IV. Computer-Aided Description of Language Systems I: Basic Problems Computergestützte Beschreibung von Sprache I: Grundprobleme

11. Theoretical Aspects of Computer-Aided Description of Language Systems

1. Introduction

Linguistics is mostly treated as an empirical science; it is only rarely denied that linguistic theories and their subparts are hypotheses which should be checked as for their relative adequacy with respect to some kind of data. Even Chomskyan linguistics with all its emphasis on generalizations to be reflected by formal frameworks, including also substantial portions of deductive reasoning, treats linguistics as an empirical domain; the question whether the observational data serving as a basis for evaluating the adequacy of hypotheses should consist (entirely, or primarily) in introspection or in recorded discourses is substantial, as we shall see later; however, it is not immediately relevant for the basically empirical nature of the science of language.

Montague's (1970) view of natural language as describable by the same methods by which logicians describe the formal languages they themselves create cannot be fully accepted: natural language exists as an object which is to be described, rather than constructed; its grammar, lexicon, and semantics have to be examined. In order to find out how natural languages are structured, it is necessary to pass over from what Chomsky calls the level of observational adequacy to that of descriptive adequacy and, whenever possible, to reach at least some stages of explanatory adequacy.

Even if we do not go so far as Yngve (1983) in assuming that a truly scientific linguistic research cannot be "disciplined by grammar", it is beyond any doubt that already the way from sound to phoneme, and the more that from the observational data to a description of structures, requires a complex progression by means of trial and error, in which the specification of grammar can only be understood as grasping prototypical cases, directly related to linguistic norms (which are fully identical neither with usage in its full range, nor with a codification of the norm by linguists and their institutions); it is not the task of such a specification of linguistic norm to handle all possible deviations, although many of them are easily found in the usage being accepted in discourses and easily understood by hearers. A theory of possible deviation may itself be grounded on core principles of a grammatical theory.

The sources of empirical material for linguistic investigations have always been limited to a few kinds of data. The following four cases can be mentioned:

(i) A corpus of written or spoken discourses and their subparts, ranging from excerpted occurrences of expressions chosen according to a specific research topic to a general-purpose set of data. As was stated e. g. by Bergenholtz and Schaeder (1977 a), a corpus can never present a fully general characterization of a living language with its varieties of different kinds; always a more or less representative choice has to be made. Under Chomskyan views of the dichotomy of competence versus performance, corpus data were relegated to the field of performance and thus believed, rather than found, to be unsuitable; see Aarts and van den Heuvel (1985, 304), who point out that since the middle of the 1960's numerous attempts to reestablish the corpus as a valuable research tool have been made, and that an appropriate relationship between corpus data and introspection can be established if corpus is understood as a test bed for hypotheses about the structure of the corpus language.

(ii) Introspection. If linguists use their own intuitions about language as the basic source of observational data, this can do for the core of the structure of language; however, as has been well known in structural linguistics, and confirmed by the recent experience of discussions within generative grammar, for more subtle questions the subjective nature of these data often leads to discrepancies concerning the "dialect" (or idiolect) that is being analyzed. Although discussions often stumble over "some speakers finding this set of sentences rather unacceptable than merely odd", etc., introspection cannot be excluded from the sources of observational data, since without it a great part of linguistic phenomena could not be properly discussed, since they occur in actual discourses so rarely that practically they cannot be found in a corpus. Thus, the formulation of general hypotheses (rules, restrictions, principles) would be impossible without using introspection together with other sources of linguistic knowledge.

(iii) Informants. When testing informants, intuition is again present as a source of observational data; however, it is important that in this case one works with the intuition of a neutral speaker, rather than with that of the linguist him- or herself (who is interested in the results, and who may become uncertain as to the appropriate use of the given item after a certain amount of introspective effort); moreover, as recalled by Rieger (1981 b, 200), it is possible and useful to observe the informants' behavior in a relevant situation or their reactions to characteristic contexts, rather than to address their intuitions directly. Also the study of aphasias and the results of various psycholinguistic tests have been helpful.

(iv) Experience with applications. Important sources of knowledge for linguistic research have always been found in the more or less successful, partial or complex results of applications of linguistic theories in language teaching and learning, translation, standardization and 'language culture', including the study of errors made by native speakers, speakers of a second language, translators, and others.

Under this view of the nature of linguistic research, it can be stated that computers will be more and more useful in what concerns points (i) — compilation, maintainance, updating and exploitation of corpora (as well as of databases arranged for the aims of linguistic research) — and (iv) — extension of the applications of linguistics. Due to the frequent and manifold cases of ambiguity, vagueness and irregularity intrinsic to natural language, it does not seem probable that computers will play a major role in the formulation of hypotheses (descriptions, theories). They can be expected to present an efficient help in preparing the data for human insight to formulate such hypotheses, and in testing them as for their adequacy and consistency. Only when a theory has achieved the shape of a formal framework, also an automatic proof system can be used for its further development. Nevertheless, although the role of computers in linguistics thus appears to be mainly that of an auxiliary tool, it can be seen that already in the last two decades the influence of the use of computers on theoretical linguistic discussions has been quite substantial. In the future the broad spread of computers and of their networks certainly will be of even much greater impact on theoretical linguistics. In the following paragraphs we want to help to elucidate this apparent paradox. Since the size of a corpus (especially of one with a general purpose) should be as large as possible, the use of computers in arranging and updating corpora is of a substantial help even if no linguistically specific software is applied. With specific programs it is possible to use linguistic (especially lexical) databases, to tag a corpus (i. e. to add grammatical and possibly other information to its individual word tokens; the basic data specified by tagging is the word class), to parse a text, to apply various simple and complex statistical techniques, and so on. All these activities have been often performed manually, and if done by computers, they generally require some kind of prearrangement (compilation of a database, pre-editing a text) and/or of postediting (emending the mistakes made by the program). It is here where the use of computers as tools for the testing (and thus partially also for the formulation) of linguistic hypotheses starts.

