13. Survey of Natural Language Corpora in Computational Linguistics

1. Introduction

Two decades ago neither natural language corpora nor programs to operate on data were available. Although ample software is now accessible it has only been recently that many texts that humanists have been converting to machine-readable form over the past twenty odd years have begun to be collected, organized, stored in data banks at research centers or in archives, and made available to the academic community. Many archives or research groups also have optical character readers and are adding to their collection by this means. It is even possible now for scholars to request that a particular text be encoded by the optical reader. Nevertheless, the most efficient and economical way of obtaining a tape is by obtaining a copy of one that has been already encoded. One warning however: it must be realized that the supplying of copies of text by research groups and even by archives, is a secondary operation to research activities, consequently correspondence directed to the archive concerning desired texts may go unanswered for a long period of time, or not be answered at all. There are nevertheless a considerable number of archives that appear to take their responsibilities seriously.

There are so many data banks of natural language corpora that some criteria had to be established for inclusion herein. The size of the corpus was one of the criteria used, although it could not be applied uniformly for every genre. A novel usually consists of about 100 000 words, (30 000 to 200 000 words) and dictionaries, newspaper articles, poetry, and recorded speech a fraction of that size. A corpus of novels containing 250 000 words would be excluded while the same size corpus of poetry or newspaper articles would be included. Regrettably, some archives are not included, such as CETEDOC, for Latin, the *Istituto di Linguistica Computazionale* with its 80 000 000 word database, for Italian, and the Department of Welsh, (listed in the Oxford Text Archive bulletin as an archive) for Welsh, all of which did not respond to requests for information sent by letter, registered letter, and, for the latter two archives, a telephone call. If it is not possible to access the data, there is no sense in mentioning it.

2. General Language Archives

2.1. The Oxford Text Archive (OTA)

The OTA of the Oxford University Computing Service, established in 1976, is the first archive created to store machine-readable natural language texts and this without any restriction as to language, or time period. Originally the texts were only literary ones, but now include material that belongs to the field of linguistics, so that one may find in the archive corpora of dictionaries and of the spoken language.

The OTA publishes a booklet that contains a list of all the texts that are available through their services. It also contains information about natural language corpora that are contained in data banks belonging to other research groups. In the booklet there are 36 languages listed of which 27 are available from the Oxford Text Archive:

Arabic
Chinese
English
French
Fulani
German
Greek
Hebrew
Icelandic
Italian
Kurdish
Latin
Latvian
Malayan
Norwegian
Pali
Portuguese
Provençal
Russian
Sanskrit
Serbo-Croat
Spanish
Swedish
Turkish
Welsh

The archive has an impressive collection of Latin and English texts, with 91 Latin titles and 283 English titles listed as of September 1984. The great majority of the texts are available without restriction, some texts are available only with the explicit permission of the depositer, and a very small percentage of the tapes are only available to registered users at Oxford. — The OTA also has available the Oxford Concordance Program, software capable of creating word lists, word indexes, and concordances, as well as identifying patterns of prefixes, suffixes, co-occurrences, etc. It operates on any language, and is a truly versatile program. One of the few things it can not do is lemmatize vocabulary, but then one can not have a program that works with all languages and lemmatizes them at the same time. For further information, one should write to: Oxford Text Archive, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN.

2.2. Cambridge, Literary and Linguistic Computing Centre

Cambridge University has an archive similar to the Oxford Text Archive. It has texts in 19 languages that are listed in the booklet put out by Oxford. The languages are:

Catalan
Danish
Dutch
English
French
German
Hebrew
Italian
Khotanese
Norwegian
Old Norse
Polish
Prakrit
Provençal
Russian
Spanish
Turkish
Tibetan
Mediaeval Latin

Cambridge appears to be particularly strong in the Germanic languages. For further information write to: Literary and Linguistic Computing Centre, Sidgwick Avenue, Cambridge CB3 9DA, England.

2.3. Humanities Research Center, BYU

The Humatities Research Center (HRC) at Brigham Young University was established in 1981 to provide research and technical support to the College of Humanities. It also serves as a distribution agent for information on ICAME (see below). — The HRC has succeeded in obtaining for its archive copies of some of the most important language corpora, e. g., the Brown Corpus, the Limas Corpus, the *Thesaurus Linguae Graecae,* and has text of languages that are not easy to come by, e. g., Portuguese, Finnish, as well as a number of dictionaries for such languages as Serbo-Croatian, and of several Central and South American Indian Languages. For further information, write to: BYU Humanities Research Center, 3060 JKHB Brigham Young University, Provo, Utah 84602.

