Association for Literary and Linguistic Computing

BULLETIN

1976
Volume 4
Number 1

Literatur und Datenverarbeitung. Ein Zwischenbericht zur Verarbeitung neuer deutscher Texte.

Im Unterschied zu der ersten Phase der literaturwissenschaftlichen Datenverarbeitung mit aufsehenerregenden Erfolgen Ende der 60-er Jahre scheint derzeit Ruhe, ja Zurückhaltung auf diesem Gebiet eingetreten zu sein. Man müßte jedoch eher von einer Konsolidierungsphase sprechen, in der, wenn auch mit beschränkteren Mitteln, die Möglichkeiten der Datentechnik für die Literaturforschung erprobt werden.

Im Rahmen dieses Berichts, der an den von mir edierten Sammelband *Literatur und Datenverarbeitung* (Tübingen: Niemeyer, 1972) anschließen kann, möchte ich auf je zwei abgeschlossene und laufende Projekte auf diesem Gebiet in Aachen hinweisen, die mir für die Entwicklung jedoch typisch erscheinen. Erschienen sind 1973 im Rahmen meiner Arbeit über *Drama im Bürgerlichen Realismus (1850-1890)* (Frankfurt a.M.: Klostermann, 1973) die Indices zu den Spielplänen des Königlichen Hof-Theaters zu Berlin (1849-1899) und des Stadttheaters Frankfurt am Main 1849-1886, die ein neuartiges Hilfsmittel der Rezeptionsforschung darstellen. Ebenfalls aufgenommen in diesen Band sind Auswertungen zur Häufigkeit bestimmter Gattungen, zur bevorzugten Aktanzahl und zur Verknüpfung von Gattung und Dramenstruktur.

Soeben erschienen ist auch die Arbeit von Hans Otto Horch mit dem Titel *Gottfried Benn. Worte-Texte-Sinn. Das Problem deskriptiver Textanalyse am Beispiel seiner Lyrik* (Darmstadt: Thesen, 1975), in der eine Reihe von grundlegenden

Problemen literaturwissenschaftlicher Datenverarbeitung anhand eines breiten, computerunterstützt aufgearbeiteten Textmaterials diskutiert wird. Es geht hier vor allem um die mit der Textzerlegung verbundenen methodischen Schwierigkeiten der Sinnentnahme (und Identifikation) einzelner, aus dem Kontext gelöster Wortkörper.

Bei den zwei noch in Bearbeitung befindlichen, jedoch kurz vor dem Abschluß stehenden Projekten handelt es sich um das Wörterbuch zu den Dramen Heinrich von Kleists und um die neue Hölderlin-Konkordanz. Das Wörterbuch zu den Dramen Kleists geht in folgenden Punkten über die bisher im Rahmen der Reihe 'Indices zur deutschen Literatur' erschienenen Bände zur Prosa Kleists hinaus:

1. Das Wörterbuch ist klassifizierend und lemmatisiert aufgebaut.
2. Es enthält Stellenangaben auch zu den Varianten, aber nur zu den Wortkörpern, die im endgültigen Text nicht auftauchen.
3. Es gibt Hinweise nicht nur auf die Stelle, sondern auch auf den jeweiligen 'Benutzer' des Wortes, d. h., ob es und von welcher Person es gebraucht wird, bzw. ob es der Regieanweisung dient. Damit enthält das Wörterbuch auch 'pragmatische' Informationen.

Das Wörterbuch, eine Analyse von Kleists Dramensprache mit Hilfe des Computers, steht vor dem Abschluß. Es soll 1976 im Druck erscheinen und die Reihe 'Indices zur deutschen Literatur' fortsetzen.

Beim Wörterbuch zu Hölderlins Werken ist der bislang für die deutsche Literatur wohl umfangreichste Bestand an Textdaten auf Datenträger gebracht worden. Erstellt ist damit eine 'literarische Datenbank', die erhebliche Probleme des Aufbaus und der Organisation mit sich bringen mußte. Neben der Wörterbuchherstellung kann sie selbstverständlich der Fülle weiterer entwickelter oder noch zu entwickelnder Schritte literarischer und linguistischer Analyse mit Hilfe des Computers dienen. Das Wörterbuch, zu dem die Druckvorlagen des ersten Bandes im Anfang des nächsten Jahres vorliegen werden, ist nicht nur klassifizierend und lemmatisiert aufgebaut, sondern wird auch durch Beigabe von Kontexten den Wert einer Konkordanz erhalten. Erprobt ist bereits ein interaktiv arbeitendes Programm zur Erzeugung von sinnvollen Kontexteinheiten, das Mängel und Aufwendigkeit bisheriger Kontextprogramme vermeidet.

Schließlich sei erwähnt, daß beide Projekte konzipiert sind in Hinblick auf moderne Vervielfältigungsverfahren, sei es auf microfiche oder auf Lichtsatz hin. Gerade hierbei zeigt sich der deutliche Vorteil der elektronischen Datenverarbeitung bei der Wörterbucherstellung gegenüber herkömmlichen Verfahren. Dem Ideal eines 'vollautomatischen Wörterbuchs' kommen beide Projekte in gewisser Weise näher. Es zeigt sich jedoch, daß um der Komplexität des Gegenstands willen und aus grundsätzlichen hermeneutischen Überlegungen die Rolle des bearbeitenden Philologen und Literaturwissenschaftlers nicht an Bedeutung verloren, sondern eher gewonnen hat.

Professor A. Zampolli (Italian Texts)

1. Texts in Machine-Readable Form. It is not possible to cite the large number of texts, of all periods, both in Italian and local dialects, which have been processed within Italy (by the users of the Linguistics Division of CNUCE, CLD) and abroad (mainly by the Italian Institute of Utrecht). It seems more important to remember that a two-way scheme is in progress in order to unify the various data banks into one single network.