2. Towards a Classification of the Applications in Linguistics

A relatively complete survey of the various ways in which linguists use computers is presented by Butler (1985 a; 1985 b, Sect. I.2), who points out to what degree even such rather simple tasks as lemmatization (assignment of the basic form, i. e. of the lexical unit corresponding to an inflected form) presents difficulties to computerization (i. e. to an algorithmic and implementable treatment). Also the issues concerning style and literature theory (disputed authorship, chronological sequencing), which cannot be given due attention in our contribution, are treated by Butler, as well as the ways in which instructions and textual data can be fed into the computer, and results extracted from it; ready-made 'package' programs determined for a naive computer user are briefly described, and some practical aspects of the use of microcomputers and of programming languages designed specifically for applications in text analysis (SNOBOL 4, ICON) are amptly displayed in the quoted book.

It should be noted that the research in computer linguistics has always been typically scattered; the existence of a research group or institution of more than 15 people has been at most exceptional. Even if some of the approaches to the use of computers spread from one group to several others, a unified broad trend or complex project never came into being. Up to the end of the 1970s theoretical linguistics seldom paid a systematic attention to the aspects of research connected with computers. The parsers included in machine translation systems, as well as Winograd's (1972) analysis were based on linguistic theories not belongig to the most widespread trends.

Transformational grammar was used as the theoretical basis for an automatic syntactic analysis (parser) since Petrick (1973) and for a system of random generation of sentences by Friedman (1969 b), both concerned with English. Later, an approach to parsing related to transformational description was developed by Marcus, M. P. (1980) in his deterministic parser. However, as Ramsey notes in art. 18 of the present volume, sect. 3.1.1., for many researchers the indirect way in which transformational description relates meaning to sound appeared less intuitive and economical than a sequence of operations leading directly from sound to meaning and from meaning to sound, although the asymmetrical dualism of linguistic sign (i. e. the existence of ambiguity and of synonymy) makes it necessary to work with a one-to-many relationship in each direction (see also Sgall, 1968).

Transformational grammar thus has had much less influence in computational linguistics than in linguistic theory. It remains to be seen whether the more recent form of Chomsky's (1981 a; 1982) theory (the 'government and binding' theory, in which the X-bar syntax and a set of general principles with differences in parameters substitute the older abundancy of syntactic rules) is more suitable as a basis for computational elaboration. Berwick and Weinberg (1984) present a fundamental analysis of the relationship between the current theory of transformational grammar and the aspects of language use (performance), which are represented by parsing; cf. also Barton (1984), where certain preconditions of a substantial adjustment of parsing to the modular theories of syntax are discussed.

Various approaches have attempted at a more immediate treatment of the relationship between linguistic competence and performance. After such older experimental systems as those of machine translation and natural language understanding, in the 1980's several approaches have endeavoured both a theoretical description of language structure and an implementable account of language use. Some of them are mainly theoretically oriented, but present also means for a computational handling of parsing and of other phenomena from the domain of performance; this concerns first of all Gazdar's generalized phrase structure grammar and Bresnan's lexical functional grammar (see the bibliographical data on these and other approaches in article 18, 31 and 32 of the present volume). Also Montague grammar, which originally attempted at a general account of syntactic-algebraic operations, rather than of mere concatenation, has found computational treatments in several projects. The procedurally based theoretical net-linguistics has been investigated from the viewpoint of parallel programming; see Schnelle (1981; 1985), Rothacker (1985).

With other approaches, such as those by Kay, Woods, Pereira and Warren, or Joshi, the computational aspects have perhaps been the first objects of research, but the theoretical issues are systematically elaborated, too, with important insights. Among these, one of the most interesting concerns the notion of unification, used by several kinds of description and characterized briefly by Karttunen (1984; 1985). Vijay-Shanker and Joshi (1985) and the writings quoted there present a valuable comparison of several frameworks.

In Europe, several approaches based on dependency syntax have been formulated; as for their American ancestors, see Hays (1964 b) and Robinson (1970). Kunze (1975; 1982 b) and Hellwig (1978 a) describe German, Apresjan (1980) works first of all (though by far not only) with Russian, Nagao et al. (1984) with Japanese, Charpin (1985) with Latin; Sgall (1967 a; et al. 1969; 1986) and Plátek et al. (1984) present a framework that should cover Czech and other Slavic languages as well as English.

Most of the just quoted approaches want to characterize the relationship between sound and meaning in a more direct way than transformation grammar does. However, due to the scattered character of both theoretical and computationally oriented research, we still are at the beginning of a deeper understanding of how a theoretical description of the system of language should be conceived to be directly implantable in a systematic description of the regularities of language use.

Coming back to the questions of how to classify the various uses of computers in linguistics (without which an account of language use hardly would be feasible), we want to point out several criteria which may be understood as a useful basis for a classification. We discuss the linguistic aspects of the use of computers here rather than the issues of computer software, so that criteria concerning the kinds, structures and functions of computer programs and their packages are not considered. This is given by the fact that the present author as a linguist does not dare to estimate the main points of the next evolution of the programming techniques, the trends of which change rapidly due to the quick development of the hardware; if the efforts aiming at a further substantial change of programming itself are successful, bringing programming nearer to the human communication in natural language, then the software techniques may differ quite substantially from those of our days. However, the empirical problems of linguistics and the nature of forming hypotheses on the basis of human invention themselves will remain.

In future research, the criteria we discuss may be compared as to their impact on the classification of the linguistic uses of computers and to their clustering properties, which would be useful if a typology of computer applications in linguistics is to be established.

2.1. The first of them concerns the fact that most linguistic theories work with several levels of the structure of language — from phonetics and phonemics (not to speak about graphemics) through morphemics to the grammatical structure of the sentence (surface syntax, etc.) and to one or more deeper, underlying sentence structures; the patterning of discourse, which belongs to the domain of human (communicative) action, rather than to the structure of language, as well as the cognitive structures of various kinds (intensional semantics, cognitive networks and scenarios known from artificial intelligence), require other means of inquiry than those proper to linguistics in a narrow sense. The complexity of natural language, which is at the source of the generally accepted principle of its modular description using a division into levels, also underlies the fact that the use of computers in natural language description is much more complicated a task than might be assumed at the first glance. There is a far reaching difference between easily accessible levels of language structure, the patterning of which can be handled by a computer program relatively easily, and other levels, an automatic treatment of which is much more difficult.