2.4. Novosibirsk University

The natural language corpora at the Novosibirsk University Computational Linguistics Center is made up of both written and spoken data in approximately 50 different languages, some of which have only an oral tradition. Many of the languages are spoken within the USSR or in areas that border on the USSR: The Finno-Ugric languages, — The Turkic languages — The Paleo-Asiatic Languages, — The Mongolian languages — The Tungus-Manchurian languages.

The corpora consists of about 1 000 texts that deal with the everyday life of the Siberian aboriginals; their tales, traditions, religions, etc. Languages other than those listed above have been added to the corpus: Indo-European, Japanese, American Indian (Chontal, Haida), and languages of Australian aboriginals. For further information, write to: The Novosibirsk University Computational Linguistics Center, Novosibirsk University, Novosibirsk — 58, 630058 USSR.

2.5. American Indian

2.5.1. The Siouan Languages Archive

The Siouan Language is spoken from central Alberta and Saskatchewan as far south as Oklahoma and from Wisconsin to Montana. There are sixteen languages in the Siouan family of which, to cite a few, are: Crow, Dakota, Osage, and Winnebago.

Shortly after 1830, missionaries began to work among the Indians and to teach them to read and write their own language. "In the last half of the 19th century, scholars became interested in the languages and cultures for their own sake and scientific recordings and analysis began" (Rood 1980, 191). The archive includes elementary reading and writing lessons, personal letters, diaries and histories, political and legal documents, songs, dictionaries and grammars, tape-recorded oral history, etc. The goal of the Archive is to establish this material as the database for a computer-operated information retrieval system. The archive is aiming at a bilingually accessible archive. — For further information, write to the Department of Linguistics, U. of Colorado, Boulder, Colorado, USA.

2.5.2. Language Analysis Project

The Language Analysis Project of the Department of Linguistics at the University of Pennsylvania, Philadelphia, Penn. 10 104, is working on the development of algorithms that can be applied across languages for morphological and syntactic analysis, especially relating to little known languages. The project has a language corpus of some 4,5 million words. The American Indian languages are:

Chorti, Mayan, Popoluca, Sayula, Takelman, and Yucatec.

The project also has other languages than American Indian, e. g., Somali, Indo-European languages including Slavic. — For further information write to the above address.

2.5.3. Yuman Dictionary Project

Margaret Langdon of the Linguistics Department at the University of California, San Diego has received a National Science Foundation grant for the construction of a comparative dictionary of the Yuman languages, a family of American Indian languages spoken in the southwest United States. Speech recordings were made of the Yuman Indians, and then transcribed either phonemically, or, for those languages that had a system of writing, orthographically. — One of the aims of the project is to reconstruct the prototype language of Yuman. In so doing, the project also expects to be able to identify and describe the grammar of the languages. The project was begun in 1984 and will run for two years, after which time the corpora of the spoken language, the dictionaries, and grammars will be available to the academic community. For further information, write to Margaret Langdon at the above address.

3. Psycholinguistic Database: Child Language Data Exchange System (CHILDES)

CHILDES was organized in 1984 at Carnegie-Mellon University (CMU), with the aim of bringing about major improvements in data collection, storage, and data sharing of child language data-transcripts. CHILDES has two centers, one at CMU and the other at the Max-Planck Institut für Psycholinguistik in Nijmegen, Holland. The system has a corpus of recorded speech of children from many countries, speaking: English, Spanish, French, Hebrew, Danish, Hungarian, Italian, Tamil (one of the Dravidian languages spoken in South India), and bilingual Chinese-English. At present count CHILDES has 25 corpora in its data bank. — For further information, write to: Brian MacWhinney, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA 152 13.

4. Individual Languages

4.1. Dutch

The most important language corpora of Dutch are found at the Vrije Universiteit in Amsterdam. The corpora are used principally for linguistic analysis:

(a) *"The Elsevier Corpus"* contains a representative selection of texts of all types from the period 1920—1975 totaling ca. three million words. The corpus is used principally for linguistic analysis.

(b) *"The Eindhovens Corpus"* contains a representative selection of texts of all types from the year 1971 with a total of approximately one and a half million words.

(c) *"Phonological and Morphological Narrowly Transcribed Dialect Texts"* — transcribed in the international phonetic alphabet. The corpus is not finished but can be used in its present form.

(d) *"Middle Dutch Dialects on the Basis of 14th Century Charters"*. Not complete but can be used as it is.

