**Lesia Ukrainka Volyn State University**

# Iryna Biskub

# APPLIED AND COMPUTATIONAL LINGUISTICS

**Lutsk - 2006**

УДК 81’33 (075.8)
ББК 81.1 я 73
Б 65

**Рецензенти:**
**А.Ґ. Ґудманян**, директор Гуманітарного інституту, завідувач кафедри англійської філології і перекладу Національного авіаційного університету, професор, доктор філологічних наук;

**П.О. Бех,** начальник управління міжнародних зв’язків, завідувач кафедри методики навчання іноземних мов і прикладної лінгвістики Київського національного університету імені Тараса Шевченка, професор, кандидат філологічних наук;

**М.П. Фабіан,** завідувач кафедри іноземних мов Ужгородського національного університету, професор, доктор філологічних наук;

**М.О. Олікова,** завідувач кафедри прикладної лінгвістики Волинського державного університету імені Лесі Українки, професор, кандидат філологічних наук

Б 65 **Biskub I.P.**
Applied and Computational Linguistics: Підручник (англ. мовою) – Луцьк: РВВ „Вежа” Волин. держ. ун-ту ім. Лесі Українки, 2006. – 335 с.
ISBN

Розглядається сучасний стан прикладної та комп’ютерної лінгвістики, проаналізовано лінгвістичні теорії 20-го – початку 21-го століть під кутом розмежування різних аспектів мови з метою формалізованого опису у електронних лінгвістичних ресурсах. Запропоновано критичний огляд таких актуальних проблем прикладної (комп’ютерної) лінгвістики як укладання комп’ютерних лексиконів та електронних текстових корпусів, автоматична обробка природної мови, автоматичний синтез та розпізнавання мовлення, машинний переклад, створення інтелектуальних роботів, здатних сприймати інформацію природною мовою.

Для студентів та аспірантів гуманітарного профілю, науково-педагогічних працівників вищих навчальних закладів України.

ISBN

УДК 81’33 (075.8)
ББК 81.1 я 73

**CONTENTS**

# PREFACE

Modern theories of language, unlike those of ancient and medieval times, are more concerned with how language works than with why it exists. They therefore tend to base their principles on the observation of language and languages. The theory will therefore depend on what is observed and how it is observed. In each field of knowledge concerned with language, there are different and often contrary ways of observing linguistic facts.

In the field of philosophy, some writers regard language as an external expression of universal thought; others would reduce all differences in philosophy to differences in the use of language. In the field of psychology, theories of language tend to differ according to both the school of psychology and the branch of psychology practiced – social, educational, or child Psychology. For some psychologists, language is a type of symbolism with many functions; for others, it is a man-made instrument of communication.

Linguists, whose special field is the study of language, maintain an even greater divergence of theories. To the linguist, language may be form and not matter; or it may be a system of arbitrary vocal symbols; or it may be a system of systems, a system of hierarchies, or even a hierarchy of systems. To some, it may be material; to others it may be mental. To some it may include only vocal symbols; to others, it may also include written symbols. If there are differences within each field, there are also points of similarity between theories in different fields – the agreement, for example, of certain linguists, psychologists and philosophers on the non-material nature of language, as well as admitting the necessity of the implementation of technology into linguistic research.

Tracing the sociocultural influence of any technology is fraught with problems. First, many of the influences cited are likely to be too large and diffuse to be tested under experimental conditions in the laboratory. Second, the technology is likely to be, at most, an accessory to many other influencing factors rather than a singular cause. Third, insofar as the technology can be isolated as a factor of influence, the direction of the influence is often two way. The technology

may cause changes in sociocultural states, but existing sociocultural states are also likely to result in the technology being used and evolved in unanticipated ways.

The use of computers in linguistic research has lead to the establishment of new disciplines, such as Computational Linguistics, Computational Lexicography, Computer Corpus Linguistics. All of them consider the notion of computational lexicon as their primary concern. The *lexicon* has come to occupy an increasingly central place in a variety of current linguistic theories, and it is equally important to work in natural language processing. The lexicon - the repository of information about words - has often proved to be a bottleneck in the design of large-scale natural language systems, given the tremendous number of words in the English language, coupled with the constant coinage of new words and shifts in the meanings of existing words. For this reason, there has been growing interest recently in building large-scale lexical knowledge bases automatically, or even semi-automatically, taking various on-line resources such as machine readable dictionaries (MRDs) and text corpora as a starting point.

One of the major tasks in building large-scale lexicons for natural-language processing systems is to establish standards for the compilation of machine-readable dictionaries. The existence of large-scale electronic corpora now makes it possible to systematize facts about the usage of linguistic items in real world. In-depth studies have already been made in the sphere of implementing systematicity in the treatment of linguistic data by lexicographers (R.Moon 1987, 1988; B. Boguraev and E.J. Briscoe 1989; B.T.S.Atkins & B.Levin 1991; M. Bates, R.M.Weischedel 1993; V. van Ooi 1997).

Variations in lexicographical decisions, as well as ambiguities in entry design, prevent successful automated lexicon building and extraction by natural-language processing systems. Electronic corpora now make possible a realistic evaluation of dictionary entries by examining the behavior of words in a real language in use.

Corpus evidence is used to support apparently incompatible semantic descriptions. Lexicographers now agree that a distinction should be made between formal semantics (FS) and common-sense semantics (CSS). No clear qualitative

difference has been made between them up to now. Quantitative distinctions between the two types of semantics are based on their idiosyncratic identifications. FS systems offer minimal codings for content words using meta-logical methods for the establishment of properties of whole systems, and the employment of properties such as decidability to establish the validity of universal theorems.

The theory of lexicographical systems has been applied in designing the Ukrainian linguistic network on the Internet, and in establishing virtual lexicographic laboratories. They may function on-line by realizing the possibility of simultaneous work by linguistic groups working in separate places. The project has as its goal the initiating of the so-called “All-Ukrainian Linguistic Workgroup”. The results of the research will be uploaded into on-line catalogue of the Ukrainian language dictionaries.

The on-line version of the integrated lexicographical system "Dictionaries of Ukraine" has been developed on the basis of the CD-versions of the integrated lexicographical systems (V.Shyrokov, 2004). The on-line version consists of 4 dictionary subsystems: Word Inflexion, Synonymy, Antonyms and Phraseology. The overall word list includes approximately 186,000 entries. The word entries located in the modules "Phraseology", "Synonymy" and "Antonyms" have been indexed. Clicking with the mouse on any word-form within the entry brings up the word's initial form and performs the immediate transition to the selected word in the system's word list.

None of the dictionary’s subsets is intended for use by machines at this point. The perspectives of the project development presuppose the creation of machine-readable versions of each on-line dictionary. The generation of such useful systems requires observation of word behavior in interactive lexicographic systems under realistic circumstances.

The problem of the semantic formalizations is especially urgent for modern Ukrainian projects in the sphere of Computational Linguistics, Computational Lexicography and Corpus Linguistics. Systematization and the establishment of standards for the formalization of linguistic information is now the focus of

attention by the Ukrainian Academy of Sciences. The mutual impact of algorithms for formalizing semantics and corpus-based lexical acquisition has led to the creation of the National Corpus of the Ukrainian Language (O. Demska-Kulchytska, 2006) based on the example of the British National Corpus. The project is still being developed.

The first Computational Linguistics Conference was held in Kiev, Ukraine (April 2006). The Conference organizers were the Ukrainian Language Institute (Ukrainian National Academy of Science) together with the Institute of Philology (T.Shevchenko Kiev National University) and the Kiev National Linguistic University.

The following fields were the main focus of the conference: 1) Corpus linguistics; 2) Computational Lexicography and Lexicology; 3) Computer linguistic resources; 4) Information retrieval. Information extraction; 5) Speech pro cessing. Spoken language recognition and understanding; 6) Computerized morphological analysis; 7) Computerized syntactic analysis; 8) Computerized semantic analysis. Semantic Processing; 9) Machine translation. Among the participants of the conference were V.Perebyjnis (computational lexicography, computational and applied linguistics), E.Karpilovska (mathematical and statistical linguistics, computational linguistics), N.Bardina (experimental linguistics, linguistic cryminalistics), O.Demska-Kulchytska (computer corpus linguistics), and other scholars.

The standardized mechanisms of semantics and syntax formalization as well as the collection of corpora of typical data for text analysis and machine translation systems are the main areas of near-term research, directed towards making significant breakthroughs in Ukrainian Computational and Applied Linguistics.

**THANKS**

The author of this book benefited greatly from the discussions with many prominent Ukrainian scholars in different domains of linguistics: prof. S.Shvachko (Sumy State University), O.Demska-Kulchytska (National Academy of Sciences), N.Andrejchuk (Lviv National Polytechnic University), O.Konstantinova (Rivne Institute of Slavonics), prof. M.Poluzhyn (Uzhgorod National University), I.Korda-Andrusiak (Uzhgorod National University), prof. R.Pomirko (I.Franko Lviv National University), O.Shpak (I.Franko Lviv National University).

I am particularly grateful to the first Chair of Applied Linguistics Department M.Kikets (Lesia Ukrainka Volyn State University), for initiating applied linguistics research at my university and for being my teacher.

I wish to express my sincere gratitude to prof. D.Sherik (Lesia Ukrainka Volyn State University) for his support and for having supplied me with useful books on computational linguistics.

Special thanks go to prof. P.Bekh, who supervised my previous work on verbal aspectuality, and kindly agreed to review the present book.

I owe a separate debt of thanks to prof. A.Belova (T.Shevchenko Kiev National University) for valuable criticism and scientific discussions we used to have, and whose advice I appreciate above all.

Finally, I express my gratitude to the reviewers: prof. A Gudmanian (The Hummanitarian Institute of National Aviational University) and prof. M. Fabian (Uzhgorod National University), who many years ago used to be my best university teachers, prof. M.Olikova (Lesia Ukrainka Volyn State University), whose comments and suggestions made at various stages in the development of the manuscript were most helpful.

Iryna Biskub

# CHAPTER I.
# APPLIED THEORIES OF LANGUAGE

## 1. Meta-Linguistic Terminology

The most obvious way in which one language theory differs from another is in the words it uses or invents to talk about language. The differences may be due to the use of (1) different languages, (2) different terms, and (3) different meanings.

Linguistic theories are not all written in the same language. The different languages in which they are written do not all have the same number of words for linguistic concepts, and even in cases where they do, the counterparts do not cover the same area of meaning. English, for example, has only the words *language* and *speech* to do the work of the French *langue*, *langage*, and *parole*. The English word *language* is not always equal to French *langue*; nor is French *langue* equivalent to the German word *Sprache*, no more than *parole* is equal to *Sprechen*. French has *signification* and *sens* to cover the meaning of English *meaning*, *sense* and *signification*. Yet neither set of terms covers the meaning of German *Bedeutung* and *Sinn.*

This state of affairs has led some linguists to speculate on whether existing linguistic theories would have been different had they originally been formulated in a different language. In translating theories from one language to another, it has become the practice to preserve the key words in the original language in which the theory was first expressed.

A second difference is in the terms themselves. Some theorists invent new words for the categories which their particular theory distinguishes; they do so as one way of overcoming the possible confusion and inexactness in the use of everyday words. For the study of speech sounds, for example, they have invented such terms as *phone*, *phoneme* and *allophone* to distinguish between a segment of speech sound, a relevant speech sound, and its variants; and by analogy, *morph*,

*morpheme* and *allomorph* were invented for the study of words. New terms such as these have filled a number of linguistic glossaries.

The different terms used in different theories, however, do not always correspond to new concepts. Linguists have not hesitated to invent new terms for well-known concepts. Thus the study of relevant sounds might be called *phonemics*, *phonematics*, or *phonology*, depending on the school of linguistics in which the term is used. The new terms are created on the grounds that they do not stand for exactly the same concepts as those of the other linguists. Of course, they are unlikely to, since the theories are not the same.

Different schools of linguistics and language theory have turned out entire vocabularies of technical terms. Within each school, however, there are terms which are the property of a single writer; for example, C. Morris's (1946) glossary contains over a hundred terms, nearly all of which are of his own invention.

Most attempts to date at compiling a general dictionary of linguistic terms have given unsatisfactory results. It is not surprising that the Permanent International Committee of Linguists (PICL) have considered it wiser to ask each school to prepare its own glossary, covering a limited span of time. Among such glossaries of linguistics are: J.Knobloch et al, *Sprachwissenschaftliches Wőrterbuch* (1961), E.P.Hamp *A Glossary of American Technical Linguistic Usage* (1950), M.A.Pei, F.Gaynor *A Dictionary of Linguistics* (1954).

The greatest confusion, however, is that created by giving different personal meanings to words in common use. Take for example the words *sign* and *symbol*, key words in many language theories. The word *sign* may mean simply an event which produces a response (K.Britton), or it may be more than a stimulus in that it controls behaviour toward a goal and means the same thing to speaker and hearer (C.Morris), or it may mean an abstract unit consisting of content and its expression (L.Hjelmslev), or it may be a class of events which produces the same reaction as another class of events (B.Russell), or any material thing having prearranged mental equivalents transferable at will (A.H.Gardiner), or a concept bound to an acoustic image (F.de Saussure). Writers in the same tradition may use the same

terms; but this is no guarantee that they carry the same sense. Both F.de Saussure and G.Guillaume, for example, distinguish between *signifiant* (the signifier), *signifié* (the signified), and *signe* (the sign); but the latter uses signifiant with Saussure's meaning of *signe*, and *signe* with part of Saussure's meaning of *signifiant.*

The word *symbol* is another example. For some, linguistic symbols are units of communication; for others they are units of thought. For F.de Saussure, symbols can even be natural phenomena; for J.Lasswell they are interpretations of communication and are opposed to signs, which are physical carriers of symbols from speaker to listener. According to C.K.Ogden, symbols are signs used by man for purposes of communication; they are signs of acts of reference. To B.Langer, a symbol is something which refers to a conventional concept and has meaning only in the mind; to C.Morris, a symbol is a sign produced by an interpreter and acting as a substitute for some other sign (an interpreter is defined as "an organism for which something is a sign"); to G.Maritain a symbol is a sensible thing signifying an object by reason of some presupposed relation of analogy; to R.Naumburg it is an expression, cultural or active, which contains an element of disguise or metaphorical allusion, etc., etc.

The words *symbol* and *sign* are by no means the only instances of the confusion of terms in linguistic theory. An equal number of different definitions could be given for almost any of the key words. J.Ries (1931), for example, has been able to compile a hundred and forty definitions of the term *sentence*

This confusion in terminology has been largely responsible for the isolation of one discipline from another in matters of language theory and for the limitation of most linguists to their own theory—sometimes supplemented by a misinterpretation of a few others.

