBASE: Main Menu.

content—to which it applies Boolean operators; see Figure 15). The software can just as freely draw up a list of words that interest the user (through a CHOICE OF WORDS program, which offers automatic lists based on suffix, prefix, or lemmatization, and which authorizes additions, deletions or cross-sorts; see Figure 16).

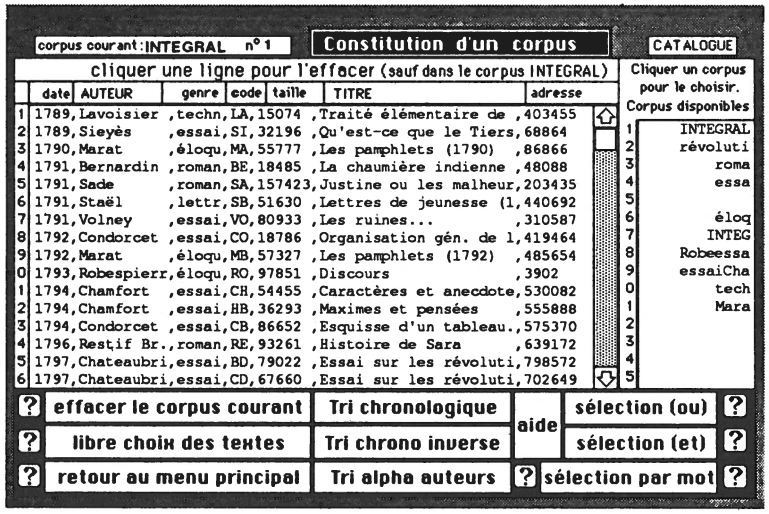

Fig. 15. Choice of Corpus.

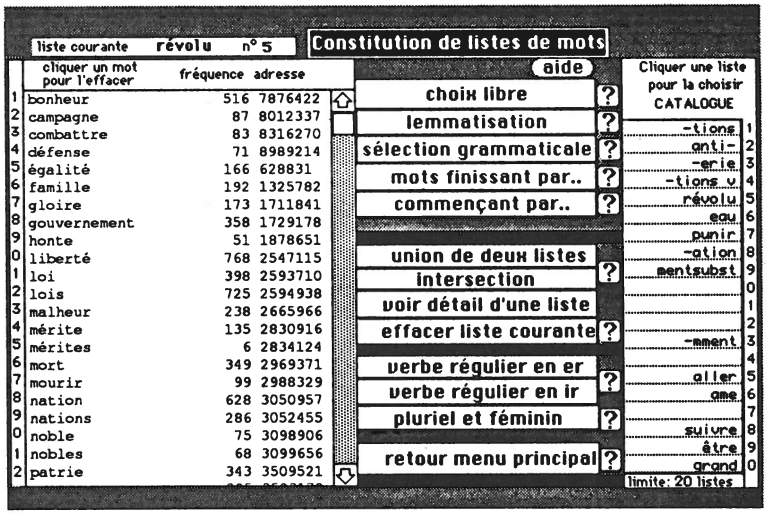

Fig. 16. Choice of Words.

The user might first of all wish to perform control operations, in order to examine the texts page by page, (SEE A TEXT program, see Figure 17), or consult the file of words, in alphabetical order or in order of decreasing frequency (SEE WORDS program, see Figure 18), or he may wish to pass instantly, by a simple 'click' function, from texts to words, and from words to texts, according to hypertext methods. Should he wish to check the contents

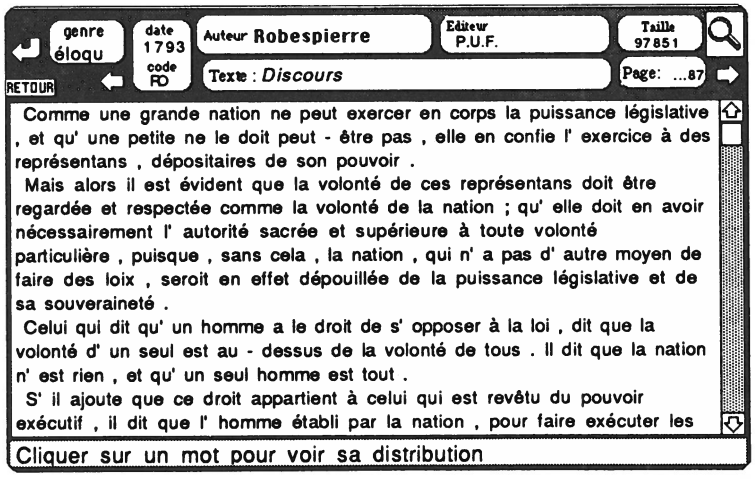

Fig. 17. A Page-Card.