Thus, much has been done already in the use of computers for statistical and structural inquiry of graphemics, phonemics and some subdomains of morphemics, especially of lexemics. Statistical studies based on corpora are characterized in article 12 of this Handbook; the whole domain of quantitative linguistics, including taxometrical and other complex methods, is relevant here. It can be seen in article 15 and 16 that even the relationship between phonemes and graphemes is far from being simple; see also Sgall (1986). In what concerns a computerized description of morphemics, there are first of all the difficult questions how to formulate a morphemic analysis so that it is capable of handling also the newly coined or encountered words, as far as they are inflected (and derived) in accordance with productive paradigms; an effective generally linguistic algorithm, which can use different sets of input data to describe different languages also belongs to the interesting goals in this domain.

On the other hand, the levels concerning syntax require first to reach a deeper understanding of the kinds of mutual relationships between the elementary and complex units of the individual levels; here the differences between the approaches to theoretical linguistics are more relevant for computer oriented research than they are elsewhere. Due to the empirical character of linguistics, it is necessary to study the observable phenomena in such a way as to understand their interrelationships and thus to go deeper, from sounds to phonemes, from morphs to morphemes and to syntax, from lexical morphs by means of lemmatization and disambiguation to lexical meanings, from morphemes and syntax to the layers of meaning (including both semantic and pragmatic aspects) which are closely connected with Frege's principle of compositionality (according to which the meanings of complex expressions are determined — directly or indirectly — by those of their parts).

As we have stated, most linguistic approaches distinguish also different levels of sentence structure (one of which, according to some approaches, can be understood as a level of disambiguated, though literal, sentence meaning). It is also generally accepted that lexical units have to be classified into word classes (though only for prototypical cases there are generally accepted criteria for this classification) and that, in the structure of a sentence, the verb occupies a central position. Not only verbs, but also nouns, adjectives and adverbs (as well as some pronouns) have their valency (case) frames, or theta role sets; some of the slots in these frames are obligatory, others are optional; some slots (i. e. some kinds of complementations) are inner participants (deep cases), others are free adverbials. In some approaches more or less tentative enumerations of the different kinds of complementations were formulated, but the underlying criteria of such a classification are discussed rather rarely.

To what extent, in what parts of these subdomains of language description, and in which ways can computers be used?

(a) Much experience has already been gained in applying computers to assemble, identify and count the units of the relatively accessible levels, using more or less standard statistical routines for an evaluation of the identified relationships; let us add that spoken input still can be handled only to a limited extent (if lists of words or of short commands for robots, and so on, are given; an acoustic analysis of connected speech has not yet been made available for computers at a broader scale), so that even the data concerning phonemics have to be gained from printed or typed input, i. e. either via an algorithmic transition form graphemes to phonemes, which is much easier e. g. for Spanish or Finnish than for English, or from phonetic or phonemic transcriptions.

(b) Computers can be used to identify the regular (frequent, or otherwise determined) combinations of units of the accessible levels — e. g. cooccurrence of graphemes, or of word forms, an so on.

(c) In accordance with the deeper relationships stated (i. e. hypothesized) by researchers, computers can handle (perhaps in interaction, or with pre- and postediting) also deeper issues; this concerns the tagging of a corpus as well as testing a (part of a) grammar by means of a system of random generation of sentences or by other means (as for the corpora, see art. 12 and 13 of this Handbook; computational testing of language models is discussed in art. 21 and 22).

Since syntax belongs to the relatively unaccessible layers of language, it is not surprising that also article 18 of this Handbook, devoted to the use of computers in connection with syntax, is much more concerned with syntactic frameworks used for experimental and applicational goals, i. e. first of all with parsers, than with the use of computers for the proper aims of linguistic research itself. Articles 31 and 32, directly devoted to parsing as simulating natural language syntactic processes, may be understood as corroborating the view that an extensive use of parsers for practical purposes, which can be expected to take place in the future decades, will serve for practically important and therefore urgent efforts devoted to testing syntactic theories and comparing them for their adequacy and efficiency. The epoch of widespread use of machine translation and various kinds of systems of human-machine communication will also be an epoch of big institutes active in large scale automatic data processing, the extensive experience in which will provide a valuable practical groundwork for future research; see Plath (1963, 52).

In such difficult domains as syntax, this indirect effect of the use of computers (designed first for practical applications, and only then to test the theoretical assumptions) probably will be more important than a direct exploitation of computers for the purposes of linguistic research in the sense of paragraph (b) above.

2.2. Thus we have come to the second important criterion of our classification: the computer is either used for the purposes of practical applications, or for the aims of linguistic research itself.

Among the applications some — concerning the lower layers of information retrieval and natural language interface to databases — are linguistically rather uninteresting. Non-trivial tasks concern either deeper layers of text retrieval, see esp. Karlgren and Walker (1983) and the systems of Thinking Machines, Inc., or general-purpose transportable user oriented query systems, which are user-friendly e. g. in that they avoid stonewalling, see Joshi et al. (1984), Kaplan (1983), Zoeppritz (1984). Grapheme-to-phoneme conversion systems, on which more can be found in articles 15 and 16 of the Handbook, are useful for linguistics itself as well as for a number of practical applications including speech synthesizers.