For this reason, the study of linguistic terminology is one of the main tasks of modern Applied Linguistics; for linguistic terms often conceal significant differences and similarities in what has been said and thought about language, its nature and aspects.

## 2. Applied Theories of Language

Many fields of knowledge have been concerned with language and some have elaborated theories to explain its workings. Since different fields of knowledge are concerned with different things, or study the same thing in different ways, it is not surprising that there is a large number of different answers to the simple question: What is language? To the philosopher, language may be an instrument of thought; to the sociologist – a form of behaviour; to the psychologist – a cloudy window through which he glimpses the workings of the mind; to the logician – it may be a calculus; to the engineer – a series of physical events; to the statistician – a selection by choice and chance; to the linguist – a system of arbitrary signs.

Modern theories of language, unlike those of ancient and medieval times, are more concerned with how language works than with why it exists. They therefore tend to base their principles on the observation of language and languages. The theory will therefore depend on what is observed and how it is observed. In each field of knowledge concerned with language, there are different and often contrary ways of observing linguistic facts.

In the field of philosophy, some writers regard language as an external expression of universal thought; others would reduce all differences in philosophy to differences in the use of language. In the field of psychology, theories of language tend to differ according to both the school of psychology and the branch of psychology practiced – social, educational, or child Psychology. For some psychologists, language is a type of symbolism with many functions; for others, it is a man-made instrument of communication.

Linguists, whose special field is the study of language, maintain an even greater divergence of theories. To the linguist, language may be form and not matter; or it may be a system of arbitrary vocal symbols; or it may be a system of systems, a system of hierarchies, or even a hierarchy of systems. To some, it may be material; to others it may be mental. To some it may include only vocal symbols; to others, it may also include written symbols.

If there are differences within each field, there are also points of similarity between theories in different fields – the agreement, for example, of certain linguists, psychologists and philosophers on the non-material nature of language.

To locate these points of difference and similarity, it is necessary to compare the theories according to their main characteristics. What are the main characteristics of a theory? A theory assumes the validity of certain basic concepts, states the nature of that part of the field of knowledge which it selects as its legitimate concern, and treats it from a certain point of view through the use of certain terms. These four characteristics, therefore, are the main lines on which we can place theories in order to compare them: *(1) the validity of concepts, (2) the nature of language, (3) aspects of language, and (4) terminology.* These are the four ways in which one theory may differ from another. By examining each of them, we can get some idea of the differences between language theories.

*2.1. Linguistics and Philosophy*

What sort of understanding does a theory of language convey? To what branch of knowledge does the study of language belong? What are its central problems? How should knowledge about them be acquired – by experience or reasoning? Should a language theory be based on a distinction between the physical and the mental? These are some of the questions which all theories of language must face. They must also face the possibility of being identified with one or other of the conflicting schools of philosophy.

Some of the best-known philosophers of the twentieth century have based their philosophy on an analysis of language. The work of B.Russell (1927) with the language of mathematics and his view of mathematical knowledge as merely verbal knowledge led eventually to the notion that much of philosophy could be reduced to problems of language. L.Wittgenstein (1960) devoted most of his philosophy to an analysis of everyday language and to a study of the function of words. Others, like E.Cassirer (1923), began to consider language as an independent mental form – scientific thinking as *Another*, religious thinking as still

another mental form. Being thus independent, language could not be understood through the concepts and methods of other sciences. W.M.Urban (1939) used the very existence of language as a proof that metaphysics and ethics could be meaningful; while R.Carnap (1934) rejected these as meaningless since they were not open to logical analysis, which he based on the analysis—or rather, reconstruction—of syntax. For R.Carnap the only proper task of philosophy was logical analysis. Philosophy became logic; logic became syntax.

The basing of philosophy on language analysis is one thing; the basing of language analysis on philosophy is quite another. The preoccupations of the philosopher are not those of the linguist. Each makes a different use of the tools of language and logic. Although both may make use of formal logic, as do R.Carnap in philosophy and L.Hjelmslev (1957) in linguistics, they use it for different purposes: R.Carnap uses it to build up a language; L.Hjelmslev, to break it down. The philosopher is interested in the direct or indirect proof of linguistic statements. Not so the linguist; indeed, many of the statements the linguist is likely to analyse will be logically irrelevant, since they have to do with feelings and images. The linguist is interested in the form and meaning of all possible statements in a language – questions, commands, value judgments – which form the bulk of everyday discourse and have to be analysed as meaningful.

Some linguists claim independence of any philosophical assumption by adopting the pragmatic attitude that only facts verified by the senses are valid and that theories can only be summaries of such facts. But this in itself is a philosophical assumption which shapes the theory.

It is such philosophical assumptions of linguistics, rather than the linguistic assumptions of philosophy, that are relevant to the conceptual foundations of language theory. And these may differ in two fundamental respects – (1) on the concept of man, and (2) on the concept of knowledge.

*Language and the concept of man.*

Since language is a human activity, different ideas on what human activity involves produce different notions on what a language is. Human activity may be regarded (1) as wholly physical (the mechanist view), or (2) as largely mental (the mentalist view).

*The Mechanist View*

This view of man considers the mind as an extension of the body, different only in that the activity of the mind is more difficult to observe. The difference between the mental and the physical, between the animate and the inanimate, is in their complexity. They are essentially the same; the difference is only in degree. All human activity, including language, is a chain of material cause-effect sequences; if one knew the entire history of a person's nervous system one would know what he would say in any given circumstances.

This chain of sequences may be studied from evidence supplied by physical experiments, mostly of the stimulus-response type such as those performed on animals. For the linguistic responses of human beings are in essence considered to be the same as the physical responses of animals to their surroundings. But since so much of the stimulus and so many of the causes, the meanings expressed in speech, happen to be in the mind and therefore unseen, they are understandably neglected in the mechanist theories in favour of the physical manifestations of language in its spoken and written forms. These are the facts of language and are treated as the facts of a natural science.

Language descriptions and language-teaching methods based on such theories tend therefore to present the language mainly as a system of forms rather than as a collection of meanings. One outstanding example of a theory based on this mechanist view of man is that of L.Bloomfield and his school (M.Joos, G.l. Trager, B.Bloch).

*The Mentalist View*

In opposition to the mechanist view, the mentalist view maintains the traditional distinction between mental and physical. Acts of language are mainly mental acts and, although they may very well be correlated with the physical acts of speech, they are acts of a different type. The difference is not only one of degree; it is essentially a difference of kind. Linguistic activity cannot therefore be classed as physical activity.

Nor can human language be studied as animal behaviour. There is a fundamental difference. The animal can be conditioned to respond in a certain way; man, in addition to this, knows the right way to go on, on the basis of what he has been taught. Analogy, an instance of this capacity, is what makes language possible. Much of human behaviour is voluntary behaviour; it is essentially different from the conditioned behaviour of animals. Language, being a human and social phenomenon, cannot therefore be regarded simply as a physical or an animal act. It must be regarded from the point of view of the ideas and feelings peculiar to man.

Language descriptions and language-teaching methods based on a mentalist view are likely to give a great deal of importance to meanings, the mental part of language, and not exclusively to the physical forms. The best-known example of a language theory worked out from a mentalist point of view is that of F. de Saussure and his school.

*Language and the concept of knowledge.*

The validity of a language theory also depends on the type of knowledge it represents—knowledge obtained through the senses, or knowledge acquired through scientific intuition.

A theory may require (1) that languages be described through the observation and classification of facts (the inductive approach), or (2) through the intuition and

construction of a model from which all possible facts may be deduced (the deductive approach).

*The Inductive Approach*

According to this approach, the only valid statements about languages are those arrived at by observing linguistic facts, classifying them and making generalizations on what is observed and classified. It is an imitation of the approach used by the sciences of observation. The linguist is to collect specimens of acts of speech, observe them, and classify the differences. Although he can obviously do this for only a small sample of all acts of speech performed in any one language, he makes generalizations on what he has observed and applies these to the unobserved remainder on the assumption that his sample contains everything of significance.

Since he arrives at his knowledge of language through the observation of its uses, there is no theoretical necessity for him to have any prior working knowledge of the language he describes. It is essentially a matter of gathering samples of the language from a native speaker and then "cracking the code" as it were, through techniques not unlike those of cryptography.

Such theories can therefore produce techniques and procedures of language analysis which are the same for all languages analysed  Any person trained in such procedures is able to make a grammar of any language of which he can get a sufficient number of samples. The approach is based on the belief that only the facts verified by the senses have any scientific validity.

Descriptions of language and language-teaching methods based on this approach are likely to give a great deal of importance to those features of language which lend themselves most readily to physical observation and classification, that is to the phonological features of language – the sounds and sound-patterns. Descriptive procedures such as those of Z.S.Harris (1951), for example, are based on this sort of approach.

*The Deductive Approach*

If the inductive theorist of language imitates the sciences of observation, the deductive theorist follows the theoretical sciences. He perceives a pattern, constructs a theoretical model, and tests to see how much can be deduced from it.

The making of the right model is a matter of scientific intuition. It is done by making explicit the unconscious rules which every speaker of the language possesses; it is the codifying of one's intuitive notions of the structure of the language. One must therefore necessarily know the language before one can codify it in this way. A deductive linguist must first possess the language he wishes to describe.

In any language, the number and variety of utterances are infinite. And since it is impossible to describe all of them, the deductive linguist constructs theories to explain all possible utterances. The best deductive theory is that which gives the simplest explanation for all the known facts and is capable of predicting most of the others.

Descriptions of language and language-teaching methods based on this type of theory are likely to stress the largest patterns of the language—the type which can be arrived at most readily through intuition—the system of the parts of speech and syntactic relationships. An example of a deductive theory is that of G.Guillaume (1963) and the psychomechanic approach to language analysis.

*2.2. Linguistics and Psychology*

Devolving from the concepts of man and knowledge are the concepts of the nature of language. These may range from the conception of language as a sequence of sounds to the conception of language as everything that can be talked about, including the means used to talk about it. Language may be conceived as including not the sequences of sounds themselves but only our idea of them. It may include or exclude the meaning of the sound sequences. If it includes meaning, it may also include the thing meant – or it may exclude it. There is so much overlapping in what different concepts include that, in order to distinguish one

from another, it is necessary to place each within a framework which includes all of them. The framework may be built on distinctions in the field as a whole, starting with the most general – the distinction between substance and form. W.F.Mackey (1983) suggests illustrating this simply by a vertical line:

SUBSTANCE | FORM

In the areas of both form and substance, there is a further distinction between what is being talked about (the content) and the means used to talk about it (expression). This distinction may be indicated by a simple horizontal line, placed over the substance-form areas, since it is included in them:

CONTENT
EXPRESSION

The resulting framework may be shown thus:

| A | B |
|---|---|
| C | D |

where:

AC=Substance

BD=Form

AB=Content

CD = Expression

This gives us four distinct areas:

A: Everything that can be talked about (e.g. things like *doors* and *gates*, and our experience of them) =content-substance.

B: The formalization of these into units of language or linguistic concepts (e.g. English *door+gate*=French *porte*)=content-form.

C: The physical media of language (e.g. sounds)=expression-substance.

D: The formalization of the media into units of expression (e.g. phonemes, letters of the alphabet)=expression-form.

Although this framework is sufficient to locate most theories, a refinement is necessary in order to make it completely functional. Area A includes not only things but also our experience of them, and Area C includes not only sounds but our experience of them. Psychologically it is necessary to distinguish between things, etc. (phenomena), and our conception of them (experience).

An applied psychological language theory may be based on the assumption that language is by nature (1) substance, (2) form, or (3) both form and substance.

*1. Language as substance*

Language may be considered as made up of things that one can see and hear, feel and think. But there is a difference between (1) the substance of things we think about or talk about (Area A—Content), and (2) the substance of what we talk with—or write with—the sounds we utter and the marks we make on paper (Area C—Expression).

Content-Substance relations in modern theories of language are not described through the identification of words with things; but even today, people act as if words and things were identical.

On the other hand, thoughts have long been identified with the words which represent them. The view that the content of words comes from universal mental concepts is responsible for the popular logical grammars which, since the early eighteenth century, have dominated the linguistic education of school-children.

Although twentieth-century linguists and psychologists have either completely overthrown or considerably modified this notion, it still remains the unstated assumption upon which many school grammars and language-teaching methods are based.

Expression-Substance in some theories of language deal exclusively with the material study of what can be seen or heard as language, that is, with the substance (e.g. sounds) used to express language.

Most of these are phonetic theories, which consider the language sounds either from a physical point of view (acoustic phonetics) as do such phoneticians as P.Ladefoged (1962) and L.Kaiser (1957), or from the physiological point of view of how speech sounds are made (articulatory phonetics) as do G.Starka, P.Fouche (1956), and many others. They insist that the sound substance of expression is the very foundation of language. This fundamental approach in now being used in automatic speech synthesis (described in Ch.6).

*2. Language as form.*

The best-known linguistic theories of the first half of the twentieth century consider language not as a substance but as a form. Language is not the same as the thoughts and things about which we speak; nor is it the sounds and tongue movements we use to speak about them.

It may, however, be (1) a labelling or classification of these thoughts and things (content-form), (2) an abstract grouping or image of the sounds and forms of the language (expression-form), or (3) the formalization of both – of what we talk about and how we talk about it (content and expression).

Language has been considered exclusively as the formalization of thought. It is considered not as thought itself, but as a separate symbolic form. This is the view of E.Cassirer, who considers language as an independent mental form, separate from other symbolic forms like mysticism and scientific thought.

Since such theories do not account for the sounds and forms of the language, and show no connection with either substance or expression, they have not been used as a basis for language-teaching methods.

Theories which consider language as the formalization of our means of expression, however, have been applied both to methods of language description

and to language teaching. The best known of these theories is that of L.Bloomfield (1933).

L.Bloomfield begins by excluding both mind and matter from linguistics on the ground that the linguist is not competent to deal with problems of psychology or physiology. Meaning cannot be analysed through linguistics only. The argument is that the totality of meaningful discourse must be “truth" – meaningful and truth being used in-their pragmatic sense. And in dealing with the nature of language, the question of the nature of truth is irrelevant. For this reason, this school of language theory classifies speech by form and not by meaning. And some of its adherents (M.Joos. G.L.Trager, A.A.Hill, C.C.Fries), in order to keep the purity and exactness of their science, have handed over meaning to the anthropologists, phonetics to the physicists, and language learning to the psychologists. But since these "linguistic appendices" are central to none of these disciplines, they have not been incorporated into the main stream of either anthropology, physics, or psychology – disciplines which still look to linguistics to supply the answers to questions concerned with language.