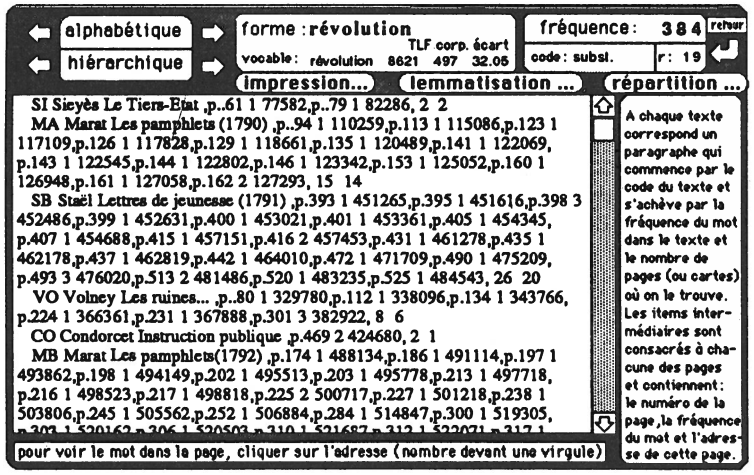

Fig. 18. A Word-Card.

of the current corpus and of the current list (which are variable and can be modified at will), a rapid-overview facility can flip through all the bibliographical cards of the texts in the corpus, or the identity slips of all the words in the current list.

If the researcher wishes to engage in documentary research, he chooses from the buttons on the right, in order to summon up an index (INDEX program), a concordance (CONCORDANCE program, see Figure 19), a list

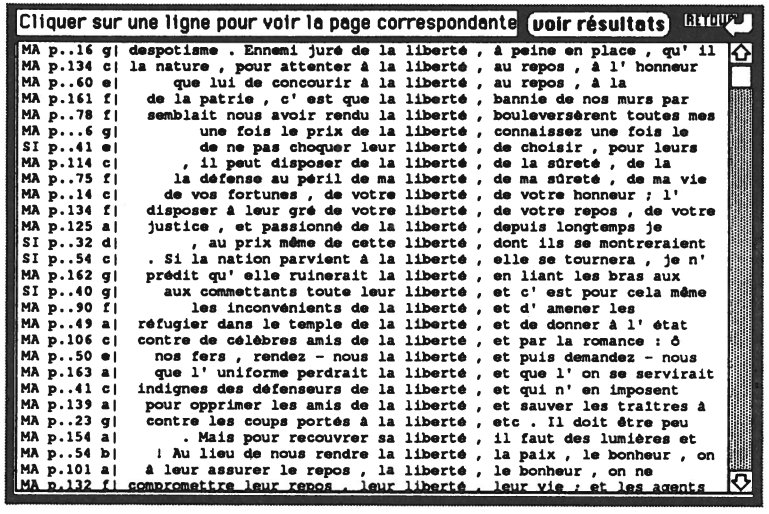

Fig. 19. Concordance.

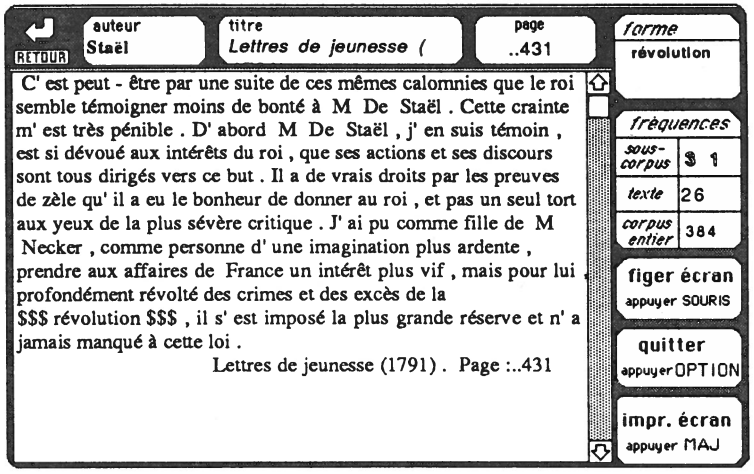

Fig. 20. Context.

of sentence-contexts (CONTEXT program, see Figure 20), or a search of co-occurrences (CO-OCCURRENCE program). In each of these operations he is free to provide either a word-token or a word-type (whose different forms will be given automatically), or else a word string, a phrase, or a word list.