Many other kinds of applications might be enumerated; among the linguistically most interesting ones there is the large domain of machine translation. The attempts at a rapid word-for-word translation, based on comprehensive lists of lexical and phraseological units, may be practically useful, although the inadequacies of their output often lead to their development into (or replacement by) other systems, especially those of machine aided translation (interactive or not, with 'automatic dictionaries', text processors and other tools) and so called second-generation systems, using linguistically based parsers; now see esp. Slocum (1985), Vauquois and Boitet (1985), Thouin (1982), Nagao et al. (1985), Wilks (1983 a), Rohrer (1986). Methodically, some urgent problems of this field concern the complexity of parsers and the difficulties connected with the incompleteness of linguistic knowledge, see Nagao et al. (1984), and with a modular handling of syntax, which should make it possible to add new rules (or change old ones) for desired amendments of the parser (while avoiding unexpected side-effects). For this purpose often a specific software is aimed at, which should keep the grammar apart from the processor itself, and thus allow the linguist to write down the rules using a usual linguistic notation. Since not only linguistic, but also factual knowledge has to be used at different stages of translation, one either works with systems requiring a considerable amount of human postediting, or one prepares experiments with systems which also include a knowledge base and/or rules of inference, reasoning schemes, etc., i. e. which work with natural language comprehension in the sense of artificial intelligence, now see Bátori (1986), Saluveer and Õim (1986); cf. also sect. 5 below. Machine translation represents an extensive repertoire of projects, some of which certainly will become important for practical aims and then also for broad practical testing of linguistic hypotheses (cf. what we stated at the end of sect. 2.1.). Even more challenging for linguistics (since requiring a deeper syntactico-semantic and pragmatic analysis of texts) are such tasks within the experimental domain of artificial intelligence as automatic compilation of knowledge bases from input texts in natural language, natural language contact with expert systems, general systems of communication with robots, interfaces to teaching systems, and so on. As Sparck Jones (1984) states, projects of these kinds probably will replace database queries as stimulating tasks useful for fundamental research on natural language understanding. An experiment with a knowledge base compiled from text with the use of a syntactico-semantic analysis and of inference rules (combining two utterances into a complex one, e. g. with a relative clause, shortening the utterances which contain unnecessary items, taking causal clauses out of complex sentences, operating with hyponymy, etc.) is reported by Hajičová/Sgall (1984); for a project using PROLOG and Kamp's discourse representation structures, see Frey et al. (1982); Wahlster (1981) analyzes several problems of argumentation and approximative inferencing in dialog systems. A characterization of the dialogue oriented projects carried out in Hamburg was presented by Hahn et al. (1980), Hoeppner/Morik (1983).

Among other kinds of the use of computers oriented to practical applications of linguistic knowledge, there is computer assisted instruction in foreign language teaching; its perspectives, as discussed e. g. by Pusack/Otto (1984), appear to offer manifold possibilities from grammar practice and vocabulary acquisition (with the new words occurring in typical context positions) to linguistic databases serving as reference utilities; on the other hand, it seems that in the domains of pronunciation and conversation the use of computers will remain rather limited in the next future; the Athena project of M. I. T. is one of the few relevant attempts. We cannot dwell here on such applications as text editing or detection of spelling errors and misprints, although here linguistics also plays an important role (cf. e. g. the IBM Epistle system for a correcting grammar).

The linguistically oriented practical applications of computers are not limited to technologies, to an intensification or routine skills. Their aims include making humans freer for creative activities. Moreover, as Burns (1984) illustrates, in such domains as computer-assisted rhetorics and stylistics the humanist's goals in aesthetic understanding are directly pursued.

There are various possibilities how to combine the use of computers for practical applications and for instrinsic linguistic purposes. One of them, namely a parser designed for interactive use during which a (restricted) linguistic description can be constructed and tested, is characterized by Philipps/Thompson (1985).

Up to now computational tools have been used by linguists mainly for applications in this or that kind of translation (from one natural language to another, or to a code concerning a database, to a knowledge representation, and so on); however, now the conditions are present for computational tools specifically geared to doing linguistics to become less rare: powerful computers are less expensive now, and many computational linguists have realized that natural language processing needs grammars and other essentially linguistic components; see Gazdar (1985). This spreading consciousness of the necessary connection between a description of language structure and the use of computers concerns also many theoretical and structural linguists, who are becoming more aware of the advantages of using specifically linguistic software for the aims of a complex description of language.

The use of computers directly for linguistic aims can again be divided according to more specific purposes; the main dichotomy can be seen here in the difference between systems used for linguistic analysis itself and those used for testing the hypotheses formulated on the basis of an analysis.

In the former domain the most important area can be seen in the computerized or computer assisted analysis (tagging) of corpora, to which we return in sect. 3 and 5 below. Other areas concern computer aids for subdomains of linguistic research. Such a computational tool designed for linguists developing grammars of the shape of generalized phrase structure is described by Evans (1985), who stresses that his system, though large and still difficult to handle, helps to 'debug' a grammar which (due to the complex character of natural language) cannot be effectively tested without such a tool. The use of computers in simpler tasks than grammar writing — e. g. in compiling concordances (even if they are more sophisticated than the known KWIC system) or dictionaries of different kinds (including frequency and reverse dictionaries) — is much more straightforward; the most interesting systems of this kind in lexicology are lexical databases, some comments on which can be found in sect. 4 below. In the domain of discourse structure computers are used for research concerning experimental systems of artificial intelligence (see sect. 5), and many of these projects and publications are of direct importance for theoretical linguistics, especially for the theories of discourse (or text) structure; here the main difference between 'sentence linguistics' and 'text linguistics' should be preferably seen in the opposition of language system versus language use (communicative action), rather than the question whether sequences longer than a sentence are examined, since a discourse can consist in an occurrence of just one sentence (i. e. in a single utterance, which differs from the abstract sentence in occurring in a specific situation, connected with reference assignment concerning the referring expressions it contains, and so on). Statistical and other empirical analyses of text structure are effectively facilitated by partly or fully computerized text corpora (see sect. 1 (i) above and sect. 3 below). Also studies concerning language development and reconstruction can use computers, see e. g. M. Johnson's (1985) computer assisted comparative dictionary, serving also to tracing regular sound changes which can be supposed to have occurred in the languages compared (the first test concerned the Yuman family in Northern America). Indo-European etymologies also have been similarly treated by various research groups.

It seems that no part of linguistics will remain untouched by a direct impact of computer use in the next years. Computers can play a role in forcing rigorous consistency of the linguist's formulations, in testing correctness and completeness of analyses, and in comparing disparate approaches to linguistic phenomena; if certain conditions of linguistically appropriate design are met (with the use of such operations as unification and others), then not only grammars can be tested by means of automatic analyses of sentences, but it is also possible to determine whether certain properties of a grammar provably obtain and whether some axiom of one theory might be provable in another; see Shieber (1985 c). Several systems have been so designed as to be useful for different tools — experimental as well as descriptive and others; besides PATR-II, constructed at Stanford Research Institute, this concerns the Lexicrunch system, described by Golding and Thompson (1985), which can be used in various kinds of work with lexical units, morphemics and word formation. Another multi-purpose system, which can be used as a parser, a hierarchical database system and as a system for queries of information written in natural language is characterized by van der Steen (1984).