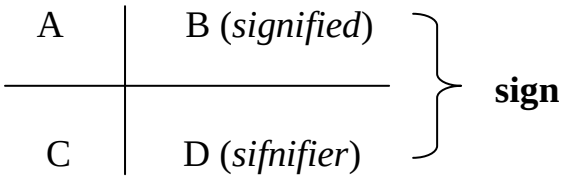
Another theory which limits itself to expression-form is R.Carnap's logical syntax of language (1942). This has been less applicable to language teaching than Bloomfield's theory, however; for while Bloomfield includes only the spoken form of a natural language, R.Carnap includes the written form of any language, but especially that of the artificial languages made up by logicians for the purposes of their study.

Some of the better-known theories of language as form do not limit themselves to expression or to content, but include both. These theories insist on the formal relationship of expressions of language with what they mean.

Most such theories trace their origin to the teachings of F.de Saussure (1915). Saussure's theory first distinguishes language as a code or system (*langue*) from the use made of it in speaking (*parole*). For Saussure, the object of linguistics is the study of the code (*langue*) which is essentially form and not substance. The thought substance and the sound substance do not concern linguistics. Language

(*langue*) comprises neither ideas nor sounds, but simply conceptual and phonic differences. The substance of both the content and the expression of a language is purely arbitrary; so are the connections between the real world, or our idea of it, and the signs used to talk about it.

A sign for Saussure is a fusion of a linguistic concept – the *signified* in Area B – with an acoustic image of the sound – the *signifier* in Area D. This may be illustrated thus:




The *signified* is made of distinctive characteristics isolated by the language from the events of the real world. The *signifier* includes acoustic images of the sounds of the language. Neither includes the physical qualities which such events or speech sounds may possess in themselves. What is relevant is the fact that one sign is not confused with any other. The important feature of a sign is simply in being what the others are not. It is the differences that count. In fact, language is made up entirely of differences. The only positive fact is their combination; it is the only sort of fact that there is in language. Any value which a sign may have lies in its opposition to or contrast with other signs.

Although this is a more comprehensive theory than Bloomfield's, there are many points of similarity. Both L.Bloomfield and F.de Saussure consider language as a form rather than a substance; they both make formal difference the main characteristic of language. For L.Bloomfield, all that is necessary is that each phoneme be unmistakably different from all others.

F.de Saussure maintains that the patterning of the substance of language (Area AC) must be arbitrary. So does L. Bloomfield.

The main difference between these two important theories is in the place given to linguistic content. L.Bloomfield places it outside the realm of linguistics,

claiming that meaning can only be described by the sciences whose object is the content in question. For the Saussurians (R.Godel, C.Bally, C.A.Sechehaye), it is impossible to analyse the expression side as language without implicitly considering its content. Linguistic content is inseparable from linguistic expression; linguistics is the study of their interrelationship.

Among the better-known theories which claim to stem from F.de Saussure is the theory of Glossematics (L.Hjelmslev, H.Spang-Hanssen, E.Richer). The object of Glossematics (1955) is the study of linguistic form; other sciences study the substance. The theory therefore follows Saussure in considering language as form and not substance – as a totality which does not consist of things but of relationships. It is obvious to the Glossematicist that the description of language must begin by stating relations between relevant units, and that the description cannot include information about the substance of these units. The actual sounds (expression-substance: Area C) and the things they stand for (content-substance: Area A) are therefore irrelevant to the language system and may be completely altered without changing the language. But unlike the Bloomfield theory, Glossematics includes the study of form in both areas (expression and content) and stresses their constant relationship.

*3. Language as form and substance.*

While considering language as form, a number of important theories insist that language, by its nature, is also substance. It may be only (1) the substance of language content – the thoughts and things we talk about, or (2) the substance of language expression – the sounds we use to talk about them, or it may be (3) both the substance of content and expression.

Theories concerned with content are interested in how the content of reality becomes formalized as the content of language (the linguistic meanings and patterns). How are the things and ideas about which people talk attached to the units of meaning (content-form) through which the listener understands the speaker? In other words, where do our patterns of meaning originate?

Some theories seek the origin of these patterns in the real world; others find them in the language itself. For some, they are determined by experience; for others, by the particular language used.

### *2.3. The nature of language in the linguistic theories of the 20th century*

At one time or other, philosophers, linguists and psychologists have seen in mind and matter the origins of patterns of meaning believed to exist in all languages. This is the view of B.Russell (1940) and A.H.Gardiner (1932). Although nature and society may be ignored in discussions of language, they ultimately determine all language content. Patterns of language depend ultimately on relations between non-verbal facts derived from nature. Countless acts of speech reflect this relation and result in the shadowy patterns we call meaning.

On the other hand, the patterns of nature may be only partly responsible for the patterns of language. This is the view of K.Britton (1939) and C.W.Morris (1938). K.Britton sees two types of patterns in language, the psychological type, which belongs to the human mind, and the linguistic type which belongs to the particular language. Of the different types of linguistic meaning recognized by C.W.Morris, only identification, that is, the location in time and space depends on the patterns found in nature. This might suggest, however, that the space and time patterns of nature as shown by the physical sciences should be found in all languages.

In opposition to the above is the view that the content of language is entirely independent of our mental or physical experiences of reality. Indeed, the content of language, far from being shaped by thought, is itself the shaper of our mental categories. It is the language content that shapes the mental content. This hypothesis was advanced by E.Sapir (1945) and developed by B.L.Whorf (1956).

E.Sapir saw language as a self-contained, creative symbolic organization which not only refers to experience largely acquired without its help, but actually defines experience for us by reason of its formal completeness and because of our unconscious projection of its implicit expectations into the field of experience.

Elaborating this view into a theory, B.L.Whorf submitted that the structure or grammar of a language is not merely a reproducing instrument for voicing ideas, but rather is itself the shaper of ideas, the programme and instrument of our mental activity, for the analysis of impressions, for the synthesis of our mental stock-in-trade.

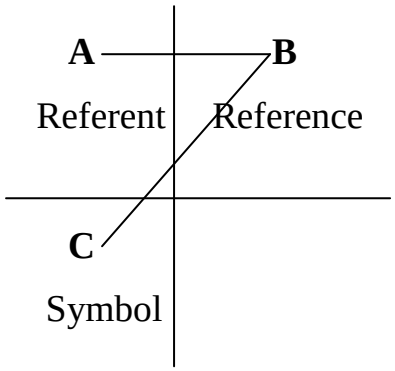
Theories of this type imply that a language is capable of expressing certain things and incapable of expressing others; and translators have not hesitated to supply examples.

Language may be regarded as simply a means of expression, composed of both the substance of expression and its formalization. It is the relationship between the hundreds of sounds we make when speaking and those selected and grouped by the language as being relevant (phonemes as reflected, for example, in the letters of the alphabet).

Among those who consider language thus are the phonologists of the Prague School, who, following the theories of N.S.Trubetzkoy (1949), study the relationship between what they call the speech act and the speech structure of Area D. According to N.S.Trubetzkoy, we can discover the structure of speech by first finding the distinctive or relevant units (phonemes, etc.), by determining exactly what in the speech act keeps one unit separate from the others (phonetic contrasts), and by charting the relation between these. Contrasts such as the voice-vibration in the sound /z/ opposed to the lack of it in /s/, or between /s/ as a continuous sound and /t/ as a non-continuous sound, are all to be found in the actual substance of expression, the physical sounds of speech. The elements of speech structure, however, consist of the way such contrasts are arranged. This arrangement, which varies from language to language, determines the phonological structure of each language.

Theories which include both content and expression do so to different extents. It may be content and expression only as regards substance, only as regards form (BCD in W.F.Mackey's (1983) terminology), or as regards both substance and form.

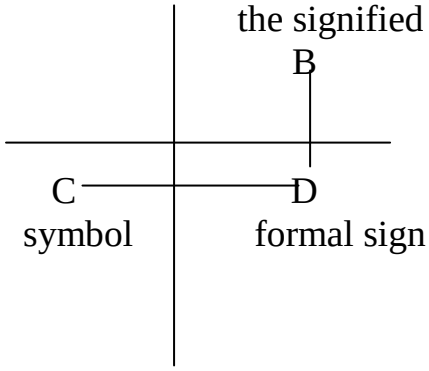
Language may be regarded as an activity in which the thing or idea referred to gets its linguistic meaning by an act of *reference* (B) to a physical symbol (C). Language is thus a continual movement between Areas A and C (the areas of substance) via Area B. It is a movement between the thing or idea referred to (the referent in A) and whatever is used to refer to it (the symbol in C). This is done through an act of reference in Area B which is peculiar to the language in question. This may be illustrated thus (llustration borrowed from W.F.Mackey 1983):




The type of movement between A (things and ideas), B (references) and C (symbols) varies from language to language. In applied language research as well as in language-teaching methods which this theory has produced, a great deal of care is devoted to establishing the right sort of connections (A-B) - (B-C) between the symbols of the language and what they stand for. C.K.Ogden and I.A.Richards' theory of reference (1949) is the original and most widely-known theory of this type.

Language may also be regarded as the expression of content form. All the sounds, words and inflections of a language exist only for the purpose of expressing this. Content form is regarded neither as substance nor as the ready-made representation of substance, but rather as a system of abstract outlines of mental operations whose use enables us to represent certain fractions of our experience. Using W.F.Mackey's terminology (1983), language may be defined primarily as a system of representation (B) which makes use of a system of expression (D); both are form, not substance. But in order that the form be perceived, a language must make use of some physical means, some substance like

sounds or letters (C). The act of speech as expression consists of a continual movement between B and C via Area D, between what is signified (content-form) and the signs used to signify it – the formal signs (expression-form) and the material used to express them (expression-substance). This may be illustrated thus (llustration borrowed from W.F.Mackey 1983):




Such is the psychomechanic theory of G.Guillaume (1963). Any applied language study based on it would presumably start from the most inclusive patterns of content form, the main outlines of representation, graduating through varying shades of abstraction from the most inclusive to the least.

There exist theories which consider language as both the content and expression of form and substance (ABCD), the best known are those of J.R.Firth (1957) and K.L.Pike (1960).

For J.R.Firth, language, being an essentially human activity, must not exclude the mind, thoughts and ideas of those who use it, nor the situations in which it is used (Area A). A language groups and abstracts elements of these situations which have constant relationship with its vocabulary and grammar; this is the "context," a term which followers of J.R.Firth prefer to "content" (Area B). It includes all internal (formal) and external (contextual) relations. There is the relation of one utterance to another, and the relation of utterances to the situations in which they are made. The formal expression of this context of situation is the vocabulary and grammar of the language working through its spoken or written forms, its phonology or graphology (Area D). These in turn are a formalization of the phonic

or graphic substance used by the language (Area C). If we plot this relationship on our framework, it appears as shown on the next page.

Students of J.R.Firth have elaborated this theory and made it more complete and rigorous. M.A.K.Halliday (1959) tries to point out that although the foundations of language are in the context of situation, the theory does not include an analysis of situations as such; it has nothing to do with the study of physical phenomena. It is concerned with this area only to the extent that the forms of the language are related to situations. It is this relationship which is the context, and this context is expressed through the grammar and vocabulary of a language by means of its phonology or graphology, through the actual sounds or script it happens to use.

The applied approach to grammar, no matter what the language, necessarily presupposes that any language is made of certain units, certain structures, certain systems of relationship, and certain grammatical classes. These exist in all languages; but their type and number depend on the particular language. For example, in English the units include the morpheme, the word, the phrase, the clause, and the sentence. These can be arranged in order of increasing size, and each can be defined in terms of the other. Structures are the frameworks into which these units fit. Morphemes fit into word structures, words into a limited number of phrase structures, etc. In a phrase structure of the type *on the table*, each position (like that of the article *the*) can be filled by a limited number of words (*this, that, his, a,* etc.) which operate as a system. Systems operate in a way that imposes limitations on the structure. For example, *the* can be replaced by *these* only if we make *table* plural. Systems group words and word-endings into classes, such as prepositions, conjunctions, etc. The number of items in any one grammatical class is limited, and these items form a closed series.

In the vocabulary of the language, however, the series is not closed. A language is always acquiring new words. The items in any word-class in the vocabulary must form an open series. Words may be arranged in two types of

classes. They may be arranged according to their range of possible combinations with other words, and according to the range of situations in which they are used.

Although it operates likewise in all four areas, the theory of K.L.Pike is quite different from the above. K.L.Pike considers language in relation to a unified theory of the structure of human behaviour. Before elaborating his general theory, K.L.Pike developed special theories for determining the elements in the area of expression (CD) – the "phonetic" elements of C and the "phonemic" elements of D Phonetic elements include all non-relevant variants found in language usage; phonemic elements are limited to the relevant ones. Extending this distinction to the areas of content (AB), and indeed to the entire field of human behaviour, in which language is included, K.L.Pike divides all activity into non-relevant elements in the area of substance (AC) and to the formalized and relevant elements in the area of form (BC).

## 3. Aspects of Language

Within each of the above areas, an applied theory may consider language from three different points of view. It may be interested in (1) how the language sounds or looks (*language as a state*), (2) how it works (*language as an activity*), or (3) how it develops (*language as change*).

### *3.1. Language as a state*

If it is interested in how the language sounds or looks, the theory will include something on language considered as a fixed state—either (1) as a state dependent on what people think and do, or (2) as one which is independent of this.

1. Dependently, language may be considered as human thought or as human behaviour.

*As Thought*

Language may be regarded as being composed of elements of thought. This is the traditional view. In the early twentieth century, F.Brunot (1926) developed this

view into a theory treating of the relation between language and thought. He first divided language into five categories of thought— beings, facts, circumstances, modalities and relationships—and then attempted to show how each of these is expressed by the French language. He made no distinction, however, between structure and vocabulary.

An example of an extensive analysis of a particular language based on mental categories is the French grammar of J.Damourette and E.Pichon (1911-52), which considers language as a system of thought. This voluminous work atomizes the language into a number of mental categories and sub-categories. Many traditional school grammars and language-teaching methods are based on this view of language as a set of mental categories.

*As Behaviour*

Language may also be regarded as being composed of units of behaviour. Not only the anthropologists and sociologists, but also certain linguists, consider language thus. K.L.Pike (1960) places language in the context of human behaviour in general, in which all social acts, including language, are divided into units of significance (*emic units*), each containing a number of non-significant variants (*etic units*).