Word indexes, concordances, and information concerning the morphology and syntax of the language are also available for the Eindhovens and Elsevier corpora. — For further information, write to either: the Department of Linguistics (Pieter Van Reenen) or the Computer Department (G. Van der Steen), Faculteit der Letteren, Vrije Universiteit, Amsterdam, Holland.

4.2. English

In addition to the following databases, one should refer to the OTA. Many scholars and research groups have passed their data to this archive, e. g., "The Dictionary of Old English" and its 600 some odd texts of old English/Anglo-Saxon.

4.2.1. ICAME

During the early 1960s, W. N. Francis, and H. Kucera of Brown University produced the well known Brown corpus, a collection of 500 texts in machine-readable form, each of about 2 000 words. The texts were made up of American prose and drawn from diverse genres that had been printed in 1961. The corpus was divided into two parts, one of "informative prose", articles from the press, government and industry reports, texts on natural science, medicine, mathematics, biographies, etc., the other of "Imaginative prose" i. e., fiction. The corpus was produced for purposes of linguistic analysis. — As with many research centers, the research team that produced the corpus has turned over the responsibility for distributing it to an archive, which in this case is the International Computer Archive of Modern English (ICAME). ICAME has the following language corpora:

(a) The *Brown Corpus, Formats* I and II. Format I is without grammatical tagging, has upper- and lower-case letters, and regular punctuation marks. It has the same line division as the original version, except that words at the end of a line are never divided. Format II is similar to format I, except that a new, longer line is used.

(b) The *LOB (Lancaster-Oslo-Bergen) Corpus Text,* like the Brown Corpus, consists of 500 texts of 2 000 words each of British English. A grammatically tagged version (each word being tagged according to its grammatical category) is being prepared and will shortly be available.

(c) The *London-Lund Corpus* contains samples of educated spoken English, in orthographic transcription with detailed prosodic marking. It consists of 87 "texts", each of some 5 000 running words. The text categories represented are spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc. — For further information write to: International Computer Archive of Modern English, Bergen University, Bergen, Norway.

4.2.2. Stanford Computer Archive of Language Materials (CALM)

CALM has prepared a version of the Brown corpus, called the Brown MARC format, in which each record of a text contains a sentence. Since, when analyzing language, the context is all important, it is of significant importance to be able to retrieve the sentence where the structure occurs that is being analyzed. Data that are formatted in this way facilitate greatly the retrieval of the sentences as the unit of analysis. For further information, write to: CALM, Department of Linguistics Stanford University Stanford, California 94305, USA.

4.2.3. Domesday Book Database

The Domesday Book is a comprehensive land survey, undertaken for tax purposes in 1086 at the request of William the Conqueror to learn about the worth and nature of land holdings in the kingdom he had just acquired. Its entries provide specific and fairly standardized data on the holder, tax obligations, agricultural exploitation, population, annual value and tenurial status of nearly every manor in 11th century England. Because of the massive size of the "Book", until it had been converted to machine-readable form, it had been extremely difficult to arrive at an understanding of its full context. In addition to being of significant value to English historians and geographers, it is of great interest to linguists. — For further information write to: C. Warren Hollister, Department of History, University of California, Santa Barbara, CA 93 106.

4.2.4. English Department, U. S. Military Academy

Col. Jack L. Capp of the English department of the U. S. Military Academy at West Point, New York is studying the works of William Faulkner, and for his research has created an author-corpus of many of Faulkner's novels: *Intruder in the Dust, As I Lay Dying, The Wild Palms, Go Down Moses, A Fable, The Sound and the Fury.* Copies on magnetic tape will be loaned to user upon written request. For further information write to Col. Capp, West Point, New York, USA.

4.2.5. English Department, Wisconsin University

Tod Bender of the English Department at the University of Wisconsin, Madison, WI 53 706, has been engaged in the stylistic analysis of literary texts. His principal interest has been in the works of Joseph Conrad, and for his research he has encoded the complete works of Conrad. Other works that he has converted to machine-readable form are: The complete poems of Gerard Manley Hopkins, and of Canon Richard Watson Dixon, the letters of Keats, *The Good Soldier* by Ford Maddox Ford, *Jane Eyre* by Charlotte Bronte and *Wurthering Heights* by Emily Bronte. — For further information, contact Bender.