The HYPERBASE software also facilitates quantitative methods. It can yield contingency tables (FREQUENCY program, see Figure 21), and it

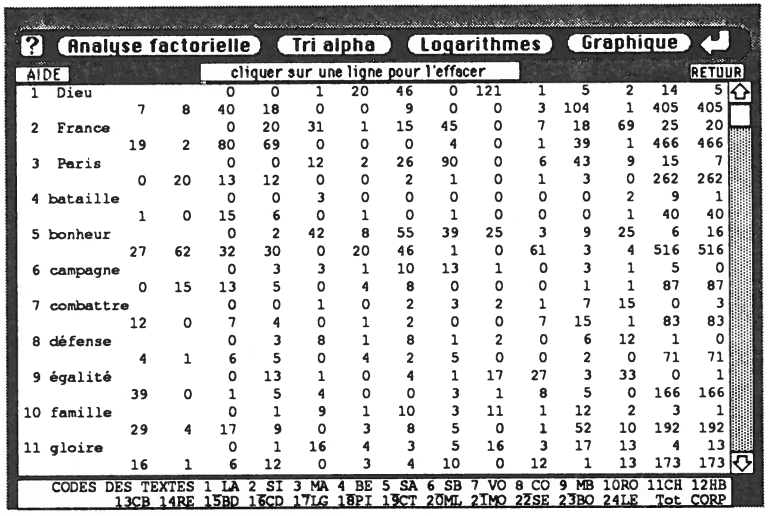

Fig. 21. Frequency.

provides graphs comparing the distribution of two words in different texts, or the profiles of two texts by means of a series of words (GRAPHICS program, see Figure 22). An interface is available for the exploitation of these tables by the methods of multi-dimensional scaling (FACTOR ANALYSIS program). Finally, several modules are offered which display the distribution of frequency-classes, in each text as well as the entire corpus (DISTRIBUTION and ZIPF'S LAW buttons), or which detail the specific vocabulary of a text or the specific properties of a word (SPECIFICITIES button, see Figure 23).

Finally, the results appear in two forms which are not always identical: on the screen, at the moment they are obtained, and in a file where they are stored before being checked and printed. The results are then subject to sorting operations, notably the concordances, which come in various layouts: chronological order, reverse chronological order, alphabetical order of authors, sorts on the left-hand or right-hand contexts. An editor with a printing module is provided, which makes finishing touches and final layout possible. It should be noted that the results are not an end in themselves, and that hypertext methods are applicable to them. In particular, it is enough to point at a line of the created concordance to cause the corresponding page to appear on the screen.

The search algorithm takes into account the limitations inherent in the use of a CD-ROM, particularly the relative slowness of disk access. For each exploration, only two movements of the reading-head are required. The accesses themselves have been optimized, on the basis of the supposed frequency of queries.

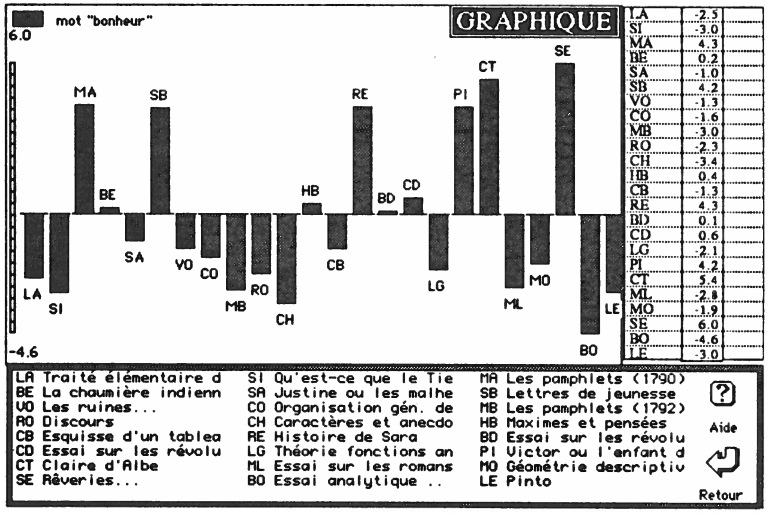

Fig. 22. Graphics.