Another linguist who regarded language as behaviour was G.K.Zipf (1949), who made the state of language dependent upon the principle of least effort.

According to this principle, the forms of the language used in human behaviour become a compromise between the desire of the speaker to get his ideas and feelings across with the minimum of effort and the desire of the hearer to understand them, also with a minimum of effort.

2. Independent

Language may be viewed as an independent state, either as structure, or as a system of communication.

*As Structure*

The idea of language as structure goes back to the teachings of F.de Saussure and his disciples, who regard language as a structure of values between which systematic relations can be observed.

Saussure's basic notion is developed in a number of structural theories. Among these is Glossematics (L.Hjelmslev, H.J.Uldall, 1957). Glossematic theory considers a language as a system of internal relations, as a self-subsistent whole which consists of nothing but relations or functions – "a web of functions. Using the methods of formal logic, it aims at describing the internal structure of a language completely, as simply as possible, and without contradiction. The language is therefore considered as an interplay of purely formal relations.

Other linguistic theories which consider language as structure are those of L.Bloomfield and N.Chomsky. L.Bloomfield (1933) considers the structure as a line or sequence of the smallest units of the language, that is, as a sequence of its phonemes. N.Chomsky (1957), on the other hand, starts by considering linguistic structure itself as a theory which generates all and only grammatical sentences. Considering language as a mechanism for generating sentences, his theory uses a chain of transformations to link the most general structures of language, the sentence patterns, to the sequences of phonemes uttered as sound patterns. The theory attempts to explain how the limited number of structural elements in a language can produce an unlimited number of sentences.

The view of languages as structures is also the basis of *typology* – the study of language types. Typology (R.Jakobson 1958, P.S.Kuznecov 1956, J.H.Greenberg 1954, P.Menzerath 1955, M.Leroy 1961) disregards the traditional family classifications of languages as descended from a common parent (Germanic, Slavonic, Romance languages), since languages of entirely different origin may belong to the same structural type (isomorphism).

Languages may also belong to the same type at one level of structure but differ at other levels. One of the aims of typology is to discover features common to all languages. All languages, for example, seem to have stop consonants like /p,

t, k/; some, however, may have no fricative or continuant consonants. There are some languages lacking syllables with initial vowels, but none lacking initial consonants. Any knowledge of features common to all or most languages has an obvious application to computational decoding of linguistic information, machine translation, language teaching, and other applied linguistic tasks.

*As a system of communication*

Although communication is not language, language can be communication. It can also be much more – and much less.

Language considered from a communications point of view is the transmission of messages; it is the choice of a sequence of symbols from a reservoir of code. Indeed we use language as if we had to choose words one after another. Once a word or sound has been chosen, the choice of the one following is governed by the laws of probability. Thus, if the word is the, the probability of the next word in the sentence being an article or a verb is very small indeed. It is in this way that information theory regards language. "Information" here is a technical term, different in meaning from its usual colloquial sense. This understanding of language has been reflected in the *mathematical theory of communication* (B. McMillan 1953, C.Shannon, W.Weaver 1949).

The language studied in the science of communication is the language of averages. It requires long statistical analyses of languages and specific methods for studying the results. These have been compiled and elaborated in such useful general studies of *quantitative linguistics* as those of G.Herdan (1962) and P.Guiraud (1960).

*Statistical analyses* are of interest to Applied Linguistics in so far as they give information on items which are most often used in a language. They are also of interest to those who base their structural analysis of a language on samples of texts. Since we cannot observe all of the spoken or written sentences used in any one language but only a relatively small sample, linguistic analysis is to that extent necessarily statistical.

*3.2. Language as activity*

In contradistinction to the view of language as a state made of elements of thought or behaviour, or of units of structure or communication, is the view of language as an activity. It is concerned with the way language operates or is operated by man. From this point of view, language may be considered as (1) an activity of the mind, or (2) an activity of the brain.

1.As an activity of the mind, language may be regarded either (a) as mental movement, or (b) as stimulus-response.

*As Mental Movement*

The study of language as mental movement is called *psychomechanics* (G.Guillaume 1963). Its basic postulate is that the mental operations involved in the use of language necessarily take a certain amount of time, infinitesimal though this may be. The task of psychomechanics is to identify these mental operations and to refer them to mental time in an effort to demonstrate the mental process involved in acts of language. The study starts by delimiting the degrees of abstraction through which our minds seize and represent the world of experience. Such an approach is fundamentally different from the traditional analysis of language as a group of static, logical categories.

Psychomechanics attempts to explain how language, the institutional system, becomes usage in the individual acts of speaking and writing. Usage is considered essentially as a process of mental expression by means of acts of abstraction capable of producing such different types of linguistic categories as the parts of speech, inflectional forms, and vocabulary.

*As Stimulus-Response*

Psychologists, and certain linguists as well, have long considered language as a verbal response to external stimulus. This trend is called *behavioristic theory*. Language is regarded as an immediate animal-like reaction to what is perceived (B.F.Skinner 1957). It is as if language were a long series of conditioned reflexes.

A number of linguists have regarded language in this way and have composed applied methods for language teaching from this point of view. Such methods present and drill unanalysed units of language as complete utterances, always given in association with the appropriate situation.

The contextual stimulus-response view differs from this in that it teaches the language as a constant variation in the stimulus to fit a corresponding variation in the language response, thus leading the learner to abstract the patterns of language by seeing the relation between each element of the situation and the corresponding element of the response.

This view of language evolved from the *contextual theory* of C.K.Ogden and I.A. Richards (1949).233 According to this theory, the stimulus of experience comes to us in repeated contexts. These may be physical events which reach our minds through our senses; or they may be events in the mind itself – memories, associations of ideas. Whatever they be, these contexts are continually associated with certain elements (e.g. words) in the language which then become symbols of elements in the context. These symbols, from the speaker's standpoint, are always subordinate to what they stand for; from the hearer's standpoint, they are equal. The hearer first perceives the sound as sound. He then recognizes sounds as distinctive units; he does so because similar sound sensations in the past were always associated with signs. He then recognizes simple referents (e.g. names of things), and finally complex ones. The complexity of the referent, however, is not necessarily reflected in the complexity of the symbol; a single word can stand for a complex idea. But it is not the single word which determines the reference; it is its interconnection with the other words in the sentence.

Language may be considered as only partly a matter of stimulus-response. It is partly concrete activity; partly abstract activity. The concrete perceives and reacts to situations in an animal-like way, through verbal responses to immediately perceived cues and associations, in an automatic type of speech behaviour. The abstract conceptualizes and categorizes. Everyday speech is a combination of both. This is the view of K.Goldstein and R.Buyssens (1948). K.Goldstein regards

abstract and concrete behaviour as only two extreme poles, with fine gradations between them; similarly he makes no clear-cut distinction between conscious and unconscious verbal responses.

*3.3. Language as an activity of the brain*

If some linguists regard language as an operation of the mind, others prefer to consider it as an operation of the brain. They study this operation of the human brain as a physical activity. In order to understand the nature of this activity, two approaches have been developed:

1) The first consists in an analysis of speech reactions during local interferences with the brain (the approach through *neurology*);

2) The second is the construction of models and devices which function as analogues of the human brain (the approach through *technology*).

Since all stimuli leave a trace on the brain and a language sign is an association between two stimuli, the acoustic image and the concept, this sign may theoretically be found in the brain. The localization of speech areas in the brain goes back to 1861, when Broca pointed out the relationship between language and the cortex of man's brain. In recent years, neuro-surgeons have been able to locate the different areas in the cortex which control hearing and speech, memory and thought, and to formulate theories as to their function and interrelation. W.G.Penfield (1959), working in this field for more than a quarter of a century, made verbal tests on hundreds of patients during brain operations. This enabled him to construct a theory based on speech areas and to assume that the organization and co-ordination of the speech mechanisms are carried out by nerve-cell connections, all within the same half of the brain, the dominant one.

Most of those who study the linguistic activity of the human brain no longer believe that it is a matter of mental images. It is rather a matter of nerve impulses traveling along networks. These seem to correspond to the statistical properties of language. Recent neurological theories have led to speculation by theorists of language in a number of different directions. A science of "speculative neurology"

has even arisen. Nerve-cells, which are all-or-nothing firing devices, operate in a two-unit system, building networks in which every linguistic form has its position. Some linguists believe that the nerve-cells are arranged in loops, around which signals circulate and may be remembered by firing one another in succession around the loop and back to the first cell, where the cycle is started anew on its next round (J.Whatmough 1956).

Similarities have also been found between certain *brain disorders* and certain fundamentals of language. R.Jakobson (1956) has compared the basic types of loss of memory with the basic characteristics of language. One type affects the ability to put words together in the right way; the other type affects the ability to substitute one unit for another.

There is still no definite answer to the question whether the dividing of speech into units like phonemes and words is done in the mind of the speaker, or only in the mind of the linguist. Applied linguists set up experiments which aimed to obtain from speakers and listeners certain responses which correspond to the theories and observations of the linguists. But this sort of experiment is more difficult than it seems, for each speaker and listener brings to the language his own special responses which are due to his peculiar nervous system and his own unique combination of memories and experiences in the use of the language; it would often seem that he interprets what he hears according to his own liking. The results of such experiments were described in the works of P.Guiraud (1958), G.Peterson (1960) and B.Manderlbrot (1962).

The second approach to the linguistic operations of the human brain presupposes the construction of models of it, theoretical models and working models.

Theoretical models have been built for the purpose of studying of the activity of the brain. One of these is the *chromatoscope,* a sort of mechanical generator of linguistic hypotheses, in which both words and concepts are regarded as "molecules of experience", particles of "meaning" being the atoms out of which these molecules are built. The atoms of meaning are considered as active "packets

of information" capable of activating other atoms. Theoretical models such as these only suggest possible approaches to the study of the linguistic activities of the human brain (G.P.Meredith 1955).

As for the working models, it is the so-called "electronic brain" that is expected to lead to an understanding of how language operates in the brain of man (V.Belevitch 1956, P.L.Garvin 1963, J. von Newman 1958). These devices, although greater in working capacity and efficiency than the human brain, are extremely limited in the variety of their activities. Their greatest achievement has been in the field of *mathematical computation* and *cybernetics* (N.Wiener 1948).

Efforts to design a machine for the translation of languages have resulted in a good deal of speculation on the linguistic activities of the brain (W.N.Locke, A.d.Boots 1955, A.G.Oettinger 1960). It has seemed likely that the construction of an efficient mechanical translator will contribute to the design of an electronic analogue of the brain. The construction of such an analogue is one of the greatest ambitions of modern science and technology.

Some mathematicians, however, have denied any close analogy between such digital computers and the construction or activity of the human brain. They point out the historical and contingent character of both mathematics and the natural languages. Moreover, it is reasonable to expect that any adequate theory of brain functioning should have statistical characteristics which display plurality, probabilism, variability, redundancy, and tolerance of small errors. No machine having a unitary mechanism, a fixity of properties, an economy of connections, a certainty of output and an intolerance of small errors can successfully simulate the brain (J.von Newmann 1958). Batteries of machines of various types, however, have been suggested as capable of doing so. If and when such an analogue with the proper characteristics is developed, its contributions to linguistic psychology could be profitably correlated with the findings of the neuro-surgeons. If in turn these could be correlated with the analytical and inferential work of linguists, our knowledge of the two extremes of the act of human communication – two minds

communicating through language – may yet reach the exactness of our knowledge of the sound-waves which occur between them.

### *3.4. Language as change*

While some scholars consider language as state or activity, others regard it as something which is continually changing (1) in time, or (2) in space.

*In Time*

A theory may cover variations in language over a period of time, either (a) in the individual, or (b) in the society in which the language has been used.

a) In the Individual

Analysis of change in the speech of the individual is generally confined to the study of the linguistic development of children. This is a field in which important theories have been developed, in the first half of the twentieth century, as a result of studies based on more and more refined techniques of analysis (G.A.Miller 1951).

b) In Society

Today language is generally regarded as an ever-changing code. The changes are not considered as inventions designed by individuals to suit particular purposes; they are systems which arise from the interrelation of the many needs of thousands or millions of people. The mutually modifying practices of hundreds of non-relevant elements in the speech of many individuals eventually bring about changes in the relevant elements which form the code of the language. We are continually altering, continually building the system of our language. It is as if the human mind were dissatisfied with the language it inherits and tries to correct and improve it. Usage seems to display a constant need to be brief, expressive, precise and consistent.

H.Frei (1929) described an advanced state of the French language in which so-called mistakes appear as attempts to simplify the system. It is through these mistakes that the language develops. *Analogy*, which reduces empty forms of no further value, is perfectly normal in the development of language. For example,

frequently heard substandard forms like we was and you was reveal a tendency to regularize the only remaining English verb with an irregular past tense by bringing the plural form into line with the singular, and the second singular with the form of the first and third persons.

There is always room for change, for the vast majority of possible linguistic items and patterns are never used. Of all the possible sounds and forms, only a small fraction is selected by a given language. These are continually varied, combined and re-combined. Any language selects certain features or procedures (like word-endings or word-order), using them more or less consistently and varying the elements in as many combinations as needed. This is how J.Whatmough's theory of selective variation (1956) explains the evolution of language in time. It is through selection and variation that languages evolve. Historical changes continually vary established patterns but only in certain ways. The variation is selective. As patterns are eliminated new ones are chosen to replace them. Each distinctive system evolves in a set pattern, whatever the phonetic or morphological process may have been which first set the pattern.

*In Space*

Variations in space have also given rise to various theories of language. Linguists have studied the variations in space of a single language in the present or of a group of languages traceable to a common ancestor. The first of these disciplines is known as area linguistics, the second as comparative linguistics.

*Area linguistics* has produced theories to explain changes in a language from one part to another of the area in which it is spoken.

Some words are used in all parts of the country in which the language is spoken; others are limited to certain regions. Of the latter, some are limited to one region only, while others cover a number of different regions. The vocabulary of each region differs in both extent and extension.

The differences found from region to region are not limited to vocabulary; they also include pronunciation and grammar. Since many of these differences can be explained neither by the laws of phonetic change nor by the creation of new

forms by analogy, some scholars, like J.Gillieron and E.Edmont (1912), operated on the theory that each word must be treated as if it had a history of its own.

