The Value of Computer-generated Concordances in linguistic and literary Research

Although the reflections set forth in this essay are, specifically, in reaction to the views of Paul Tombeur as he expressed them in an article on the "Application des méthodes mécanographiques à un auteur médiéval,"* regular readers of this periodical will recognize that his views closely parallel those of the research team of the Laboratoire d'Analyse statistique des Langues anciennes. All the greater is my gratitude, then, toward M. Louis Delatte for allowing me to present in these pages my thoughts on the desirability of concordances, as distinct from indexes, in computer-aided linguistic and literary research.

While the bulk of M. Tombeur's study consisted of the presentation of his findings on the language of the medieval Belgian author Raoul de.Saint- Trond, its conclusion is a veritable indictment of concordances of the key-word- in-context type, and at the same time a vigorous defense of the method of morphological and syntactic analysis employed by the LASLA. Since the first fascicle of this Revue, M. Delatte has insisted, "Nous ne voyons pas l'intérêt d'établir des concordances,"** and M. Tombeur's condemnation is even more vehement: "Nous écartons . . . ces concordances, parce qu'elles sont fallacieuses, inutilisables dans certains cas, inutiles dans beaucoup d'autres."*** Although the methodology of the LASLA is usually irreproachable, it seems to me that the facts which motivated such a severe judgment of the value of concordances might with profit be reviewed.

Key-word-in-context concordances provide a listing of all

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

the words in a text, together with their context, and an indication of where they are to be found in the original work. The amount of context included may vary: some machines give up to 144 columns of print- out, so that, reserving eight of these for the mention of location and up to fifteen for the word being concorded, one will often be able to obtain a context of around 120 characters. The key-word can be positioned anywhere on the line, with the result that the printed context may be limited to what follows the key-word (with the latter, in this case, situated on the extreme left of the page), or to what precedes it (key- word on the right), or, and this is usually the most convenient arrangement, a combination of the two. In addition to these features, some programs provide for the sub-sorting of either the preceding or the following context. Thus multiple occurrences of a word-form may be listed according to the alphabetical order of the following context, and one will have a concordance of phrases or sentences, and not merely of words; if this is undesirable, the occurrences may simply be arranged in the order of their appearance in the base text.

Since concordances of this type have been of great use to many researches and are widely employed by linguistic and literary scholars in the United States**** as well as in other countries of the world, I was a bit surprised at M. Tombeur's unequivocal condemnation of the key-word-in- context technique. The LASLA type of index which he advocates is based upon an input which has first been analyzed morphologically and syntactically by a scholar, and the words of the text, accompanied by a coded form of this analysis, are ordered under their lemmas rather than under the form in which they appear in the original work. It has also been the practice of the LASLA to append to the word-list a frequency count of the vocabulary and a series of statistical tables relating to sentence length, morphology, grammatical categories and word order.

M. Tombeur's objections to key-word-in-context concordances may be divided into two types: those with which he endeavors to prove that they are lacking in usefulness, and those which deal with technical problems related to their publication.

The key-word-in-context concordance is, of course, highly dependent on alphabetization, which is not always the most efficient manner of arranging a word-listing. Such grammatical entities as verbs with multiple roots are distributed among several different alphabetical entries, so that the researcher must look in more than one place before he can gather together all the forms of the verb. Thus, to cite M. Tombeur's own example, some forms of the verb tollo would be found under S. But such an objection assumes that the researcher will wish to unite the alphabetically divergent forms of the verb: another approach is not only conceivable, but often desirable. Suppose, for example, that one were to undertake a study of verbal prefixes, using a computer- generated index of the LASLA type. In the case of the prefix sub-,one could search in the word-list under SUB to find subaro, subausculto, subbibo, and so forth, and then go farther on under the S heading to such entries as succedo, suffero, suggero and even sumo to find phonological variations of sub- which are the result of the combination of this prefix with the initial sounds of various root verbs. Surprisingly enough, however, it would also be necessary to search under the heading T, for there one would find sustuli and sublatum, listed as forms of the verb tollo! One man's advantage is another's inconvenience: in this particular case one could more easily, and with less likelihood of an error due to incompleteness, work from an alphabetical listing of unanalyzed forms of the kind found in key-word-in-context concordances. By no means am I saying that the LASLA type of analysis is inferior to concordances: on the contrary it is of the greatest practical value. But it is not adequate to the needs of all researchers, and to assume that it

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

is would be to underestimate the variety of needs that a word-list will be called upon to fill.