Thus the division of computerized systems into those serving for practical applications of linguistics and those useful for linguistic research itself is not quite clearcut, which certainly cannot diminish the importance of such a classificatory criterion.

2.3. From what was stated in the previous paragraphs it follows that also the relative complexness of the aims pursued by the use of computers in linguistics may be used as a classificatory criterion. With respect to such subdomains of language as graphemics, phonemics, or (partly) the lexicon, it is possible to distinguish several degrees of complexity of the tasks of automated research. Statistics is applied rather easily by computers, even if complex mathematical routines are used, which often is highly useful. Also a compilation of indexes and concordances of different kinds (from KWIC to those connected with lemmatization, i. e. determining word forms as belonging to a certain lexical unit) can be classed as not much complex; even so, a certain amount of human interference is useful, whenever ambiguity is present. Methodically much more complex are such tasks as those concerning a classification of items according to their intrinsic features (as far as these are not immediately accessible to computers), and almost all interesting tasks concerning the levels of the syntax of the sentence, not to speak about discourse patterns.

2.4. Another important opposition, the impact of which has been brought to my attention by Karen Sparck Jones (personal communication), consists in the difference between systems helping the linguist to learn something about the object of her/his study (i. e. about something else than the computer) and systems serving to find out how computerized language processing itself should be done. Systems of the latter kind are those belonging to basic research, rather than to the domains of application; they are useful for testing hypotheses concerning the nature of language and its use. If e. g. automatic reference resolution procedures are designed and processed, they belong here if they are used for research in psycholinguistics, rather than for practical aims of information retrieval; this concerns also experiments in summarizing texts, in machine translation, and so on.

For the future development of computational linguistics, systems of the latter kind certainly are of major importance. As we have seen in sect. 2.1., computer assisted research of those levels of language which are not easily accessible to automatic procedures hardly can become efficient on a large scale without these systems, and, moreover, a secondary effect of extensive practical applications probably will be substantial for the study of such levels, as well as for the research in discourse structure, i. e. in the use of language, and in its psychological background.

2.5. Classification itself is just a useful tool for a better orientation in the methods appropriate for individual subdomains, and its criteria certainly will be enriched on the basis of future development of computational linguistics itself and the surrounding domains from hardware to cognitive science. In the given situation, when the interest in the use of computers in linguistics keeps growing and the possibility of large-scale projects may be foreseen, it is important to prepare a division of labor and of partial tasks between the relevant domains of science. Linguists should be prepared to cooperate with specialists in computer science more broadly and intensively than has been possible up to now, since the programming of complex procedures, which are necessary even for the processing of restricted domains of natural languages, needs special insight into look-up methods, ways how to master the combinatorics of syntactic relations at various levels, how to choose and handle appropriate interactive means, and so on. On the other hand it should be kept in mind that a thorough exploitation of all the results of traditional and structural linguistics is highly useful for computers to be used in a rational, economical manner; a proper treatment of the individual features of various languages requires good knowledge of what descriptive and contrastive linguistics, as well as the typology of languages can tell us about the languages of the world. A suitable starting point for a systematic cooperation between linguists and logicians seems to consist in recognizing a language specific patterning of literal meaning, be it called the underlying structure of sentence, the tectogrammatical level, or otherwise. The tasks of linguistics then concern the description of the relationship between this level and the outer shape of sentences and of discourse, while the logicians' cooperation is needed to handle the correspondences between meaning in the linguistic sense and truth conditions, or, perhaps, other fundamental layers of semantics. A treatment of pragmatics and communication activities cannot be complete without collaboration with psychologists, whose role in cognitive science and artificial intelligence should be further increased. Thus, a varied interdisciplinary research will be typical for the future of the domain.

In the following paragraphs we want to characterize briefly the issues of those subdomains of linguistics which have already achieved a significant level of computerization.

3. Corpora

It can be easily understood why English is the language best equipped with general-purpose corpora available in machine-readable form. The specific problems of constructions and representativeness of corpora can be found in article 12 of the Handbook, and a survey and typology of the existing corpora is presented in article 13; for a concise information and bibliographical data on the research based on the main corpora, see also Butler (1985 a, 59). Here we comment briefly upon some questions of their use in the context of computational linguistics.

It should be noted that for the first large text collection, the Brown corpus, containing 500 texts in different styles of printed American English (in total length of approximately one million word tokens) a concordance and a tagged version are available on magnetic tape. While the Lancaster-Oslo-Bergen corpus is its counterpart for British English, sharing the quoted parameters, the London-Lund corpus is smaller, its importance consisting first of all in that it contains samples of spoken discourse (in educated British English), derived from the Survey of English Usage, see Svartvik/Quirk (eds.) (1980); the samples are equipped with prosodic marking and cover different styles from spontaneous conversation to lectures.

Corpora have different uses in linguistics and its applications, from helping "ordinarily working grammarians" to testing hypotheses. For tasks concerning the less accessible levels of language and language use, a tagged corpus is necessary. Tagging itself can be computerized to a high degree, since the bulk of supplying the words or word forms by appropriate labels of part of speech membership and of other morphemic (or some syntactic) properties can be made on the basis of algorithms; for a language different from English, see Karlsson (1985). In some typical cases a tagging program employs look-up in a dictionary and, for word forms not found there, in a list of endings and productive suffixes; whenever more than one tag can be chosen, the most frequent one can be automatically preferred; also heuristic rules on context frames were formulated. However, manual testing and correction are always needed.

Problems connected with the compilation of a large corpus were discussed, on the basis of the experience with Brown corpus, by Francis (1979); one of the main points consists in the specification of future inquiries for which the prepared corpus should be used; in the case of such a generally usable collection as with Brown corpus, the shape of the corpus has to be adjusted to the needs of all tasks which can be foreseen. Further requirements on a corpus are formulated by Maegaard and Ruus (ms.), who state that if a corpus should represent a text type as a whole, a systematic influence of individual texts must be excluded (by including only short segments of single texts, chosen by means of random numbers); moreover, a corpus should be homogeneous (representing a single text type).