4.3. French

4.3.1. Institut National de la Langue Française

The *Institut National de la Langue Française* (INaLF) is the umbrella organization for a number of research groups inside and outside of France that are conducting research on the French language. The *Trésor Général des langues et Parlers Français* is the group within INaLF that deals directly with the natural language corpora and its related services. The *Trésor Général* has a language-corpus of 17th, 18th, 19th, and 20th century French literary texts constituting 120 million words. The 19th and 20th centuries comprise over 1000 texts and 70 million words. The texts in machine-readable form from the last two centuries was created during the making of the *Dictionnaire de la Langue des XIXe et XXe Siècles.*

Contrary to the Oxford Text Archive, the *Trésor Général* only allows access to texts in machine-readable form at one of its locations or through the telephone networks, wherein services are available that will provide the scholars with word concordances, word indexes, without lemmatization or with partial lemmatization, and theme indexes. INaLF also provides information concerning norm word frequencies. The frequencies are broken down not only into time periods: century, half century, and every fifteen years, but also by genre: prose, verse, prose poetry, soliloquy, and dialogue. The frequencies are calculated on the lemmatized word. — For further information, including fees, one should contact: Services des Prestations Documentaires, INaLF, section de Nancy, 44 Avenue de la Libèration, C. D. 3310 Nancy, tel. (8) 3 96 21 76.

4.3.2. American and French Research on the Treasury of the French Language (ARTFL)

ARTFL, working in collaboration with INaLF, has obtained copies of most of the texts that are contained at Nancy. The ney also have the same restrictions that INaLF has. To have access to the corpora, one's institution must subscribe to the ARTFL data base. There is also a charge for CPU time. The ARTFL system is interactive and is intended principally for literary research. Users can work either on location at the University of Chicago, or from any location in North America, using a computer terminal and a moden. As at INaLF, programs are available that will retrieve information from the text. Software such as the Oxford Concordance Program, and ARRAS will retrieve information that is not only of a literary or lexical nature, but, on an elementary level, linguistic also, i. e., one can obtain information concerning prefixes, infixes, suffixes, co-occurrences, etc. — For further information, write to: ARTFL, The University of Chicago, Department of Romance Languages and Literature, 1050 East 59th Street, Chicago, Illinois 60 637, USA.

4.3.3. University of Manitoba

Paul Fortier of the Department of French and Spanish at the University of Manitoba has, over the past twenty years, created a language-corpus of 20th century French novels, as well as of a few texts from the 18th century (Montesquieu, Chenier, Marivaux). The texts from the 20th century are by: Robbe-Grillet, Beckett, Bernanos, Blais, Sartre, Malraux, Céline, Gide, and Camus. — Fortier's field of research is the computational stylistic analysis of the themes of twentieth century French novels. — For information concerning the terms for obtaining copies of these texts, write to Paul Fortier.

4.4. German

Any one who is interested in German machine-readable texts should consult *Dokumentation Textkopora des neueren Deutsch.* Institut für deutsche Sprache, Mannheim, 1982, which lists 75 collections of machine-readable texts.

4.4.1. Institut für Deutsche Sprache

The *Institut für Deutsche Sprache* (IdS) was established in 1964 with the specific responsibility of studying and describing the present-day German written and spoken language. The IDS has five principal language corpora, three of them of the written language, two of the spoken, covering a period from 1947 to 1978.

The Written Language

(a) *"The Mannheim Corpus 1"* contains 32 texts with about 2.2. million words. The texts, from the period 1950—1967, are drawn from: belles lettres, popular literature, memoirs, scientific literature, newspapers, and periodicals. —

(b) *"The Mannheim Corpus 2"* contains 47 texts with approximately 300 000 words. The texts, from 1950—1967, are drawn from edicts, statutes, legal judgements, textbooks, popular literature, scientific literature, newspapers, and periodicals.

(c) *"The Bonn Newspaper Corpus Part 1"* contains a selection of articles from the years 1949, 54, 59, 64, 74, from the newspapers "Die Welt" (W. Germany) and "Neues Deutschland" (E. Germany).

The Spoken Language

(a) *The "Freiburg Corpus"* contains 224 texts from 1968—1974 with approximately half a million words of discussions, interviews, recitations, narratives, and news reports.

(b) *The "Dialoguestrukturen Corpus"* includes texts of interviews and discussions from the period 1968—1978.

For further information write to: Institut für deutsche Sprache, Abteilung Wissenschaftliche Dienste, Postfach 5409, 6800 Mannheim 1.