   (a) Foreign institutes adopt, if possible, the standards of the Bank of Italian Linguistic Data (located at the CLD). This has been done, for example, in the case of the eighteenth century Italian newspapers

processed at Vancouver.

(b) In other cases the CLD produces the software and technology necessary for the transformation, into its own standards, of texts processed abroad (e.g. those from Utrecht).

A questionnaire, compiled in order to gather information concerning projects underway, is soon to be sent out in Europe and America. The results will be sent to all those who answer. We are counting on the collaboration of the national representatives of the ALLC and of other associations for assistance in supplying us with address lists for this purpose.

2. Linguistic Analysis. If the standardization of texts in machine-readable form is almost accomplished, the task of normalizing the linguistic analyses within the texts seems to present more difficulties. For the moment, we aim to obtain the comparability of the results (consider, for example, the impossibility of a statistical study of texts lemmatized by different systems). The analysis of the differences between various types of lemmatization has led to the establishment of a general system for the definition of lexical units which can be used as minimum common denominators. The Italian machine dictionary (MD) of the CLD reflects this general system and it should allow the automatic comparison of texts lemmatized with different criteria. In order to increase both the detail and also the automation of the linguistic and statistical analyses of the texts, the MD is being continually enriched with new information. A hundred and fifty thousand lemmas have been completed with a semi-formalized definition of their meanings. We use these definitions to construct a formalized semantic network and also to study the chains of synonyms. (The Institute of Utrecht is also working in this field.) A model of lexical derivations has been begun. The classification of Italian verbs (more than one thousand six hundred), on the basis of their possible syntactical constructions, is now at a good point. All this information will be used in an algorithm which, we hope, should help the disambiguation of the grammatical homography.

3. Linguistic Statistics. The main concern is that of making the best use of the numerous texts now available in machine-readable form. We aim also to fill in the present gaps in the corpus for contemporary Italian. The work now in progress aims at:

(a) Making comparable, by means of operations of re-lemmatization, the statistical data produced by projects which used different linguistic criteria (e.g. *Lessico di Frequenza della Lingua Italiana Contemporanea*, Bortolini, Tagliavini, Zampolli, and the *Frequency Dictionary of Italian Words* of Juilland and Traversa).

(b) Examining the frequency distribution of linguistic units at different levels (phonological, lexical, etc.) with the aim of correlating their stability and variations with the various production factors: literary genres, topics, etc.

(c) Formulating, with observed data, a model capable of giving to linguistic statistics the theoretical foundation that it has lacked since the falsification of the most recent models.

Among current research projects, I could mention: at the phonological level, the work on a group of conversations recorded in Tuscany; at the lexical and morpho-syntactical level, a study on the language used in the eighteenth century (Vancouver), and contemporary newspapers (Pisa).

4. Applications. The Italian MD is used for the pre-analysis of texts in certain automatic documentation projects using full text. Mention can be made of projects in progress for the Italian Parliament and the Corte di Cassazione.

Dr W. Ott (Textual Editing Techniques)

1. From within the ALLC there does not seem to be much interest in the specialist group. There has been, for example, until last week only one reaction to the bibliography published in Volume 2, Number 1 (1974), containing a hint (by the author) to an article on methods of text input. Other information I had to extract, for example, from bibliographies in the *ALLC Bulletin*, *CHum*, etc. Also communications of other kinds had not occurred (except some hints by the Secretary of ALLC). I therefore doubt if there exists at all a specialist group within ALLC which goes beyond our activities at Tübingen. Maybe that this non-existence is the fault of the chairman (who, in this case, would propose to ALLC to look for a better organizer for this specialist group).

   On the other hand, this non-reaction and lack of enquiries could be a sign of the fact that there is no longer any need for a specialist group, since techniques for photocomposition of computer output are in the meanwhile easily accessible for any scholar who wants to publish his output; and a sign that there are not enough scholars who think of a computer solution for the other editorial problems.

2. This view seems to be supported by the fact that the only contacts with scholars from abroad (which went through other ways than the ALLC specialist group) concerned the application of computerized photocomposition on DIGISET with a Hebrew character font, which does not seem to be available at other centres. From this, I suppose that for literary and linguistic applications, photocomposition has become a standard (off-line) output facility.

   It seems, however, that one is not yet accustomed to the fact that photocomposition from machine-readable 'manuscripts' allows not only for the elimination of typesetting errors, but also (owing to the fact that photocomposition enables automatic page make-up, which becomes a standard feature more and more also in commercially available software) for a real integration of book production, including automatic index and register generation and the automatic setting of more than one apparatus at page-bottom in critical editions.

3. Since I last reported on our Tübingen programs in 1974 at Cardiff, these programs have been improved in this direction. The automatic index generation has become more flexible, as far as concerns the typographical representation of the single entries in an index. The automatic page make-up has been improved and allows the handling of up to ten different kinds of apparatuses on a single page to be co-ordinated with the text.

   These features make the programs more attractive also for the application with critical editions, and there are some edition projects running at Tübingen on the basis of these programs.

4. My own experience shows that computerized photocomposition is for larger edition projects the gate of access to computer applications also for the manuscript preparation (including the use of word indexes for the reconstruction of the text, and the use of programs for the automatic collation of different

versions of a text, for the automatic construction of a first, raw, but complete apparatus, and so on).

5. The easy manipulation of such projects presupposes that the single programs are not 'standalone' packages. The typesetting program, for example, must be considered only as one output module of a whole system of compatible programs - compatible in the sense that they are working on an identical format, so that the output of one program can serve as the input of any other, including the previously run program. Apart from the greater perfection of the single modules, it is, in my view, this integration of different kinds of text handling programs into text handling systems which is the real progress of the last years.