In any area in which a language is spoken, however, we can find different forms of the same word, each representing a different phase of development; some of these are identical with words and forms found in areas in which a different but related language is spoken. In these areas too a word may have a number of different forms shading off into those of still another language area so that there is no clear-cut distinction between adjacent languages like Spanish, Portuguese, Provencal, and Italian, or between German, Dutch, Flemish, Frisian, Plattdeutsch and certain dialects of English. For this reason, the delimitation of languages is arbitrary and, according to some area linguists, purely political.

Theories of area linguistics like that of M.G.Bartoli (1945) have tried to establish principles for arranging these shades of difference and for determining the form from which these arrangements should start. According to M.G.Bartoli the older forms are found in areas which are either isolated (islands and mountains), extensive, marginal (language boundaries), first settled, or areas in which the language is disappearing.

In order to record the difference in words and the shades of differences in forms and pronunciation, samples of language usage have been gathered from all parts of the area in which the "same" language is spoken. These are plotted on maps of the area (often one map per word), and the result is a linguistic atlas. There are linguistic atlases for France (by J.Gillieron and E.Edmont (1912)), Germany (by F.Wrede, W.Mitzka, B.Martin 1953), parts of the United States (H.Kurat 1939), Switzerland, Italy, and other countries. In still other countries, notably in Scotland, England, Ireland, Spain, Canada and areas where Romance, Slavonic and Germanic languages are spoken, scholars have been building extensive dialect archives of usage in the various parts of their respective areas.

The relevance of area linguistics to Applied Linguistics lies in the possibility of deciding what forms to investigate on the basis of proven usage. It helps the researcher to distinguish between the regional and the national. It also enables

them to make use of the regional peculiarities which the native language may have in common with the foreign language.

Along with historical linguistics, *comparative linguistics* profited greatly from the nineteenth-century studies of evolution and from the demanding techniques required to prove the origin and relationship of biological species. Proven relationships were formulated into scientific laws, like Grimm's Law and Verner's Law, some of them admitting of no exception. Genetic theories of the origin and spread of related languages were developed and refined from the comparative studies of F.Bopp, A.Schleicher, K.Brugmann, B.Delbruck, and others. Schleicher's *Stammbaumtheorie*, or pedigree theory, has long given way to Schmidt's *Wellentheorie*, which disclaims the abrupt fusion into language families in favour of a gradual, wave-like spread from the centre. But in order to explain the nature of the dialects on the edge of the area covered by a family of language, this Wellentheorie had later to be modified into *peripheral theory*. Although the theories themselves have little application to modern applied linguistic research, contemporary linguists have made use of comparative linguistics to create formulas for the recognition of words which were common in the parent language.

## 4. Formal Language Description

All modern methods for formal and computational decoding of linguistic information are necessarily based on some sort of analysis, for the very process of the formalization of linguistic data involves the breaking down of the language into the elements which are to be recognized. Automatic language analysis depends ultimately on the recognition of these elements. The more we know about what a particular language contains, the more we can analyse the mechanisms of its automatic processing..

Since the descriptive analysis of a language is the basis for the analysis of language recognition and matching, it is important to determine (1) exactly how

one description of a language may differ from another, and (2) what each type of description contains.

The descriptive analysis of language is of great antiquity. Although the ancient grammars were independent of any universal technique of linguistic description, the extension of Greek and Roman culture throughout the Western world resulted in the application of these classical grammars to the analysis of different languages. This is the origin of the traditional grammars which still form the basis of computational language-decoding methods.

Modern methods of language description differ from the traditional ones; they also differ considerably from one another. This is because they are based on different theories of language or on different techniques of analysis used within the same theoretical framework. These are responsible for the four fundamental differences in the description of a language:

1) in the linguistic levels described;

2) in the units used to describe them;

3) in the direction or order in which these units and levels are treated;

4) in the material on which the description is based.

Knowledge of such levels of description as the vocabulary, grammar and pronunciation of a language is obviously important both for its formal description and successful computational decoding. A method based on a detailed description of the pronunciation of a language will differ from one based mainly on its grammar.

The description of a language may be based only on its grammar, or it may be mainly a treatment of its pronunciation, or of its vocabulary. It may include any of these three levels or all of them. Or it may include more than three, dividing grammar into morphology and syntax, and pronunciation into phonetics and phonology. Descriptions of a language may therefore differ (1) in the number of levels described, and (2) in the contents of each level.

The number of levels into which a language description is divided has varied anywhere from the two of Z.S.Harris (phonology and morphology) to the fourteen

of V.Brøndal (1943). It has been the tradition to recognize three – phonetics, vocabulary, and grammar. Many modern descriptions maintain these three levels; others reduce them to two or increase the number through subdivision or additions. By subdividing the traditional levels and adding new ones a procedure of language analysis may indeed produce more detailed descriptions than it otherwise would. The legitimate scope of interest permitted by a language theory also determines the number of levels in which a description will be made.

The number of levels, however, is no indication of what a description includes. The six levels of J.R.Firth contain just as much as the fourteen of V.Brøndal. Although K.L.Pike, N.Chomsky and S.Ullmann have each three levels, the contents of these are quite different; morphology, for example, which is a separate level in the first case, is combined with phonology in the second, and with semantics in the third. Some linguists restrict their analysis to one area of language, that of linguistic form, analysed exclusively from the point of view of expression and treated in detail by division into such levels as phonematics, phonotactics, morphomatics, morphotactics, inflection and construction.

The relationship between the number of levels can best be illustrated by a comparative table giving the contents of some of the more recent types of description. As the following table shows, by using V.Brøndal's compilation as a basis, and with slight changes in order, we can get some idea of the differences in both the number and the content of levels of language analysis as delimited by a few contemporary linguists. Since the types of analysis are not comparable, however, the horizontal correspondence between levels cannot of course be complete; for what to one linguist is one level may to another have to be distributed throughout several levels. A few of the levels are sometimes considered as inter-levels. Some linguists have used the concept of language level as something that must include other levels; others, like J.R.Firth, have preferred to regard the levels as being interlinked.

Whatever enters into any of the above levels can only be analysed or described through some sort of unit. For the description of vocabulary a unit like a

word is needed; for pronunciation – a unit like the speech sound; for syntax – a sentence unit.

The linguistic units of a language, however, are neither clear nor self-evident. This is because language is a continual flow of sound in which one unit merges into another .Whatever units do exist, they are not perceived as units, any more than one perceives the individual frames of a motion picture when one goes to the cinema. That is why descriptions of the same language differ in the number and type of units used.

Some descriptions use a large number of units; some, only a few. Some have traditional terms for them, terms like sound, word, phrase, sentence; others need special ones like phone, phoneme, morph, morpheme, tagmeme, and seme. But even these special terms, invented to avoid the confusion caused in the many meanings of the popular terms, are themselves used with a number of different meanings. Two descriptions of the phonemes of the same language are not necessarily identical. This is true for most linguistic concepts.

Differences in units and what they mean are determined not only by the linguist's choice of levels, but ultimately by his ideas on the nature of language. Because of these, he may admit units (1) only of expression, (2) only of content, (3) of content and expression.

Language may be described as a system of units of expression. These may be considered as physical units of sound or movements of the speech organs. Or they may be groups of these, formalized into the basic and relevant elements of the language, as when an alphabet represents all the relevant sounds of a language, but only the relevant ones, that is, its phonemes.

Basic units like phonemes may be used to describe all the other units of the language – syllables, affixes, words and word-groups.

In opposition to this are the units based on content, on meaning or reference. They may themselves be units of meaning or content, as for example, the concept of plural in English considered as the same whether it be expressed as *-es* in *foxes* or *-en* in *oxen*; or the classification of all questions as interrogative sentences, no

matter what their forms of expression may be. Or the content may be used to identify the units of expression, as when a native speaker of an unknown language is asked whether two similar-sounding words mean the same thing.

Those who consider language as a form of expression and of content seek their basic units in the relation between these two areas, but without reference to their physical substance. A language is considered as forming its units out of two formless masses – experience and sounds. From each of these, it extracts what is relevant for the content and its expression; and by relating the one to the other, it creates linguistic signs. These are the basic units of a language. Each is composed of a concept (the content) and a sound image (its form of expression).

The levels recognized may be described in different order. In other words, descriptions may proceed in different directions. The direction may be of no theoretical importance and predetermined only by some concrete applied task. One may start with a description of the words, or of the sounds, or of the sentence types, and state any relation observed between one level and any other level. This is the practice of J.R.Firth, M.A.K.Halliday and their collaborators. The same approach is now used in formal language descriptions for automatic language processing by computers.

On the other hand, the theory or technique may require that the description of levels and units follow one and only one direction. This may be (1) upward—from sound to sentence, (2) downward—from sentence to sound, (3) across—from word to word-position to pronunciation.

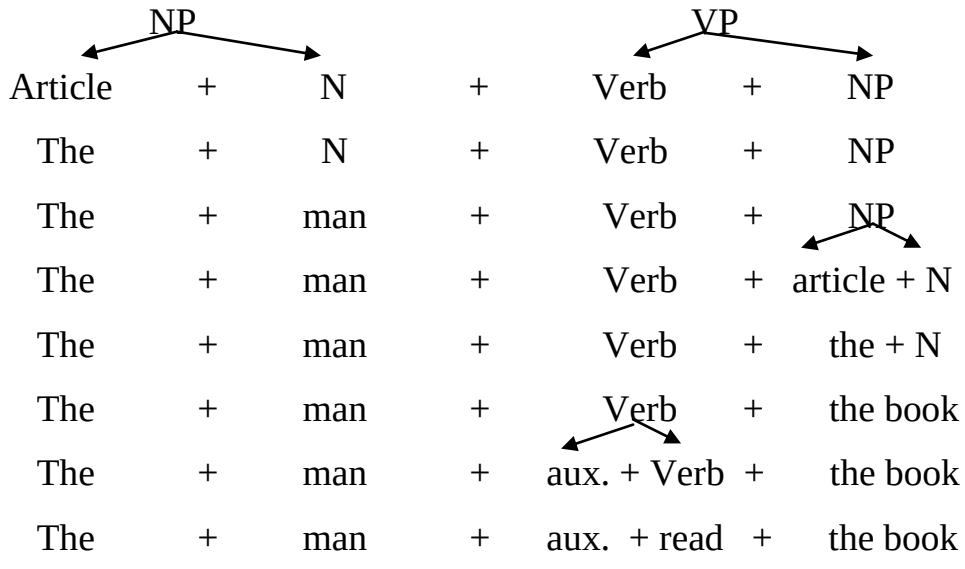
Following the first direction, one starts by establishing the relevant sounds of the language (phonemes); one then proceeds to study how they combine into words, how the words combine into larger units, what the rules are for combining forms and words together, and so on, until the main sentence-types have been determined.

Levels are analysed separately and in the ascending order of complexity, starting with the phonology and ending with the syntax. There is often a strong injunction against anticipating the next higher level, against using material from a

higher level to explain items in a lower level. The phonological description of the language must not only precede the morphological description, it must also be entirely independent of it. Among those who follow this direction are A.A.Hill (1958), G.L.Trager and H.L.Smith (1951), in their descriptions of English. This method is now used in automatic speech synthesis and recognition.

Methods of description using the downward direction start with the largest units and work down to the smallest. The description may begin with a series of texts in the language. These are first broken down into sentences and sentence-types. With the sentence-types as a framework, word-classes (roughly equivalent to nouns, verbs, adjectives and adverbs) and groups of function words (articles, prepositions, etc.) are established. This is the technique used by C.C.Fries (1957) in his description of English structure. The same technique is now applied in modern automatic text parsers representing a top-down approach to the analysis of linguistic items.

The sentence-to-sound direction is all-important for the transformation theory of analysis. Proceeding in this direction, N.Chomsky establishes the basic sentence-types and then moves gradually, through a series of transformations, down to the sequences of sounds. Take, for example, the sentence-type Noun+Verb, or rather Noun-Phrase (NP)+Verb-Phrase (VP). This can become a sequence of sounds through a series of transformations which follow definite rules. For instance:



And a few morphological transformations of the verbal elements convert the sequence into the sentence:

*The man has been reading the book.*

Further morpho-phonological transformations produce the sequence of phonemes. It is in this way that an inventory of sentence-patterns is able to produce an endless text.

If transformation theory produces a text from an inventory, Glossematic theory produces an inventory from a text. Application of Glossematics results in such descriptions as K.Togeby's grammar of French (1951). In this grammar, the French language is first considered as if it were an endless text, to be analysed by dividing it into units and putting these into classes. The text is first divided into two planes – content and expression. Each of these is broken down into parts and described separately. The content is analysed into units according to the ways they combine and the systems they form (e.g. tense, number, case). The expression plane is also analysed into units according to the ways they combine (e.g. according to which sounds occur together) and according to the systems they form (e.g. the vowel system of the language). The larger units are then classified and sub-classified to give a list or inventory of the elements of the language and an outline of the systems which unite them. This appears as a series of tables. The phonological tables of French, for example, show 27 phonemes in all. The same technique was applied in first Machine Translation systems.

The main process of linguistic analysis is therefore the division and classification of the text (language) as a whole and of its expression and content separately. The analysis may be pictured like this:

| *Language* | *Division* | *Classification* |
|---|---|---|
| Expression-Content | Syntagmatic | Systematic |
| Content Plane | Syntactic | Morphological |
| Expression Plane | Prosodic | Phonological |

Distinct from this, is the technique of description which relates each unit of language to all other units at all other levels—phonological, morphological, grammatical. A word is described in terms of its pronunciation, its endings and its place in the different types of sentence. This is the technique advocated by K.L.Pike and his school. The theoretical reason for advocating this technique of analysis is the view of language as a system of three hierarchies – lexicon, phonology and grammar. They are hierarchies because, within each, there is a number of levels, each more inclusive than the other; phonemes are included in syllables, syllables in stress-groups, stress-groups in pause-groups; morphemes are included in morpheme-clusters, morpheme-clusters in words, words in phrases; similarly with grammatical units like tagmemes (gramemes) and utteremes. We may picture the analysis thus:

| *Lexicon* | *Phonology* | *Grammar* |
|---|---|---|
| morpheme | phoneme | tagmeme |
| cluster | syllable | (grameme) |
| word | stress-group | syntagmeme |
| phrase | pause-group | (utterance) |

The difference between this and the Glossematic technique is that it analyses language both as a system of hierarchies and as a hierarchy of systems.

Descriptions of language may differ in the material on which they are based. A description of General American English differs from one of Southern British. An analysis based on the spoken language is not likely to be the same as one based only on written materials. Materials of language description may vary in four respects: (1) in dialect, (2) in register, (3) in style, and (4) in media.