4.4.2. Institut für Kommunikationsforschung und Phonetik at the University of Bonn

The *Institut für Kommunikationsforschung und Phonetik* has been engaged in research in the field of linguistic data processing and computational linguistics for the past 20 years. A number of data bases were created for this research:

Author Corpora

(a) The *"Kant-corpus"* contains the complete works of Kant, less his correspondence (19 volumes of the published version), and comprises some 2.8 million words. Indexes, frequency lists, ranking lists, etc. have been established for the corpus.

(b) The *"Heine-corpus"* contains the complete works of Heinrich Heine — some 2.5 million words. The corpus is not available for use except on location and by special request.

Language Corpora

(a) The *"Corpus of Middle High German texts"* consists of complete texts or works of different middle high German authors, e. g. "Heinrich Wittenwiler, Oswald von Wolkenstein, Hartmann von Aue, Konrad von Würzburg and Gottfried von Strassburg" (Lenders 1985, 13).

(b) The *"Corpus of Early Modern German texts"* contains 40 texts, each 30 pages in length from the period 1350 to 1700, from different places of origin: Ripuarisch, Hessisch, Mittelbairisch. Every noun, verb, and adjective in the corpus has been indicated by specific markers, so that lists of nouns, verbs, and adjectives have been generated.

(c) The *"LIMAS-corpus of Modern Standard German"* is considered to be a representative corpus of modern standard German from the year 1970. It has been constructed in accordance to the BROWN-Corpus and has 1 million words.

(d) The *"The Limas-Totalkorpus",* is a collection of texts: novels, specialized books, reports, etc. It contains 4 millions words.

Word Corpora (Dictionaries)

(a) The *"LIMAS-dictionary",* is a list of more that 130 000 words. Every word is combined with a set of linguistic markers, expressing one or more word classes plus morphological information.

(b) The *"Mackensen-dictionary",* is a machine-readable representation of the dictionary of *Standard Modern German* by Mackensen. It has 117 370 entries.

(c) The *"Cumulated Word Data Base"* consists of several machine readable dictionaries, e. g., the MOLEX-dictionary, and the SADAW-dictionary, which have been integrated and can be used on-line.

For further information, write to: Institut für Kommunikationsforschung und Phonetik, Universität Bonn, Poppelsdorfer Allee 47, 5300 Bonn, West Germany.

4.4.3. Arbeitsgruppe für Mathematisch-Empirische Systemforschung

The Arbeitsgruppe für mathematisch-empirische Systemforschung (MESY) has two language corpora:

(a) *"Student poetry"* — from some 150 anthologies of minor German students from the period 1822—1966, comprising about 300 000 words in approximately 3 000 poems.

(b) The *"Prose Corpus"* — 40 texts of fiction and non-fiction of German and non-German writers containing some 4.2 million words.

Although the information provided by MESY for the *Dokumentation Textkorpora des neueren Deutsch,* (IDS 1982, 50) states that the prose corpus is available with certain conditions, correspondence with a member of MESY claims that the texts are "not available to the scientific public for research purposes due to the publishers special contract conditions." It is reasonable to suppose that the prose corpus can be accessed on site.

4.4.4. University of Adelaide

The "Adelaide corpus" consists of seven sets of texts comprising, in all, 70 texts of written, present-day German. It contains 100 000 words. The corpus was created for research on a basic vocabulary of the Arts and Social Sciences, principally of a statistical nature, concerning the word's distribution, usage, range and frequency of use. — For further information, write to: Director of the Language Laboratory, The Universtiy of Adelaide, G. P. O. Box 498, Adelaide, South Australia 5001.

4.4.5. Arbeitsstelle für wissenschaftliche Didaktik of the Goethe-Institut, Projekt Phonethek

The Arbeitsstelle für wissenschaftliche Didaktik Projekt Phonethek, of the Goethe-Institut has created a language corpus "Wissenschaftsdeutsch-Corpus" and a word corpus. The *"Wissenschaftsdeutsch Corpus",* which was created for a project on scientific German, contains 102 texts of 34 subject areas, with an overall word count of 250 000. Each subject area is reprensented by 3 texts of the following type: textbook, professional journals, and the popular press. — The word corpus contains a list of 8 003 words "der Mindestfrequenz 101 auf ca. elf Millionen lfd. Wortformen von F. W. Kaeding's Häufigkeitswörterbuch der dt. Sprache, Steglitz bei Berlin (Selbstverlag des Hrsg.) 1898" (IDS 1982, 61). — For further information write to: Arbeitsstelle für wissenschaftliche Didaktik Projekt Phonethek, Goethe-Institut, Postfach 20 10 09, D-8000 München 2.