It is necessary to remove the prejudices against the use of corpus data that resulted from the one-sided preference of linguistic competence, rather than performance, as the object of linguistic research. The intuitive data gained by introspection play an essential role in linguistic inquiries, but the corpus appropriately serves as a test bed for the linguist's hypotheses; see Aarts and van den Heuvel (1985). A computerized corpus requires the means of description to lend themselves to automation, i. e. to be explicit enough for implementation; thus, corpus-based research belongs to computational linguistics, with an emphasis on large quantities of data to be analyzed. As the authors further conclude, corpus linguistics is descriptive, rather than theoretical, in that its primary objective is the description of individual languages, rather than an inquiry into the language faculty (in Saussurean or Chomskyan terms). Moreover, a corpus is a collection of historical linguistic events, so that corpus linguistics deals with actual language use, and thus also with language varieties. Ideally, also the linguistically relevant aspects of the situation of the discourse should be dealt with, and the phenomena of register (fiction, drama, primary dialogue, ...), medium (spoken, printed), and style are to be investigated, e. g. by means of a set of corpora distinguishing different language varieties (codes), rather than representing cross-sections of different varieties of the studied language. The basic tool for the analysis of a corpus (tagging included) is a formal grammar (rather than a computer program), which is relatively independent of the computer and therefore allows a discussion about the hypotheses embedded in it. The authors add that the grammar should be automatically convertible into a parser, which also is the case of their 'extended affix grammar', working with two subsequent syntactic components, each of which is context free, though the pair as a whole is more powerful. Such generative power is proper also to the description characterized by Plátek/Sgall 1978 and Plátek/Sgall/Sgall 1984.

Since the tagged or analyzed corpus should serve as a linguistic database the users of which come from various linguistic backgrounds and are committed to different kinds of language research, the classification of words, morphemes and relations in the corpus should be acceptable to a large variety of linguists and therefore fairly traditional labels are preferable, as Aarts and van den Heuvel (1985, p. 310, 333) argue. Moreover, consistency in the analysis of the utterances in a corpus has to be guaranteed, even though the analysis is typically executed by a team. To this aim the authors (pp. 318 ff.) have designed a 'linguist's workbench', with which the decisions made by the linguist at the terminal are recorded in order to make the process that has resulted in a set of analyzed utterances fully reproducible. The grammar underlying the analysis cannot be changed from within the workbench; the latter organizes the communication between the corpus, the grammar, a specifically designed database system (language independent, menu driven and standing close to that of relational databases), the lexicon, a local lexicon (making it possible to extend the lexicon for the duration of one session, e. g. in proper names, jargon words and other expressions likely to return in the subcorpus under consideration, although not included in the main lexicon), and the 'logbook' file, in which the linguist's actions are automatically recorded. The linguist has a set of automatic functions at her/his disposal here (such as 'analyse', 'look up in lexicon', 'extend lexicon'), as well as compound functions, mainly cycles, going through the necessary steps of the analysis of an utterance or of a sequence of utterances. The linguist's interventions are necessary with words not found in the lexicon, perhaps with the specification of sentence boundaries, and first of all whenever the analysis of a complex expression could not be completed automatically (for this case it is useful to include an automatic indication up to what point in the utterance the analysis has proceeded).

It seems that often the unclearness of classification criteria and the fuzziness of linguistic phenomena lower the reliability of the traditional kinds of tagging and its usefulness for the testing of theories. — It should be mentioned here that the Longman Dictionary of Contemporary English (see also Michiels et al., 1981) now is accessible on-line; a machine readable version of the Oxford Second Language Learner's English Dictionary, where case frames are associated with the lexical entries, will be of great importance.

Semiautomatic tagging (with pre- and postediting), its developments and perspectives are discussed on the basis of the experience with the Lancaster-Oslo-Bergen corpus by Leech et al. (1983), see also Beale (1985). An interactive tagging procedure in four levels, concerning the London-Lund corpus, is described by Svartvik (1984): about 100 word-class tags are assigned to the word occurrences in the text at the word level, five kinds of phrases are distinguished at the phrase level, five syntactic positions are identified at the clause level, and — no sentence level being handled in this approach to spoken texts — such expressions as *I'am sorry, never mind* , etc., are classified as appologies, smoothovers, and so on, at what is called a discourse level.

4. Lexical Databases

Lexical systems in which the entries are organized according to principles of database structure, with associated information accessible at different levels, make it possible to use these data in ways for which the traditional sequential organization in alphabetical order is not suitable (although sequential dictionaries on computer media certainly are useful for various aims). As characterized in Calzolari (1983 d) and in the writings quoted there, the Italian Machine Dictionary (DMI), consisting of a set of more than 100 000 lexical units, a set of more than a million word forms and a set of 'definitions', with morphological, syntactic and semantic data on each of these items, substantiates the view that the lexicon can be represented as a dynamic system belonging to the core of a computational description of language. Calzolari (1984 a) points out how hyponymy, and also certain syntagmatic relations between lexical units can be interactively evidenced. This aspect, as well as the nature of search strategies in DMI and the easiness of frequency counts, concordance compilation and other look-up tasks are stressed by Calzolari and Picchi (1984).

Perspectively, computerized dictionaries will be used to a variety of aims, such as identification of inflected forms, correction of misspellings, finding lexical definitions, synonyms, etymologies, and so on, as Kay (1984 b) points out on the basis of the implemented version of the American Heritage Dictionary. A comprehensive computerized dictionary of English is being prepared by the Oxford English Dictionary project at the University of Waterloo.

An account of the major machine readable dictionaries of German (which were compiled for the aims of machine translation, question answering, information retrieval, and so on), is presented by Heß/Brustkern/Lenders (1983), where the ways are discussed how to integrate these dictionaries and to provide the resulting database with a homogeneous structure. The implementation of the collected lexical database of German in the Institute for Communication Research of the Bonn University in cooperation with the Institute for German Language in Mannheim and with the Saarland University in Saarbrücken, is described by Brustkern et al. (1986); the database, organized in accordance with the system SESAM, contains 300 000 word forms, a great part of which is equipped by relatively rich grammatical data.