A related matter is the question of homographs. In the LASLA system the philologist who performs the original analysis takes great care to distinguish among the various meanings of words which are spelled the same. But would it not be much more simple to have all the examples of a given spelling before one's eyes, together with the context of each, so as better to be able to distinguish and classify all the meanings? This would be especially useful for languages which are being submitted to scientific analysis for the first time, or about which there is great controversy, such as the Linear B material, but it might also be advisable for works written in medieval Latin where new meanings, sometimes deriving from the various vernacular languages, are of frequent occurrence. M. Tombeur himself tells us that in the matter of semantic evolution, the Totius Latinitatis Lexicon of Forcellini, upon which, in large part, the lemmatization and the distinction among homographs of the LASLA analysis is based, "contient parfois de fâcheuses inconséquences".***** A scholar whose interests are lexicographical might be more grateful for the key-word-in-context concordance of "raw" forms than he would for the work of a philologist who, basing himself upon the dictionary of Forcellini, separates the diverse meanings of homographs. If key-word-in-context concordances were provided for all the authors of the latin middle ages, lexicographers could work at their leisure on the many problems which would then be open to resolution.

M. Tombeur disapproves too of the small amount of context provided by concordances. How, he asks, can one determine the sense of a verb when the subject of the object, or at times both of these, are outside the printed context? Surely this is a serious problem for some researchers, but again it presents no difficulty at all for others, namely those who

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

are dealing with a language where word-order is a dominant feature of syntax, and inflections are minimal. I can state with assurance that 120 characters of context are more than adequate in the case of Old French, and this is no doubt true too for the vast majority of living literary languages, although admittedly not for all, nor for every author who wrote in them. Other factors must be taken into consideration here: where long Proustian or Ciceronian periods are involved, it is most desirable, for certain highly significant words, to include as context the whole sentence in which the word is found. There are concordance programs which will perform this operation, such as UNICON, DISCON, and TRICON****** as well as the one offered to researchers in the pages of this Revue.******* For other texts, in Latin as well as in the modern languages, such as Catullus, the Goliardic poems, the chansons de geste, or the poetry of Blake, it would be unnecessary and even inefficient in most cases to make the context co-equal with the unit of discourse. But a concordance will save time even in cases where it is not completely adequate. Every mention of a word such as fortuna will not be of equal relevance for the scholar working on the history of this concept, but he will, with a glance at even a relatively short context, be able to separate those occurrences which interest him from the unimportant ones, and then refer to the edition for their full context.

As for periphrastic constructions whose elements are widely separated in the text, these will present difficulties in any case, but they are much more of a problem in Latin than in most living languages (German excepted). Likewise for misleading contexts: the conjunction cum followed by an ablative will more than likely be misinterpreted as a preposition if its verb is outside the printed context, but this rather unusual problem is also peculiar to a highly inflected language. And, once again, one may use a concordance in which the context consists of the entire discourse block.

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

Two more objections of a more technical nature are raised.

The first is that if all the words in a text, with their context, are included in the print-out, the concordance will be extremely large. What is more, M. Tombeur tells us, the listing of function words will be completely useless. But one should not be overly hasty in making this judgment. Frederick Mosteller and David L. Wallace have established the identity of the authors of the American Federalist Papers mainly through an examination of the use of key function words such as from, to, by, upon and because.******** The "insignificant" words in a literary text are sometimes revealed to be highly significant when viewed in the proper light. I am not advocating the publication of all the et's in Belgian Latin texts, which, as M. Tombeur tells us,********* would result in ten volumes of 400 pages each; but certainly for the more important works of each literature all the function words should be included. One solution to this problem, for less important texts at least, is to treat function words separately, using a concordance program which is equipped with an exclusion subroutine to eliminate all the function words from the print-out, and then, this time employing an inclusion subroutine, appending a frequency list, or even an index, of these words. A similar method would be especially appropriate for a word like cum, discussed above, which could be provided with a long following, and little or no preceding, context.

Would it be feasible to publish an output which is 144 columns wide? Yes, provided that the means employed is photo offset reproduction, through which the size of the final product can be reduced to prevent its proportions from becoming unwieldy. This process is used extensively for the publication of computer output, including the pages of this very article.