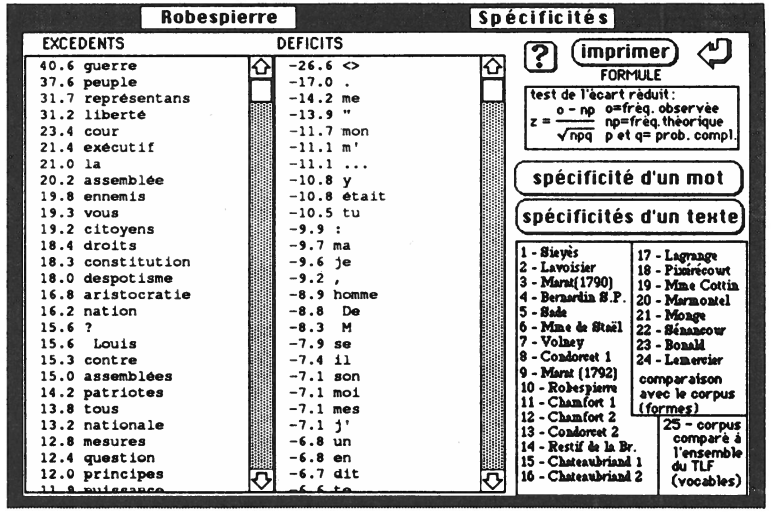

Fig. 23. Specific Vocabulary.

The necessity for speed has not resulted in the kinds of sacrifices often made by documentary products, which normally disregard function-words. These words are nevertheless valuable in linguistic applications that aim at being exhaustive. The reverse file contains all the words, and also incorporates narrative markers, particularly punctuation marks.

The entire program has been written with the intention of remaining user-friendly. Even if it is aimed at a specialist public, the presentation is not dull. On the contrary, we have tried to exploit all the graphic resources of the screen (for example, a portrait of the writer or a facsimile of the original edition are displayed). Full use is made of the operating facility of scrolling menus or control buttons, as well as the spontaneity of use provided by the 'mouse'. Even sound has been introduced. And for each function a help-menu has been provided at the moment when it is selected: all you need to do is to use the question-mark which accompanies each command.

This concern for user-friendliness and versatility explains the choice of an object-language for the whole of this software, written in HYPERTALK, and supplemented by external controls and functions. Because we are dealing with an interpreted language, the software is open to the addition of all sorts of complements.

In its present state, the prototype makes use of a database of twenty-four complete texts from the revolutionary period, from 1789 to 1800. Robespierre rubs shoulders with Sade, Siéyès with Marat, Madame de Staël with Chateaubriand and Chamfort with Condorcet. The corpus may favour essayists (such as Volney, Marmontel, Bonald), but novelists are not forgotten (for example, Restif de la Bretonne, Madame Cottin, or Bernadin de Saint Pierre),

and nor are dramatists (Pixérécourt, Lemercier), or scientists (Lavoisier, Monge, Lagrange). In total, 1,300,00 words are available.

The exhibited version—which comes in colour on a MAC II—works on a hard disk (of 40 Mb). The database was also installed on a WORM compact disk, in order to experiment with laser technology. But the CD-ROM version is not expected to come out immediately because HYPERTALK has not yet been stabilized. Great improvements are expected in the coming months, in terms of speed and security. (There have been four versions of this language in one year. The example of SUPERCARD shows that the development of multiple windows, large screens and colour is a possibility, and we can look forward in the future to the advantages of compilation.) Since the HYPER-BASE software is under a development contract from APPLE, it is hoped that we may be among the first to profit from the future improvements in the language.

Of course, other data has been considered, like Middle Ages texts, or nineteenth century poetry. There have also been plans for versions without data, but equipped with programs for preparing the data. Finally, for texts still under copyright, a version of the software has been devised that avoids the complete text in favour of contextual passages. To sum up, the planned versions are oriented in turn towards the two aspects hitherto considered by the creators of databases: the first is the complete text, which is denser, and perhaps more challenging because discoveries are more possible there; the second, the structured database, is more channelled, more signposted. The way has been cleared in advance, and one moves forward more quickly.

Lexical statistics, where copyright does not apply, moves from one aspect of documentary processing to the other. The qualities of these studies are modest, discreet, and always available. The statistics are inherent in the data, but they are hidden in the text and they only appear after the express order has been given, when the user requires figures and calculations. Providing as they do this modest service, it is to be hoped that one day lexical statistical studies will be given their due.

References

Brunet, E., *Le Vocabulaire de Giraudoux. Structure et évolution* (Geneva, 1978).
——— *Le Vocabulaire français de* 1789 *à nos jours* (Préface de Paul Imbs), 3 vols. (Geneva, 1982).
——— *Le Vocabulaire de Marcel Proust* (Préface de J. Y. Tadié) (Geneva, 1983).
——— *Le Vocabulaire de Zola* (Préface de Henri Mitterand) (Geneva, 1985).
——— *Le Vocubulaire de Victor Hugo* (Préface de Charles Muller (Geneva, 1988).
Muller, C., 'Méthodes quantitatives et informatiques dans l'étude des textes', *Computers in Literary and Linguistic Research*, Vol. 1 (Geneva and Paris, 1986), 11.

Part Two: Text Analysis Tools