4.5. Greek

4.5.1. Thesaurus Linguae Graecae

The *Thesaurus Linguae Graecae* (TLG) is a computerized data bank designed to hold the entire corpus of ancient Greek literature, from Homer to a. d. 600 (and in some cases, beyond). Begun in 1972, the TLG now contains approximately 40 million words of ancient Greek text. — Specified segments of the TLG can be provided to users in a variety of tape formats. For further information write to: TLG project, University of California, Irvine, CA 92 717.

4.5.2. Data Bank and Research Tools for Septuagint Studies

The goal of this project is to establish a data bank for the study of Jewish scriptures in their Greek and related forms. Included in the database is the Greek text of the Septuagint with all textual variants, the Hebrew text with its textual variants, morphological analyses of the Greek and Hebrew material, and a file representing the alignment in parallel vertical columns of the Greek materials with the Hebrew. — As materials become ready for distribution, they can be obtained by qualified parties, at cost, on magnetic tape by writing to the IBM facility at the University of Pennsylvania, School of Arts and Sciences.

4.5.3. Duke Data Bank of Documentary Papyri (DDBDP)

The aim of this project is to create a machine readable base of all Greek words found in the documentary papyri, estimated at 6 million, found in approximately 35 000 texts ranging from the fourth century B. C. to the seventh century A. D. The work ist being done in collaboration with, and as a complement to, the database of the Thesaurus Linguae Graecae. — Data available from Phase I (which covers some 84 published volumes of papyri) will be available by the end of 1985. The material will be provided at cost. — For further information write to: John F. Oates and William H. Willis, Department of Classical Studies, Duke University, Durham, NC 27 706.

4.5.4. Data Bank for Ancient Greek Inscription

"The inscribed texts of the ancient Athenian public decrees from 403 to 318 B. C. have been encoded in Beta Format (the standard transliteration of the Greek alphabet used by the Thesaurus Linguae Graecae), with coding for epigraphical symbols, ... A data set of 18 837 lines, developed from 521 inscriptions of this period, has been included in the base ... Combined with the existing file of texts of the decrees and laws to 403, the new material builds a data bank of 784 inscriptions from the sixth century to 318" (West, Unpublished paper). — For further information write to: Prof. William C. West, III, Department of Classics, The University of North Carolina at Chapel Hill, 212 Murphey Hall 030 A, Chapel Hill, N. C. 27 514, USA.

4.6. Latin

4.6.1. Index Thomisticus

Literary and linguistic computing can trace its origins back to Father R. Busa and his computational studies of the works of Saint Thomas Aquinas, begun in 1949. To produce the word index and concordance to the works of Aquinas, Busa created a data bank of ten and a half million words of Aquinas' works, as well as three million more words in three alphabets (Greek, Hebrew, and Cyrillic) of texts of eight more languages, this in order to test procedures. The index and concordance are lemmatized. The concordance includes "about two and a half lines of context for each word" (Hockey 1980, 68).

In 1953 Busa founded the Center for the Automation of Literary Analysis at Gallarate. In 1966 the processing of the *Index Thomisticus* was transferred to the Centro Nazionale Universitario di Calculo Elettronico (C. N. U. C. E.), at Pisa, and control of the operation now appears to have been taken over by the *Istituto di Linguistica Computazionale.* To obtain information about the *Index Thomisticus* one might try contacting the director of the Center for the Automation of Analysis at Gallarate.

4.6.2. Lessico Intellettuale Europeo, Rom

The *Lessico Intellettuale Europeo* is working on a project to create a *Thesaurus Mediae et Recentioris Latinitatis.* So far, twelve works and text collections have been filed of Medieval Latin translations of philosophical and scientific texts. — For further information, contact Tullio Gregory, Lessico Intellettuale Europeo, Via Nomentana, 118, I-00161 Roma, Italy.

4.6.3. A. P. A. Repository of Greek and Latin Texts

Stephen V. F. Waite has been collecting and storing texts in machine-readable form of Greek and Latin literature for the past twenty years. Although he maintains Greek texts in his repository, he has deposited copies of the texts at the *Thesaurus Linguae Graecae,* so that consequently, the texts that are of principal interest to scholars are the Latin ones. Latin texts from 27 different authors can be found in the repository, and this number is constantly increasing, as is the case with most archives. One can find in the list of works provided by Stephen Waite, texts by Caesar Cato, Cicero, Euclid, Horace, Ovid, and Vergil. — For further information, contact Stephen V. F. Waite, Logoi Systems, 27 School Street, Hanover, New Hampshire 03755, USA.