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

Not only have the limitations of key-word-in-context concordances been exaggerated; I would go so far as to say that some studies are impossible without them. An entire class of texts which could not be fully evaluated by the critic from a stylistic point of view, were it not for these tools, is the extensive body of oral poetry which includes some of the masterpieces of world literature: the Homeric poems, the Mahabharata, Ramayana and Rig-Veda of India, the Gilgamesh epic, the Byzantine Digenis Akritas, the chansons de geste, Beowulf, the Niebelungenlied and the contemporary Serbocroatian epic. All of these poems and many others are composed in oral formulas, more or less fixed word-patterns which recur throughout the poem and serve as an aid to the singer for his task of rapid composition before an audience. The identification and subsequent classification of formulas cannot be accomplished with word-lists or indices of the type which M. Tombeur advocates. These ends can be met, however, through the use of key-word-in-context concordances with right-hand-sort subroutines, and any attempt to attain them by manual means is doomed to incompleteness, inexactitude and stultifying tedium. The same can be said of any attempt to isolate patterns of words: the study of repetition is a field to which the concordance is particularly well adapted.

But it is often very practical for other uses. Let us suppose that a lexicographer wishes to analyze the vocabulary of a literary work. Is he to wait until a team of philologists, supplied with the most advanced equipment available, gets around to treating his author, who may very well be a minor one or unfashionable in academic circles at that moment; or is he to avail himself of the generous hospitality of M. Delatte, whose resources are, after all, not unlimited, and launch himself upon an exhaustive study, under the guidance of the LASLA, of the morphology and syntax of his text? It may well be that neither of these solutions is practicable. As "inutiles" as key-word-in-context concordances may appear

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

to M. Tombeur, our hypothetical philologist will undoubtedly find that only a concordance of this kind can quickly and easily furnish him with an error- free listing of all the vocabulary of his text. Given the advantages of time and accuracy that accrue to him if he follows this procedure, he will be more than willing to search out the various forms of irregular verbs, if he finds this necessary, and to take special care in distinguishing among the various meanings of homographs, particularly since, if he had done his work in the traditional manner, he would have had to carry out these two operations anyway.

It might be objected that these considerations do not justify the trouble and expense of publishing key-word-in-context concordances. But how much time and effort must be expended for the preparation of complete analyses such as those undertaken by the LASLA? In the last two years alone, giving only a fraction of my time to the task, I have been able, with limited funds, to prepare fourteen computer-generated concordances of Old French narrative poems, together totaling more that 45.000 verses. The morphological and syntactic classification of a body of material this large would take a whole team of philologists, working at the rate of 250 words a day, twice as long to produce, given the fact that Old French spelling is not normalized and that works written in this language are therefore not amenable to automatic analysis. Should the immense body of information made available by this means be allowed to sit on a shelf waiting to be exploited? Would one be justified in keeping it for oneself, when a single scholar's interests cannot begin to exhaust even a tenth of this data? A complete philological analysis of the most important texts is highly desirable, but who is to assure us that it will be completed even in our lifetime? The key-word-in-context concordances will satisfy the needs of many researchers, and they do not, any more that any other philological enterprise, preclude more detailed analyses. But they represent an enormous savings of

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

time and energy and are, by that fact alone, an immense advance over traditional means.

Between the literary text as it is found in a scholarly edition and the evaluation of this text by exact methods, there lie several possible intermediate states. The card-index, assembled patiently and laboriously, often inflicting most inhuman effects upon the humanist, is a thing of the past. Its successor, I believe, is the computer-generated concordance, which takes the raw material of the text, and, without altering it in any way, places it in an order which is easily exploited and suitable to the most diverse of needs. The morphological and syntactic analysis proposed by M. Tombeur and so ably perfected by the team of the LASLA represents yet another step, neither negating nor supplanting the concordance. It is at the same time more sophisticated and less versatile. But both methods have their place in the developing field of mechanolinguistics.

University of California, Berkeley Joseph J. DUGGAN.

NOTES

* Bulletin Du Cange, XXXIV (1964), 125-160.

** P. 5.

*** Op. cit., p. 159.

**** See the special supplement to the Newsletter of the American Council of Learned Societies, entitlet "Computerized Research in the Humanities, a Survey," June, 1966.

***** Op. cit., p. 130.

****** See Sydney M. Lamb and Laura Gould, Concordances from Computers (Berkeley: Mechanolinguistics Project of the University of California, 1964).

******* I (1965), 6.

******** "Inference in an Authorship Problem," Journal of the American Statistical Association, LVIII (1963), 275-309, and Inference and Disputed Authorship: The Federalist (Reading, Mass.: Addison-Welsey Publishing Company, 1964).

********* Op. cit., p. 156n.

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.