In a project of computational lexicology in Aachen a procedural model of "spreading activation", which uses cluster-analysing methods, has been prepared, see Rieger, B. (1985); as the examples suggest, even with an enormous corpus it would be difficult to distinguish pure or partial synonyms from such frequently cooccurring pairs as *Geschäft* vs. *Kenntnis,* or *Elektron* vs. *Computer.* On the other hand, in language understanding systems the representation of semantic dispositions resulting from the system's function can serve as connotative default values in solving ambiguity and vagueness problems, as the author states. A numerical measure of meaning is supposed to have to be based upon structural properties of open sets and dynamically organizing systems. Similar procedures have also been elaborated at Thinking Machines, Inc.

5. Syntax, Meaning, and Discourse Structure

An appropriate basis for computer assisted inquiries into the less accessible levels of language structure and into text patterns consists in tagged corpora. Several reports of their exploitation for the aims of syntactic analysis were presented at the 6th ICAME Conference on English Language Research on Computerized Corpora, see the summaries published in ICAME News 10, 1986, 10—61. A possibility how to analyse relative clauses on such a base is outlined by de Haan (1984 b). Several further projects using tagged corpora for specific tasks of linguistic description are reported in other volumes of

ICAME News. How the way from tagging to a parser can be thrust by means of a calculation of probabilities for competing structures, is illustrated (on the basis of techniques successfully used in tagging the Lancaster-Oslo-Bergen corpus) by Leech and Garside (1985). A possibility to build a parser with the aid of a computer was suggested already by Kulagina (1962). A way in which parsing can be based on lexical information has been characterized by the Word Expert Parser of Rieger, Ch. and Small (1979); now see Eimermacher (1986) on one of its recent versions, which allows for using the contextual properties of words to the aim of restricting the necessity of backtracking as well as of abundancy of syntactic rules. Berwick (1983 a) points out how word expert parsing can be handled and evaluated in the context of contemporary, constraint-oriented transformational grammar. The use of non-linguistic expert systems in connection with parsing (and also with synthesis) for the aims of machine translation is discussed by Boitet and Gerber (1986). Another possibility how to cope with the complexness of syntactic analysis, and also with its procedural character, was outlined by Schnelle (1984). One of the ways how to avoid the use of complex frameworks handling the sentence syntax by means of trees with unnecessarily many nodes can be seen in the use of dependency syntax, applied on parsing of English by Kirschner (1982). A model of a relatively easy parsability of a language is investigated as for its connection with a model of easy learnability by Berwick (1984 b), the result being that certain constraints ensuring efficient parsability (in a sense close to Marcus' deterministic parser) also guarantee easy learnability (for which an explicit criterion is used).

An approach to the algorithmic treatment of the fuzziness of lexical meaning was characterized by Rieger, B. (1984; 1985); as we already remarked in sect. 4 above, also here it seems probable that computerized procedures can only be used in combination with human insight. One of the aims of such an interaction in interpreting lexical as well as syntactic units may be to find out which choice should be made among the spectrum of such layers of analysis as those that are necessary for intensional and other opaque contexts; see Hobbs (ed.) (1985).

Since one of the main trends in contemporary linguistics is a shift of the center of attention from language system to language use, it is not surprising that several attempts have been made to use computers also in the research of discourse structure; computational models of aspects of discourse have been often constructed in close connection with the research in the experimental domain of artificial intelligence (natural language comprehension, knowledge representation, natural language interfaces to expert systems and intelligent robots, automatic compilation of knowledge bases). Scenarios or scripts are used as the basis for the description of typical complex actions, see esp. Schank (1973), Bobrow/Winograd (1977). For such systems to have a general basis, it is necessary to exploit the patterning of natural language, the universal character of which may ensure that not only ad-hoc procedures are formulated, i. e. that when the time comes to adjust a system to a new domain of application, it will not be necessary to rebuild it radically; an attempt at a procedure using inference rules for a transduction of linguistically based underlying representations of sentences into semantic networks is outlined by Hajičová/Sgall (1984), cf. sect. 2.2. above.

The linguistic background of the inquiries into discourse structure and human-machine communication is often studied with a specific attention paid to phenomena concerning the coherence of a text, especially to coreference (anaphora). The communication function of natural language has its impact also in the structure of the sentence, the topic-focus articulation of which is of primary importance for text coherence. This articulation (relevant for presuppositions and for the scope of negation) and the communicative dynamism (semantically relevant in what concerns the scopes of operators and related phenomena) has to be systematically accounted for, since only such a description of the sentence can be considered adequate, that can be directly embedded into an account of the use of language in communication; see Sgall et al. (1973; 1986); Hajičová (1983); Hajičová/Sgall (1985). How this articulation and the coherence of the discourse interact with the degrees of salience (activation, prominence) of the items in the stocks of knowledge of the speaker and of the hearer, was further studied by Hajičová/Vrbová (1981) and, under such headings as 'focus of attention', by Grosz (1981), Grosz/Sidner (1986), McKeown (1985), and others. Topic and comment (focus) are used as basic ingredients of a text analysis aiming at automatic summarization by Hahn, U. (1984 a). An experimental system using some of the fundamental approaches of cognitive science to the aim of knowledge representation and connecting them with a procedural determination of sentence topics on the basis of activation of subparts and/or elements of the cognitive space is characterized by Hofmann (1986).

The possibility to use a tagged corpus for a study in text linguistics was illustrated by Gustafsson (1982). Enkvist's approach to topicalization was applied here in a statistically based inquiry which has shown that e. g. free adjuncts are much easier to topicalize (in the sense of Chomskyan linguistics) than obligatory ones (valency adverbials); another, less clear result is that among the main factors for topicalization there is the textually unmarked, though grammatically marked type (in which a 'given', i. e. contextually bound item is fronted). An example of using a tagged corpus in combination with multivariate techniques for the purpose of a description of variation in language is presented by Biber (1985), who points out how factor analysis, if based on results of empirical research, can help to identify clusters of features cooccurring in individual text types; as for English, the oppositions of interactional vs. informational (edited) and of abstract vs. situated texts were established in this way.