The area from which the material comes makes a difference in the description which results. The American varieties of spoken English, Spanish, Portuguese and French differ from the European. A. description of the pronunciation of Canadian English will not be the same as that of Australian English. In England the speech of the North differs from that of the South. And in the United States, the Southern

accents differ from those of General American. The dialects included in the description may vary according to (1) the size of the area covered, and (2) the size of the sample used.

The analysis of a language may cover one or a number of the areas in which it is spoken. It may attempt to cover all areas or limit itself to one. In a single area it may include the speech of a large number of persons, as does either the French Grammar of J.Damourette and E.Pichon (1952), which includes some 850 speakers analysed, or C.C.Fries' (1957) description of English based on the speech of 380 persons. Or it may limit itself to the speech of an individual. The reason for this latter type of coverage is that it is likely to yield more complete, more accurate and more consistent results, since the language of each individual is regarded as a self-contained system. The technique has been clearly elaborated by H.Frei (1953). It has been applied in F.Kahn's (1954) description of French and Alemanic tenses.

Descriptions also vary in the amount of material on winch they are based. For H.Frei, 2,000 sentences were sufficient; for C.C.Fries, fifty hours of telephone conversation were necessary. The size of the sample may vary from the relatively small amounts of English analysed by G.L.Trager and H.L.Smith to the masses of material used by O.Jespersen and H.Poutsma. The samples may also vary in the range of time covered. H.Frei covers a few weeks or months; A.A.Hill, a few years; O.Jespersen, a few centuries.

There are descriptions, however, which are based on no samples at all, but rather on the judgment of the author. This is sometimes a sample of what he himself would say, a sample of his own speech; but in most cases it is simply based on his rationalizations about the language – not so much on what he says, as on what he thinks others should say. This sort of thing soon ceases to be a description of what the language is, and becomes a prescription of what it should be.

Prescription is not to be confused with legitimate attempts to do away with the use of samples altogether, to eliminate from linguistic analysis the quantitative approach whereby samples of the language are divided into units and categories to

be classified according to their relationships. What is proposed is a qualitative approach based on a theory of how a particular language works. This approach has been compared to that part of chemical theory concerned with the discovery of structurally possible compounds. It is a theory which can generate all physically possible chemical compounds. In the same way, a grammar should be able to generate all grammatically possible utterances. This is the approach to linguistic analysis, advanced by N.Chomsky, to replace the gathering and breaking-down of samples of language.

Register is a term employed by some linguists to indicate the uses to which a language is put—occupational, emotive, informative. A description based on samples of one register may be quite different from one based on samples of another. Hundreds of pages of scientific writing might be analysed without revealing a single instance of the first person pronoun; one page of a private letter might reveal several. A description of the occupational vocabulary of farmers will differ from that of fishermen or factory workers.

The style of the material analysed is likely to be reflected in the description of it, especially in languages where social distinctions are heavily marked. A description of the highly literate speech of a secondary school teacher and that of the language used by an illiterate, unskilled labourer would hardly be the same.

Whether the material was collected through the medium of speech or the medium of the written language also makes a difference. A description of French based on its written form would put the French adjectives *fier* and *premier* into the same *-ier* category; but if the description were based on speech they would appear in two different categories, for although they are written alike they are pronounced differently. Many of the older descriptions like those of English by O.Jespersen, H.Poutsma and E.Kruisinga, are based on the written language. Some of the more recent ones, like those of C.C.Fries, G.L.Trager and H.L.Smith, are based exclusively on speech.

If techniques of description can differ in materials, direction, units, and levels of analysis, so can any part of the actual description, its phonetics, grammar, vocabulary, or meaning.

## Conclusion

The situation in language theory is not entirely unlike that prevailing in other fields of knowledge. In the beginning of the 20th century W.James remarked that so many rival formulations are proposed in all branches of science that no single theory is absolutely a transcript of reality. Moreover, as sciences develop, it becomes evident that most of their laws are only approximations.

The great contrast in twentieth-century linguistic theory is between those who try to relate everything together and those who do one thing at a time. This is less a matter of doctrine than of method.

The consistent application of any one of these theories, however, has far-reaching practical consequences, not only for the linguistic descriptions on which the actual texts are based, but for the development of such applied linguistics-based disciplines as Artificial Intelligence, Machine Translation, Automatic Information Retrieval Systems, and Automatic Speech Synthesis and Recognition. These aspects of Applied Linguistics will be described in the next chapters.

# CHAPTER II.
# LANGUAGE AND TECHNOLOGY

## 1. Language and Literacy

*1.1. Writing*

With the ascendance of new information technologies, the significance of writing has, it seems, slipped from view, in spite of the fact that the conceptual and cognitive implications of the newer technologies is a matter of enthusiastic speculation rather that serious research. On the other hand, it is now reasonably well established that the invention of the first "information" technology, namely writing, has had a profound effect on the ways in which we think about language, the mind, and the world, effects which have taken millennia to unfold. "Effects" is perhaps too strong a term as it is less a matter of how technology affects people than a matter of the ways in which people in different cultures have used and applied the technology and the ways they have altered the technology to suit their purposes. In the West, some of these uses have involved institutional change; thus, to make use of a technology such as writing requires the development of monasteries, schools, and other institutions. Indeed, some of the cognitive effects we usually attribute to schooling are better thought of as consequences or implications of literacy.

The commonsensical view of writing is that writing is the transcription, i.e. the putting down, of speech. This is a view traditionally attributed to Aristotle's *De Interpretatione* but seconded in this century by F.de Saussure (1916/1983) and L.Bloomfield, who saw writing as "a way of recording language" (L.Bloomfield 1933). However, some more recent evidence suggests that writing, in fact, was invented (and is learned by children) primarily as a means of preserving information—the relation to speech was secondary and incidental. R.Harris in his book “The origin of writing” (1986) was the first to point out that the traditional transcription theory assumes what it needs to explain; that is, how did the writers come to know about the properties of their speech? They were, after all, talkers,

not grammarians. Thus it leaves unanswered how knowledge of such properties of speech as phonemes, words, and their literal meanings arise. These sorts of knowledge about language cannot be assumed to exist beforehand and to be available for transcription.

The *linguistically-based writing systems* in the West appear to have evolved from the token system developed for accounting purposes in Mesopotamia beginning in the ninth millennium BC. The system – developed by the ancient Sumerians living in what is now southern Iraq about the time that hunter-gatherer societies were giving way to an agricultural way of life – consisted of a set of clay tokens of distinctive shape and marking used to keep records of sheep, cattle, and other animals and goods of various kinds such as oil and grain. Three vats of oil, for example, would be represented by three vat-shaped tokens on a string. It is important to note that the tokens, like the pictures discussed above, represent things not words or sentences. They convey information; they do not represent or transcribe speech.

About the fourth millennium BC, roughly at the time of the growth of cities, the variety of tokens increased greatly, presumably because of the increasing number of types of items to be inventoried, and shortly thereafter the tokens were impressed into soft clay tablets to provide a record of commercial transactions. Yet there was no attempt to record verbal statements; rather, the tablets filled administrative needs (M.T.Larsen 1989).

The shift from the representation of things, commodities, and quantities into the representation of speech began around this time. The crucial step, R.Harris argued, was the shift from what he called token-iterative to emblem-slotting systems. A system which represents three sheep by three symbols for three sheep (i.e., sheep, sheep, sheep) is categorically different from one which represents the same three sheep by two tokens, one representing sheep, the other the number three. These two signs are now related syntactically. Just as syntax is what makes a language a language, it is the syntax which makes a graphic system "generative," for it permits the combination and recombination of symbols to express a broader

range of meanings. And it is the syntax that turns tokens into word signs. The objects denoted are no longer things but words in a language. Thus, this elementary script has a syntax and could be taken as a model for – a way of representing – the lexical and syntactical properties of a reading of that tablet.

On this view, writing did not presuppose an explicit knowledge of words and syntax. Economies of writing led to the scriptal changes which in turn, when read, could be seen as a tokening of the lexical and syntactic properties of that reading. Concepts of words are the end product of the evolution of the writing system, not preconditions for its development. Writing historically brought words into consciousness.

A parallel story can be told about *children's discovery of words*. It is not obvious to children that writing represents words; they are more aware of the meanings – the story being told. E.Ferreiro and A.Teberosky (1982) first explored this understanding, and some of our own observations confirm that children at first do not understand that printed words represent pieces of speech (that is, language) rather than meanings directly. Consider the following experiment: a child is shown a card on which is written "three little pigs." The text is then slowly read to the child as the adult points at the sequence of words in the text. The child is then asked what it says. The usual response is "three little pigs." Finally, the final word is covered up, and the child is asked, "Now what does it say"? A common, though far from universal response is, "Two little pigs." Each symbol is taken as a symbol of an object and not as a symbol of a constituent of speech. Again, one can infer that writing is important in bringing speech into consciousness.

In reading an alphabetic language like English, the child must be able to segment the words he knows into the phonemic elements that the alphabetic shapes represent. In order to do this, he needs to be consciously aware of the segmentation of the language into units of phonemic size. His competence in speech production and speech perception is of no direct use to him here.

Evidence that pre-readers lack this segmentation knowledge comes from tasks such as that of counting the "sounds" in a word or adding or deleting phonemes

from words: deleting the /f/ from "*fish*" to yield /ish/, or the /h/ from "*hat*" to yield /at/, or adding /s/ to /pit/ to yield "*spit*," for example. Readers find such tasks relatively trivial whereas pre-readers find them extremely difficult. It now seems clear that this difficulty is not merely developmental. The general finding is that people familiar with an alphabet, regardless of age, hear their oral words as composed of the sounds represented by the letters of the alphabet; those not familiar are much less likely to do so. J. Morais, L.Cary, J.Alegria and P.Bertelson (1979) found that adult Portuguese fishermen living in a remote area who received even minimal reading instruction were able to carry out such segmentation tasks while those who had never learned to read could not. Similar findings have been reported for Brazilian non-literate adults, and for both child and adult non-literates in India. Experiments have shown that Chinese readers of traditional character scripts could not detect phonemic segments whereas those who could read Pinyin, an alphabetic script representing the same language, could do so. Japanese first graders, learning to read a syllabary, were less able to manipulate phonemes than were American children learning to read an alphabet. Such findings underline the effects which knowledge of a script can have on one's knowledge of speech. But that is not to say that such awareness is merely a byproduct of learning to read. Rather, to learn to read any language-based script is, at base, to grasp that writing represents speech, and second, to detect those aspects of speech that can map onto or be represented by elements of that script. To learn to read is to learn to apprehend one's own speech in a new way.

Knowledge of the segmental structure of speech arises from the attempts of speakers to map the sounds associated with alphabetic characters onto the sound patterns of their own speech. One can hear this every day as children attempt to spell new words – they slowly articulate speech attempting to hear sounds corresponding to names of particular letters of the alphabet. Eventually, they come to examine their speech in terms of how it is or would be written, thinking that "*cat*" has three sounds because the written form has three letters or that "*pitch*" has

a /t/ sound in it while "*rich*" does not, even though phonologists assure us they are pronounced the same.

### *1.2. Literacy*

*Literacy*, the ability to create and use the written documents of a culture, allows the accumulation and distribution of information far in excess of the traditional means of collecting, storing, and distributing information. Traditional cultures, cultures without a documentary tradition, accumulate information in cultural practices such as farming or cooking, and distribute this collected knowledge through participation, demonstration, and talk. Writing, although developed and exploited in widely different ways in different cultures, has, in the West, increased by an order of magnitude the amount of information available and distributed to increasingly large audiences. But writing – creating and storing documents – not only increases the amount of information available, it also gives rise to a level of representation quite different from that acquired through talk and action. All cultures provide means for children to acquire knowledge of the world, of persons, and of the culture through active exploration and social discourse. Literate cultures, of course, do the same. But they introduce an additional level of representation—those representations constructed and accumulated archivally that make "a paper world."

The "*world on paper*" is an appropriate metaphor for analyzing the conceptual implications of literacy, for by creating texts which serve as representations of the world, one comes to deal not with the world but with the world as depicted or described. Indeed, there may be no world to represent at all, just the set of representations, as in the case of pure mathematics. That such representations can be used to stand for an aspect of the world is just a byproduct of the mathematical tradition. That we deal with representations of reality, rather than reality directly, is rather obvious from maps and formulae for which linguistic descriptions seem inadequate. But it is less readily seen for written texts which appear simply to

transcribe speech. Yet the same relations hold there as well; written texts "represent" the world quite differently than does speech.

The notion of a paper world was not accepted enthusiastically even by those who most directly contributed to its creation. A common refrain among Renaissance writers, Galileo included, was the importance of turning away from books to study the things in themselves. It was the accumulation of information in books, maps, and diagrams that made possible the rapid growth of knowledge that we associate with the Early Modern (that is, 17th century) science.

*Printing* played an important role in the establishment of an accumulative archival tradition. This accumulative archival tradition, storing knowledge produced by many minds in a common representational format, however, was preceded by a new understanding of texts and a new way of reading and writing them, namely, of seeing texts as representations.

The term “representation” requires some elaboration. As long as knowledge was thought of as in the mind and expressed in speech, the usefulness of writing was limited; writing could only be seen as reminder, not representation. To create representations is not merely to record speeches or to construct mnemonics; it is to construct visible artifacts with a degree of autonomy from their author and with special properties for controlling how they will be interpreted: it is to give texts a kind of autonomy from their creators. This transformation from mnemonics to representations began in the 12th century but became conspicuous and dominant in the 17th century. At first, written documents were augmented by iconic signs. Texts made the shift from mnemonics to representations, from memory to written record.

For written texts to bear the burden of official documents – that is, to carry the weight of meanings they came to represent – also required the formation of a new attitude to signs. By the 17th century, language and other sign systems had come to be seen in a new way. Signs were no longer seen as natural to their object but as conventions: not copies or mimesis, but as representations in a medium.

In the Middle Ages, texts were seen as a boundless resource from which one could take an inexhaustible supply of meanings. The assumption thrived in part because the texts in question were religious, primarily the Scripture. Other writing was pagan and consequently of little significance. The arts of rhetoric, various devices for inventive readings for devotion and edification, provided means for rich reading – readings of such diversity and marginal validity as to raise general alarm in the 12th century.