4.6.4. Seminar für Klassische Philologie, University of Berlin

The Seminar für Klassische Philologie has a language corpus of works by Stiblinus, Sallust, F. Bacon, Thomas More, Minucius, Plinius, Vergil, J. V. Andrea, T. Campanella, Vegetius, Jordanus Rufus, Chiro, Ypokras Indicus. — For further information write to D. Najock, Seminar für Klassische Philologie, Ehrenbergstr. 35, 1000 Berlin 33, West Germany.

4.7. Norwegian

The Norwegian Computing Centre for the Humanities, Bergen University, has four language corpora, three of which are composed of newspaper articles, used mainly for vocabulary and frequency studies, and one of novels use pricipally for frequency and orthographical studies:

(a) "Newspaper articles from three central newspapers" — about 900 000 words from 1980—1981.

(b) "Samples of Newspaper Articles at 25 year intervals" — about 900 000 words from the years 1900, 1925, 1950.

(c) "Newspaper Articles from 1982—83" — about 500 000 words.

(d) "Sixty Novels" — Twenty each from the years 1937, 1957, and 1977 — with approximately one million running words. Half the novels are written in "Nynorsk", the other half in "Bokmå".

For further information, write to: The Norwegian Computing Center for the Humanities, Box 53, N-5014 Bergen-Universitet, NORWAY.

4.8. Spanish

4.8.1. Dictionary of the Old Spanish Language (DOSL), University of Wisconsin

DOSL has two principal corpora:

(a) an author-corpus comprising the Alfonsine-corpus,

(b) a language corpus comprising the Corpus of fourteenth-century Aragonese dialect works.

The origins of DOSL date back to 1930 when the Seminary of Medieval Spanish Studies at the University of Wisconsin began the compilation of an Old Spanish dictionary, using a core vocabulary drawn from the thirteenth century texts of Alfonso X, el Sabio, King of Castile and Leon (1252—1284). In 1970 it was decided to computerize the operation and the DOSL project was begun. It was also decided to include computer generated concordances of new transcriptions of some 250 representative manuscripts and incunabula dating from 900 to 1500. By 1978 the Alfonsine-corpus — all five million words — had been encoded. Following this, the fourteenth-century Aragonese dialect works completed under the aegis of Juan Fernandez de Heredia was converted to machine-readable form. In 1979 work was begun on a data base of actual dictionary entries, and lexicologists may now conduct a search of the corpora for old Spanish vocabulary and retrieve information consisting of a citation and linguistic details about each lexical entry. — Copies of texts in machine-readable form can be obtained for those texts that have already been formally disseminated by the Hispanic Seminary of Medieval Studies, Ltd. and will be distributed on an ad hoc basis. For further information write to: John Nitti, 1132 Van Hise Hall, Dept. of Spanish and Portuguese, University of Wisconsin-Madison, Madison, Wi 53706.

4.8.2 Départment de Langues et Linguistique, l'Université de Laval

In 1976 work was begun at Laval University to create an inverse dictionary of the Spanish language — *Diccionario inverso de la lengua española* (DILE). A word corpus of dictionaries was established as the data bank to draw upon to create the DILE and was given the title of "Sixteen Dictionaries and Lexicons of Spanish". The DILE project is also in the process of creating a language-corpus: "The Lexicon Database", of which at the present time, one text is available, a play: *Una Libra de Carne* by August Cuzzani.

The DILE project has chosen for its database, dictionaries of present-day Spanish, representative of various geographical areas and social strata. From this database, linguistic information of a statistical nature will be retrieved concerning (and here from an unpublished paper on the project shall be cited):

(a) *las finales de las palabras,*

(b) *las categorias gramaticales representadas,*

(c) *las categorias gramaticales y las finales de palabra,*

(d) *las finales de palabra y las fuentes y ellas fuentes y las,*

(e) *categorias gramaticales contenidas en cada una de ellas.*

DILE contains approximately 181 000 words as well as the word's grammatical category and information about its source. Approximately 99 000 words are drawn from the *Diccionario de la Real Academia española.* For further information, write to: Silvia Faitelson-Weiser, Department of Language and Literature, Faculty of Letters, University of Laval, Quebec, Qc. G1K 7P4.

4.8.3. International Electronic Archive of the Romancero (AIER), Madrid

The AIER is a cooperative project with its main headquarters at the Menendez Pidal Archives, Madrid, Spain. A number of institutions throughout the world contribute to the project whose purpose is to collect, store, and study the Pan-Hispanic ballad. In order to do this the AIER has established a language-corpus of ballads, as well as other corpora that contains related information.