The fact that a discourse belongs to the domain of human (communicative) action is duly reflected in J. F. Allen's (1983) attempt at a procedure modelling the way in which a hearer may identify the speaker's plan and goals, where certain concepts of speech-act theory are used. This concerns also Rothkegel (1986), who attempts at combining text understanding with machine translation. Another aprroach to text organization within a theory of action was presented by Mann (1984 a). Human communicative activity, to which discourse belongs as a specific (though prototypical) case, is determined by much more complex questions than just the knowledge of a language; discourse is influenced by factual knowledge and/or beliefs and other attitudes, aims and psychological motives of all kinds. Therefore, discourse should be viewed as a sequence (or even a more complex collection) of utterance tokens together with their sense (which includes reference assignment to the referring expressions contained in the utterance), rather than as a sequence of sentences (where the sentence is a unit of language systems, of linguistic competence). A formal model of discourse thus belongs more to the domain of simulation of the use of language than to that of description of language system. A detailed characterization of this domain can be found in art. 23—36 of this Handbook.

6. Conclusions

6.1. In this article of the Handbook we are interested in the description of language system, rather than in the simulation of language behavior. The structure of natural language has been influenced by the fact that languages exist and develop in the environment of human communication. The impact of the main function of language on its structure may be seen in the anthropomorphic character of the basic ontological patterns underlying word classes (with individuals or objects, their properties and mutual relations as the primitive units) and of the sentence patterns (with actor — action — objective and bearer — property as the archetypes) as well as in the procedural character of the sentence with its 'dynamic' topic-focus articulation. The latter is derived from the communicative function as being advantageous for the 'given-new' strategy; it enables the sentence to be used as an instruction for the hearer to concentrate on some items which are easily available in her/his memory and to put them into more or less new relationships among themselves or with other (perhaps not so easily available) items. Thus, linguistic descriptions should be so formulated as to account for these properties of the structure of the sentence and its parts.

The approaches to linguistic description formulated in structural as well as in modern theoretical linguistics, before the use of computers has been really wide-spread, will maintain their basic, not only historical, importance. The development of the empirical linguistic research can only proceed in further steps of trial and error. This can hardly be changed by even a massive application of computational techniques, since the computer can only be used as a tool. The main progression of linguistic research always leads from data observation through hypotheses to their testing on some other or larger data sets, amending them or formulating new empirically based hypotheses, i. e. theories, which again should be tested. Observational data nowadays are organized in machine-readable corpora, in lexical and other databases, for the compilation and complementation of which (including tagging and other kinds of classification) computers can be used only within interactional systems; thus, human insight will remain crucial, even if in the future it becomes possible to use computers to a larger extent also for formulating or amending (generalizing, refining) hypotheses.

A linguistic theory will be needed in this sense as a basis for any classification of the phenomena observed (or of a set of deeper units derived from such phenomena). It can be argued that a well motivated linguistic approach, a sophisticated, though preliminary description is a prerequisite for the next stage in the increase of explanatory power. One stage of linguistic theory is a precondition for gaining another, higher level of deducing an underlying linguistic system from language phenomena.

6.2. However, the help of computer for a quick exploitation of the massive of data which today can be made accessible for automatic analysis is extremely effective, if organized on the basis of systematic linguistic research. Therefore the apparent paradox quoted above in sect. 1 need not be surprising: the computer is nothing more than a tool in linguistics, but it is a tool powerful enough to make the next steps in the cycles of formulation and testing of linguistic theories substantially more far-reaching than what we are used to. Even so it is important to keep in mind that each of these — however large — steps can find useful inspiration in the preceding stages of the large-scale development of research (since the regular waves of attention concentrated on another issue in each of the steps may be intensified, rather than lowered by the increasing use of cumputers). Thus, the importance of a non-interrupted continuity in the development of theoretical linguistics does not diminish. Not to be directly, intensively interested in what was done in our science some decades ago, where the methods we have now were not yet available, this means to undergo a risk of falling into one blind alley after another.

6.3. Due to the universal character of natural language, in which every human thought can be encoded with the desired degree of precision (since natural language always can be enriched by the necessary expressions, without changing its basic structure), a systematic description of the language system is also of great use for the practical applications. For every subarea of experiments and applications in computational linguistics and in the linguistically oriented domains of artificial intelligence it holds that to ensure a safe basis for future generelizations, for changes of the kind of texts to be processed, and so on, it is useful to remain as close to the structure of natural language as possible. The irregularities, ambiguities and other intricacies of the outer shape of natural language can be handled economically, if we succeed to identify its underlying structure, which retains the richness of the language without carrying with it all the burdens. As we have seen in sect. 2.1 above, up to now the research in computational linguistics has been rather scattered among small groups, and the non-trivial applications mostly have not reached a larger scale. However, it can be expected that practical needs will require wide-spread processing of natural language texts on the basis of linguistically based systems — for the aims of translation, of knowledge bases compilation and of other kinds of human-machine communication in which the encoding and decoding duties will be transferred more and more completely to the computer. This practically conditioned broad testing of linguistic descriptions, if consciously used by linguists, will be decisive for linguistic theory to gain a new level of adequacy.

Acknowledgement

I am indebted for valuable remarks to the manuscript of this contribution to Robert Berwick, Karen Sparck Jones, and Eva Hajičová.

7. Literature (selected)

J. Aarts/T. van den Heuvel 1985 · J. Allen 1983 · J. Bátori/H. J. Weber 1986 · H. Bergenholtz/B. Schaeder 1977 · R. C. Berwick 1983 · R. C. Berwick/A. S. Weinberg 1984 · C. Boitet/R. Gerber 1986 · C. S. Butler 1985 a · N. Calzolari 1983 · A. R. Golding/H. S. Thompson 1985 · B. J. Grosz 1981 · E. Hajičová/P. Sgall 1984, 1985 · P. Hellwig/H. Lehmann 1986 · K. Heß/J. Brustkern/W. Lenders 1983 · S. Johansson 1982 · A. Joshi 1982 · A. Joshi/B. Webber/R. M. Weischedel 1984 · J. Kunze 1982 · G. Leech/R. Garside/E. Atwell 1983 · M. Marcus 1980 · M. Nagao 1985 · J. P. Pusack/S. E. K. Otto 1984 · B. B. Rieger 1981 · H. Schnelle 1984 · P. Sgall 1984 · S. M. Shieber 1985 · H. Slocum 1985 · B. Vauquois/Ch. Boitet 1985 · W. Wahlster 1981.

Petr Sgall, Prague (Czechoslovakia)