A few other characteristics of such reading should be mentioned. Reading was not distinguished from memorizing, the committing of important texts to memory for subsequent pondering and reflection. Reading was more of an ascetic activity than an intellectual one. Hence, the best preparation for imbibing or drinking in the sacred word was self-purification and mortification rather than theological study. The purpose of reading was to find a truth, a revelation, the light behind the words, the spirit behind the letter. The written texts played a somewhat secondary role to oral speech and oral memory in this period. The written text was not itself thought about; rather the written text was used to permit memorization and to check memory. Difficult as it is for us to imagine, it appears that most thinkers in the Middle Ages and in the Medieval period did their thinking and composing orally and publicly as speeches and sermons, rather than in writing. Students could record what was said but composition was oral. St. Thomas Aquinas, for example, composed his magisterial “Summa Theologica” orally, pacing around a large room dictating to a bank of secretaries, each of whom took responsibility for transcribing a part.

It was this way of reading and writing that gave way and set the stage for the early modern period. Perhaps alarmed by the richness and diversity of interpretation, church fathers beginning with Hugh of St. Victor in the 12th century, and his student Andrew, began to search for ways to weed out the more luxuriant interpretations which were being generated. A new way of reading was in the making.

This new form of reading was at the basis of the Reformation, and it is immediately recognizable to ordinary readers. It required close analysis of the verbal form and its context as well as an analysis of the author, his intended audience, and his choice of expression. It was this kind of reading which M.Luther came to see as providing the one true meaning of scripture, its historical meaning, open for all to see. Reading a text according to its literal meaning, the meaning "grounded openly in the text," was sufficiently radical that reading the book of scripture yielded new heresies, one species of which succeeded as the Reformation; and reading the book of Nature according to the same principles produced Early Modern science.

*1.3. Print*

Tracing the sociocultural influence of any technology is fraught with problems. First, many of the influences cited are likely to be too large and diffuse to be tested under experimental conditions in the laboratory. Second, the technology is likely to be, at most, an accessory to many other influencing factors rather than a singular cause. Third, insofar as the technology can be isolated as a factor of influence, the direction of the influence is often two-way. The technology may cause changes in sociocultural states, but existing sociocultural states are also likely to result in the technology being used and evolved in unanticipated ways.

Various historians of media (E.Eisenstein 1979, W.Ong 1958) have used historical methods to begin to answer what print did to, and for, sociocultural processes. The question is too large for the laboratory and is more often addressed through historical methods. This is because there is a significant gap between the evolution of print technology from an engineering perspective on the one hand (that is, the history of mechanical print) and the metaphysical questions about consciousness and consciousness-raising that print, since J.Gutenberg, was supposed to have spawned. Cultural historians associate print with the evolution of spatial representations of knowledge, with a greater attentiveness to textual accuracy and fixity.

The inferences about consciousness are drawn from historical records but they also incorporate theoretical assumptions that can not themselves be justified by these records. More specifically, the inferences are underlain by an assumption which is called *technological determinism*. Technological determinism is the belief that the human cognitive architecture is reliably affected by the external technologies that are used to augment and extend cognitive processes. Applied to print, technological determinism means that print had a signature impact on mental processes and not just the knowledge held by individuals. The theorist advocating technological determinism envisions the history of literacy as a history of evolving communication technologies (J.Bolter 1991).

The weakness with technological determinism is that it places technology – in this case print – at the center of sociocultural change without taking into account the many other processes of sociocultural change (e.g., population size, migration, language, cultural specialization, differentiation, etc.) in relation to which technology is a factor. Technological determinism also tends to assume that the causal direction between technology and sociocultural processes is one-way, when the relationship seems more realistically to be one of mutual influence.

Any tracing of print's influence must use the historical record, yet it must also make assumptions that elaborate the historical record. Print changes the physical character of face-to-face communication. Some of these assumptions have been elaborated by theorists of written literacy (D.Brandt 1990, M.Nystrand 1989) who note that texts do not strip communication from context; rather, they require one to understand the meaning of communicating in contexts that lack proximity. Like writing, print creates an externalized agent called a text. The word "agent" is used in the sense of artificial intelligence, as an encapsulation of an author's cognition. The fact that the agent is "externalized" means that it is disembodied from the speaker and can engage in interactions in the speaker's absence, even after the speaker's death. The fact that the agent is the product of mechanical print makes it distinct from writing in the multiplicity of its interactions. Because of the

numerous copies made possible by print, print interaction allows one-to-many communication; it allows many readers to interact with the text at the same time.

### *1.4. The Constructural Theory*

Reading the historical record from nothing more than this physical framework for print, constructural theory and simulation methods are used to develop hypotheses about print's potential influence on sociocultural processes.

*Constructural theory* (J.Turner 1988) presupposes that there are three phases to interaction: motivation, action, and adaptation. *Motivation* involves how agents decide with whom to interact. *Action* involves the details of the interaction itself. *Adaptation* involves the longer term structural consequences of interaction. Constructural theory embeds all three of Turner's phases within a dynamic theory. All individuals are involved in a continuous interaction cycle in which individuals become motivated to interact, take action (communicate), and adapt in response to the consequences of this action.

Constructural theory is a process-based theory that relates, in mathematically explicit ways, all three phases. It is a theory designed to show how the cumulation of ongoing concurrent interactions at any time period can impact the society over time. According to the theory, the concurrent interactions of individuals at any time period produce aggregate patterns of cognitive change in individuals, patterns of change that themselves depend on characteristics of both the larger environment (or context) in which the individuals interact, and the technologies through which they interact. Constructural theory, in sum, relates aggregate sociocultural change and the concurrent interactions of arbitrary individuals. In this sense, it offers a specific theory for making the link between micro-action and macro-structure that has long occupied theoretical sociologists.

Within the assumptions of the theory, whether individuals communicate depends on the communicative distance between them. The greater the communicative distance between individuals, the less likely they will interact.

Communicative distance depends on the relative availability of other potential communication partners and their relative similarity. Relative similarity refers to the extent to which the two communication partners share information with each other relative to what they share with everyone else. In other words, two communication partners are more likely to interact with each other if, regardless of how much information they share, they share more with each other than they do with others in the society. Availability and relative similarity determine whom or what the individual actually chooses as an interaction partner.

Individuals communicate using various communication technologies, such as face-to-face and print media. These technologies vary in their synchronicity, fixity, durability, and multiplicity, all of which affect the communicative distance between possible communication partners. Face-to-face interaction tends to be synchronous, low in fixity (oral messages are produced and quickly distorted or forgotten), and low in multiplicity (accommodating few receivers in relation to the population. Print interaction, on the other hand, is asynchronous (reception lags after transmission), high in fixity (the message endures even after the life of the author), and multiplicity (accommodating many receivers in relation to the population).

Printed texts in general can be thought of as the author encapsulated, an extract of the author's knowledge at a point in time that is unalterable. From an information processing perspective, our approach develops the notion that print creates artificial agents – texts – with knowledge, a set of communicative properties, and a set of positions in a sociocultural landscape that is distinct from human agents. For example, unlike people, texts have knowledge that is bounded because they cannot learn. Unlike people, texts can impart information but they cannot acquire it. Unlike people, texts (assuming unlimited copying and feasible costs) are universally open for interaction.

As a result of the concurrent interactions taking place within a sociocultural landscape, some involving only human agents, some including print agents as well, the society adapts, leading not only to new patterns of knowledge throughout the

society but to new sociocultural positions (and hence roles). Through concurrent and recurrent transactions, changes across individuals collectively construct social and cultural changes. In response to the reciprocity between interaction and cognition at the individual level, social structure and culture co-evolve at the societal level. Technological, social, and cultural variations across agents, human and artificial, affect the rate and nature of sociocultural-historical evolution.

The constructural theory is realized as a *simulation system*. Using this system, the logic of the theory for societies with different sociocultural-historical-technological landscapes can be played out. This allows the researcher to engage in a series of historical *gedanken* experiments. This system is used to explore how print may have affected modern society. These simulations both enforce and stimulate a logical framework for thinking through some of the key issues surrounding print. This framework establishes some firm logical relationships between print and other sociocultural-historical variables, and discriminates better and worse explanations about print current in the extant literature. The simulations also rule out possible explanations of sociocultural-historical change, and they generate a series of propositions about the impact of print that are capable of being tested using other scientific methods.

This method has been used to examine the general impact of print as well as the impact of print on the professions, on academe, and on intellectual migrancy. The measures for impact that are used are stability, consensus, and diffusion. Each of these has formal definitions but for present purposes informal ones will serve.

Stability is the degree to which the sociocultural landscape cannot shift. It can be measured as the fraction of available information shared by any two individuals, as averaged across all pairs of individuals in a society. The intuition here is that the more stable the society, the less new patterns of knowledge can form and so the less the society can change as a result of interaction.

Consensus is the degree to which individuals share some belief about a focal concept or decision point. It is a more sensitive measure of shared knowledge than stability because it depends on common patterns of shared and unshared

information across individuals, rather than only on the absolute percentage of shared information. Two individuals who are stable relative to some piece of knowledge can still fail to agree on decision because of other knowledge they do not share, which leads to different decisions.

Diffusion is the fraction of the population that has received some percentage of a communication at some time period. Over-time measures of diffusion involve the amount of time that elapses before some percentage of the population knows some percentage of a communication.

The work of the simulations in this research has been to determine the relative effect of print (compared to face-to-face) interaction on 'societal impact measures' such as stability, consensus, and diffusion within a society. The purpose of this simulation work is not to test empirical hypotheses, but to develop such hypotheses from a mechanistic framework without resorting to metaphysics. In these simulations, it is possible to vary the sociocultural-historical landscape, for example, by altering the size of the population, the complexity of the culture, and the degree of cultural integration.

Print can, however, speed diffusion, stability, and consensus. However, the ability of print to speed diffusion and consensus depends on the extent to which the texts contain knowledge and norms already familiar to the readers. The completely novel text has little impact on society. These results suggest that for any technology supporting communication at a distance to be more than a "novelty" medium, it must accommodate assimilated information and facilitate the community-building aspects of language. Further, we find that the advantage of print over face-to-face communication in affecting the rate at which information is shared is greatest in large societies with complex and highly integrated cultures. Print, however, is not a panacea. Indeed, when the population is small, or the culture simple, or the population highly heterogeneous, print may have little ability to effect sociocultural-historical change.

Using the constructural perspective, it is possible to give the definition of profession. A *profession* is defined as a collection of individuals who are more

culturally integrated than the population at large and who have access to information not generally shared by outsiders. Using simulation, it is possible to explore the impact of print on the reality of the profession. In the absence of print, professions composed of few individuals and with relatively simply cultures (little specific professional knowledge) are quite stable and able to generate consensus among their members quickly. Print makes it possible for a professional group to have more members and a more complex culture and still attain stability and consensus as rapidly as a small simple professional group without print. Print makes the growth and stability of a profession less dependent on its relative size.

Print also helps a profession maintain itself when its members are few or isolated, even if it is culturally complex. Face-to-face communication encourages professions to remain small, as in this case consensus is achievable. Print, however, makes consensus within a profession less dependent on the group's absolute size. Moreover, print facilitates the growth of professions as it confers a decided advantage to professions that are relatively large. These results suggest that print may have allowed professions to grow larger than ever before even in the absence of high cultural integration and may have allowed professions to become more specialized than ever before without paying the price of cultural isolation.

The impact of print is also essential for the formation of academic discourse. Academic specialties can be thought of as professions. However, the academic specialist is distinguished from the ordinary professional in that the academic is concerned with innovation, the generation and diffusion of new ideas. By considering the relationship between discovery and diffusion, we examine the role that professional communication can play in building the professional scientific community. Simulations are used to examine how the sociocultural landscape of the scientific specialty affects the rate of information diffusion in that specialty.

### *1.5. Print and Face-to-Face Communication*

The dominant effect of print on academic structure is to increase the rapidity with which ideas diffuse. Print is simply more efficient than face-to-face

communication for the diffusion of new ideas, regardless of the shape of the sociocultural landscape. Through simulation, some potentially interesting relationships between print and the size, absolute or relative, of a specialty are isolated. The advantage of print to speed the flow of information increases as the absolute size of a specialty increases. Face-to-face communication (i.e., word of mouth) breaks down as a vehicle of dissemination as the absolute number of individuals who must be reached goes up. Face-to-face communication also degrades with the relative size of a specialty. Under face-to-face assumptions, larger is slower, whether one refers to absolute or relative numbers.

However, there is no such simple story about print's advantage when relative size is considered, that is, the size of a specialty relative to the rest of society. Under either face-to-face assumptions or print assumptions, diffusion within a specialty requires the help of some knowledgeable outsiders, individuals who share knowledge with insiders but not as much as insiders share with one another. Within the context of the simulations, specialties could be too small, relatively; that is, the number of outsiders may be so high that communication from the outside distracted the diffusion of new ideas by the internal specialists. Specialties that were too large, on the other hand, may have too few outsiders to make a useful difference. The communication of internal specialists may be too inbred, preoccupied with ritual ideas that delay communication of a new idea. Under print assumptions, new ideas diffuse fastest when the specialty is only moderately sized relative to the outside society.

Beyond the number of individuals, print also has a decided advantage when one considers the number of ideas in a specialty, its complexity. The more complex is the specialty, the more advantageous is print communication. This follows directly from the multiplicity of print. Print allows many ideas to be communicated at once with less risk of forgetting.

These results about print in specialties suggest two reasons why written texts have remained a medium of choice for scientific communication, both converging on the timely diffusion of new ideas. First, historical changes in science (whether

by choice or accident) created an environment where scientists could benefit from the rapid diffusion of discoveries: print made such rapid diffusion possible. Second, the scientific text has evolved to a form, especially through the use of citation that allows authors to engineer the diffusion of their ideas to the scientific community. Being able to engineer a text allows the researcher not just to diffuse the new idea, but also to diffuse it to the "right" people (and so establish prominence).

To understand the effect of print on sociocultural-historical change, one must understand how it plays out relative to the face-to-face technological context and across a variety of sociocultural-historical landscapes. Further, to understand the impact of print, one needs to understand the communication process more generally. This process, however, is sufficiently complex that it is difficult for humans to trace through the ramifications of even simple changes in communication technologies or sociocultural-historical landscapes. As an aid to such thought experiments, we employ a simulation model based on constructural theory. Using this model, it is possible to generate logically plausible hypotheses about print, professions, academe, and scientific migrants and to call attention to logical inconsistencies in other hypotheses. These analyses suggest that simple factors that make print-based communication different than face-to-face communication can result in major sociocultural-cultural impacts.