(a) The "AIER Database" — Each record consists of 47 variable length, comma delimited fields of which the last is the ballad text itself. Other fields register the ballad theme's identification number, its title and prosodic data, other titles by which the ballad is known, etc.

(b) The "General Exemplified Index of the Romancero" (IGER) — Each record in IGER contains the *romance's* identifying number, its geographic dispersion, old and modern incipits, other titles by which the ballad is known, and one or more complete versions of the ballad.

(c) The "General Catalogue of the Pan-Hispanic Romancero" (CGR) — Each entry comprises 11 fields: the IGER number, title and prosodic data, detailed geographic distribution of versions, narrative traces left in other ballads, *contrafacta,* Spanish and English summaries with plot and *fabula* variants geographically identified, etc.

(d) The "Traditional Hispanic Romancero" — This corpus contains editions of individual regional collections (Sephardic, Portuguese and Canary Island ballad texts).

(e) The "Sources for the Study of the Romancero" — This corpus contains thematically organized editions of ballad texts.

For further information, write to: (In Europe): Menendez Pidal Archives, Instituto Universitario "Seminario Menendez Pidal", Universidad Complutense de Madrid, Menendez Pidal 5, 28036 Madrid, Spain. (In the United States): Diego Catalan, Dept. of Literature, D-007, University of California, San Diego, La Jolla, CA 92093.

4.8.4. Logotheque, Göteborg

Logotheque is involved in various collaborative projects, one of which is the creation of a reference corpus of two million running spanish words from newspaper material from which a concordance has also been prepared. This is in collaboration with Per Rosengren and Mr. David Mighetto of the Department of Romance Languages at the University of Göteborg. The information concerning access and dissemination of the data is the same as in the following section.

4.9. Swedish

4.9.1. Språkdata, Göteborg

Språkdata is the abbreviated form for *Institutionen för språkvetenskaplig databehandling* which is the Department of Computional Linguistics at the Göteborg University. Logotheque, the Swedish language bank, a service branch of the linguistics department, was established by the Swedish government in 1975. It is responsible for the collecting, storing, processing, and providing of linguistic material in machine-readable form. The language material is basically Swedish, although data on other languages are also included. Data at the Institute are encoded on the spot, acquired from other researchers, or, when the text is available in printed form, converted into machine-readable form by an optical character reader.

The language corpora at Logotheque is synchronic, and except for a corpus of Strindberg's works, dates from post 1965 and is principally of the written language. The language corpus contains some 30 million running words: novels (69 published in 1976 and 60 in 1981, comprising approximately nine million words), legal texts (about 500 000 words), reports of the proceedings of the Swedish Parliament (1978—79), about four million), daily newspapers (1965, 1976, about 2.3 million), and weekly magazines.

An author corpus of between six and seven million words of the complete works of August Strindberg is also being encoded at Logotheque. — Word corpora within Logotheque comprise about 200 000 entries, including the vocabularies of the Word-list of the Swedish Academy and the Frequency Dictionary of Present-Day Swedish based on newspaper material. A project at Logotheque called "Lexin" is creating dictionaries for various immigrant groups. Along these lines a Swedish dictionary that serves as basis for translation has been developed and will be published by Språkdata. It contains about 15 000 words — copies of data in machine-readable form, as well as of concordances of encoded texts are available for academic research. For information, write to: Logotheque, Språkdata, Göteborg Universitet, Norra Allégatan 6, 3—413101 Göteborg, Sweden.

5. Literature (selected)

S. Hockey 1980 · IDS 1982 · W. Lenders 1985 · D. Rood 1980.

Robert F. Allen, Piscataway, New Jersey (USA)

14. Segmentierung in der Computerlinguistik

1. Problemstellung

Das Problem der Segmentierung in der Computerlinguistik hat seinen Ursprung darin, daß

erstens sprachliche Phänomene, besonders in der gesprochenen Form, nicht von vorneherein hinreichend deutlich in ihre Segmente zerlegt und in ihrer Funktion bestimmt sind, daß

aber *zweitens* für die Arbeit mit dem Computer auf diese Segmente, nämlich auf die Laute, die Silben, die Wörter, die Sätze, die Sinneinheiten und ihre Beziehungen zueinander zugegriffen werden muß.

Zwar wird in der geschriebenen Form mancher Sprachen eine Segmentierung be-