Special attention should be given to the enormous importance of two trivially simple physical factors about print. First, print increases the availability of the communication partner with a novel idea and so facilitates diffusion; that is, even when there is only one individual with a novel idea who can communicate with others, if that individual happens to be an author, the same idea can appear in multiple books. This multiplicity increases the chance that new information will diffuse, but only if the book is written in such a fashion that some of the knowledge in it is already known to members of the society. Diffusing new ideas depends on the audience and reader sharing a good deal of knowledge. Thus, mechanisms of print foster, and are in turn fostered by, the norms of diffusing new

ideas and social stability. Second, books, unlike people, cannot learn. Thus, the chance of learning the novel idea from a book remains constant; however, the chance of learning it from another individual decreases as that individual inevitably learns more information over time. The result is that in an oral society many communications become ritualized, and the ability of new ideas to diffuse is decreased. On the other hand, in a print society, such ritual time (insofar as it is spent seeking new ideas) can diminish because print eases the search for new information.

Our work has tried to reveal the lofty potential of print by means of the trivial mechanics through which it altered face-to-face communication. It remains for future historians and social scientists to tell us whether and how print has fulfilled this potential across time and place.

## 2. Literacy On-Line

At the same time, however, writing – even more than texts and languages – constitutes the soul of a society, precisely by its testimony to the immutable: It is difficult to bring about change, even orthographic change. This means that the disappearance of a system and its replacement mark the death of a civilization.

A shift in the practice of literacy such as is suggested by the term, literacy online, has epochal significance: The death of one world (that of print) and its replacement by something else (the online world) signals not just a change in communication or technology but a change in civilization itself. For the sake of simplicity it is possible to suggest the following definition: "Online literacy" means reading and writing with a computer. But reading and writing what and how? In theory, we can read a dense, complex novel like Henry James' Golden Bowl—a work created in and designed for the privacy, solitude, and deep interiority of print culture – sitting in front of a computer terminal. Of central importance, however, is not the theoretical possibility of being able to read anything from a computer

screen but whether or not readers and writers fully acclimated to the computer screen as the primary source of literacy would ever think to read, much less produce, such a work in the first place.

The fundamental problem here is the migration of the meaning of the essential term literacy, involving the dismissal of the original referents themselves, reading and writing; this shift is a problem of special concern for the compound term under current consideration, *"literacy online."* Most contemporary discussions of literacy dismiss the traditional concern with reading and writing by equating those terms with simple coding skills, knowing all along, and proving in the very books that these scholars write and ask their readers to comprehend, that coding skills are little more than a minimal and not a sufficient condition of being literate.

The assumption here is that the really important literacy skills (whatever they are) are built on top of minimal coding skills, and furthermore, that as the world becomes more technically advanced (for example, with many more computers), students are in ever-increasing need of more advanced skills. Technology, however, also has the ability to simplify work, and computers – especially with new graphic interfaces, touch-sensitive screens, and voice-recognition capability – seem to have the capacity to allow the complex transfer of information over time and space with fewer, perhaps no, traditional coding skills. People separated from their families with access to a telephone, for example, do not have to learn how to write letters, another historical use of literacy.

Literacy today regularly appears in compound-form to refer to a host of other, more important functions, suggesting that to get ahead, or survive, people need more than just coding skills. Plain "literacy" will no longer do; there is instead "functional literacy," "cultural literacy" (terminology of E.D.Hirsch 1987), or, more recently, "critical literacy" (C.Lankshear and P.L.McLaren 1993) – all presumably concerned with language competence as well as other issues – as well as literacies that may have little or nothing to do with language such as "geographic literacy" and "computational literacy" (or numeracy). "*Computer literacy*" would seem to be one of these terms not specifically related to language

competence, although it thankfully seems to have fallen out of favor, in part because hardware and software developers keep lowering the skill level required to use their products, making it, more difficult for educators to identify a discrete set of skills to teach under this new rubric.

Since "*online*," unlike many other compound terms, refers not to another domain of competence, but to a place, albeit a virtual one (one's virtual location while connected to a computer), the compound term online literacy would seem to have a far closer connection to traditional notions of literacy and the original referents, reading and writing. In this sense, online literacy would refer to the reading and writing one does at a computer. As such, any understanding of this compound term remains dependent on rehabilitating the terms reading and writing, or at least resisting the tendency to dismiss them as trivial and unrelated to serious concerns of literacy *per se*.

It is not helpful to see literacy as just reading and writing, ignoring the advanced demands that modern, industrial cultures have routinely placed on students in terms of these skills. It is not reasonable to expect that merely by virtue of silent engagement with the text ("reading"), people should be able to comprehend that with which they have had no prior experience, or that all people, and not just an elite, should be able to create texts that others (the society generally) would find interesting and informative (two extraordinary demands inconceivable in preindustrial cultures where there were few texts, and those mostly religious with rigidly controlled interpretations). At the heart of the problem is seeing the full complexity of the connection between literacy and technology. It is not just the alphabet that transforms literacy and, by extension, the world, but our entire technological struggle with nature. It is not the technology of printing *per se* that led to the deep reading and writing practices of the modern world but the emergence of technology generally, or, put differently, the emergence of printing within the context of our massive, conceited effort to extend infinitely our technological control of the world.

It’s a common knowledge that implicitly and often simplistically link literacy with the promotion of modernity – and not with the concomitant tradition of critical resistance. Literacy, or perhaps more accurately, the culture of print, may be seen as a major contributor to the two great motifs of modern history: the technological domination of nature and the political domination of indigenous peoples. Literacy is entirely associated with the most mechanical aspects of coding skills or language instruction in those skills; hence, it is completely disassociated from the prolonged, intense efforts at deep understanding and creation that constitutes the kind of reading, writing, study, and research that go into creating their own work.

The work of the sociologist, A.Gouldner (1976), remains unsurpassed in detailing the role of print literacy, what he calls the culture of critical discourse (CCD), in defining one of the most prominent and important features of modern, or print, culture, that of the distinct intellectual class charged with maintaining a critical distance from mainstream practices.

The school of thought that has been most promising lately is what might be called the study of book culture in the modem world - a phenomenon that begins with the Renaissance and hence with the earliest aspirations of technology and modernity.

*2.1. Computer literacy*

Like the typewriter a century before, the personal computer in just a few years has changed the means by which most people write. So phrased, one might wonder why there has been so much commotion about the personal computer when there has been so little interest in the impact of typing (something studied almost exclusively in the area of office automation). The answer lies in the fact that it is difficult to conceive of typing as having a truly transformative impact upon the practice of writing. It is obvious to everyone – even those who cannot type – just what typing does and does not accomplish. With computers there is so much interest, at least in part because of what we do not know, the mystery, the sense

that computers can transform practice, transform us, solve any number of our most basic problems—for example, how to write. It is not surprising, given this argument, that there is a relative lack of interest in what is by far and away the most dominant computer practice, word processing. M.Heim's “Electric language: A philosophical study of word processing” (1987) is an attempt to say something serious about the new technology of writing, and basically a metaphysical dead end, since word processing is ultimately an attempt to perfect the older writing practice that lies at the center of print literacy, helping individual writers compose discrete texts more proficiently.

S.Zuboff in his book “In the age of the smart machine” (1988) suggests a much more fruitful approach by looking at the new conditions of interaction, especially in the workplace, presented by the computer. What S.Zuboff realized was that the computer, once connected to other computers through networking, created an entirely new, more open, more collaborative world of human interaction. Computers were to have their greatest impact by allowing all people using electronic mail and other network software packages within a single complex institution to communicate directly with each other, that is, in alternative ways from the traditional hierarchical structure of that organization. Interesting work in this area is being carried on in the United Kingdom under the acronym CSCW, *Computer-Supported Cooperative Work* (D.Diaper and C.Sanger 1993, M.Sharples 1993). Here it should be noted that a major ideological issue concerning literacy inside the academy—the drive to redefine, or even eliminate, private authorship of texts, sometimes referred to under the provocative rubric of "the death of the author" (B.Agger 1994, S.Burke 1992) – draws little controversy in the world of work where collaborative writing and corporate (often bland) authorship has long been the norm.

Computer literacy is now being implemented into college composition writing. The United States is one of the few industrial countries that provides systematic instruction in composition in its native language to college students, and thus it is not surprising that people within this field of college composition should

have been early students of computer-based writing. One might argue that the traditional paradigm of print literacy was already on the wane in theoretical discussions of college composition, and perhaps practice as well, before the advent of personal computers. Traditional print literacy separates readers and writers, connecting them essentially through the mediation of the text. What is of greatest value is neither the other person's words, ideas, or thoughts (or even our own), nor insight as an abstraction; rather, the value is in the expression of insight, its embodiment and formal.

The value of the text resides in its origins apart from conversation, apart from the way people routinely discourse with each other. People could get together and talk, exchange their ideas, express their feelings, etc., but for the world of print literacy, such exchanges could never take the place of the deeper reflection and understanding available by separation and exchange of deep texts.

Working with texts and exchanging texts (called "process") became more valued than perfecting them (called "product"). This shift developed in part as a means of enhancing intimacy and also as a means of gaining a greater understanding of, and sympathy for, the psychological processes of other people.

Concomitant with the network classroom has been the emergence of a new way of conceptualizing the text itself, not as the intricate working of a single author (for example, not as a short story or novel), but as the amorphous collection by many contributors (that is, as an anthology or catalogue). In this sense, the class conversation can be seen as a text, one created by many discrete, individual pieces, potentially by different authors all linked together.

On one hand, the concept here is rather prosaic, even pedestrian, especially in light of the elevated notion of the text in print culture as a highly complex, intricately organized, and deeply personal entity. The new, computer-mediated text is instead to be a collection of information, perhaps even disparate information, accessible through a common set of commands or interface writes M.B.Spring in his book “Electronic printing and publishing: The document processing revolution” (1991) writes: "A new form of document is emerging. It is based on

the ability to store complex webs of documents and document components in a computer and to manage them with the computer with relative ease. This new form of 'document' in its most basic form is a document database or *docubase*"

On the other hand, this new form of computer "writing" continues to generate immense interest. Such interest draws on the pervasive dissatisfaction with print culture, especially among educators influenced by broader cultural criticism (and sensitive to the charge that print culture has been heavily involved with the subjection of not just indigenous peoples but minorities generally). Interest has also been generated by the marketing genius of T.Nelson (1992) in dubbing these textual databases as hypertexts. Hypertext seems to offer the promise of an even more radical transformation of print literacy than networked instruction, in part since its name suggests that it retains all the benefits of the former practice of print literacy (it is still a "text"), only now a better ("hyper") one. Not surprisingly, teachers and writers with close contacts to the world of print literacy first saw hypertext largely as a heightened form of traditional writing, a mode of expression that gave authors all the benefits of traditional print literacy and then something more. Michael Joyce took the lead in formulating a new poetics of hypertext, creating in *"Afternoon"* the most widely known hypertext short fiction, and in *"Of two minds"* (1995), a full-fledged poetics of hypertext. J.D.Bolter, a co-developer of the authoring software (*Storyspace)* that M.Joyce used to "write" his story, became the main spokesperson for this new mode of writing. In "Writing space" (1991), J.Bolter makes the almost Ptolemaic argument that all texts are essentially hypertexts and thus that, in abandoning linearity, literacy is finally returning to its true form. Meanwhile, G.P.Landow (1992) led the way in promoting hypertext as an invaluable adjunct to traditional literary study, a means for students to gain an infinite knowledge base while reading traditional texts (in Landow's case, mostly Victorian texts like the novels of Ch.Dickens).

Beneath all the excitement of hypertext is the more prosaic notion of the hardworking database, the single huge collection of information that can be accessed in different ways by users with different needs. A hypertext is essentially

a means to organize an enormous amount of information nonsequentially. T.H.Nelson (1992) himself argues for considering all information as part of a single hypertext – what he calls the "*docuverse*," something like the electronic equivalent of an immense traditional library, and hence something that is likely to be organized and accessed via impersonal search engines and indexing programs. Hypertext technology involves computers for storing huge amounts of seemingly unconnected information as a single database, and thus accessible by a unified set of commands. The essential insight here was fundamentally technological: As long as the information was stored electronically and appeared on a screen instead of in a bound page, one could move seamlessly through all the information. With electronic information there is no next page; one can move to any "page" next, any piece of information within the entire database, without having to reshelve one's current book, etc.

*2.2. The World Wide Web*

An early futurist H.G. Wells in his book “World Brain” (1938) foresaw the need to link all information, not just in a single box to be the personal tool of a lone user, but in a central location to be accessible for all users, in what H.G.Wells called the "World Encyclopedia." This notion offered an early but accurate description of the linking of information on the Internet to form the World Wide Web (WWW), or Web for short. Wells' World Encyclopedia would be the mental background of every intelligent man in the world. It would be alive and growing and changing continually under revision, extension and replacement from the original thinkers in the world everywhere. Every university and research institution should be feeding it. Every fresh mind should be brought into contact with its standing editorial organization. It would do just what our scattered and disoriented intellectual organizations of today fall short of doing. It would hold the world together mentally.

The Internet has come to be a one-word reference to the organization of a single global hypertext, largely through the protocol of linking and naming con-

ventions at the heart of the World Wide Web. Interest in the Internet and the Web remains incredibly high, fueled in part by the same technological optimism one sees in H.G.Wells. Its importance for defining online literacy derives from three interrelated factors. First, the Web is fundamentally a superb instrument (perhaps the greatest ever invented) for browsing, that is, for moving through a vast amount of information stored all over the world to find the information one wants and, in the process, to stumble across lots of other things that may be even more interesting. In the world of print literacy, the browsing or locating of information had far less status than the critical analysis, what might be called the formal comprehension, of the text. Finding a text was only a preliminary step to reading (understanding) it. There is, however, nothing to stop a migration in the term reading itself over the next few years, especially as more and more students gain regular access to the Web, in the direction of reading as locating material (a usage already in place in the sense of "reading" or overviewing a field).

Second, material on the Web is already often in highly graphic format, with pictures fully integrated with texts, with sound and even film clips available, and the trend can only be for more multimedia forms. Reading such materials, therefore, will be akin to reading cultural signs and artifacts generally – a process already being organized for academic purposes under the heading of cultural studies.

Third, the Web is also a vehicle for writing and, something with radical potential, direct self-publication, albeit in electronic rather than print format. The Web is not just a means of organizing all the information of the world but of allowing individual users and all over the world to contribute directly and instantaneously to this immense project (that daily grows more immense). Contributors use a common format for laying out (or "marking up") text (called HTML, "hypertext markup language") and for adding links to their own contributions, no matter how profound or banal. The Internet via the World Wide Web is a fast, inexpensive means of publishing, akin in terms of the broadcast world to allowing all users to be the originators and hence broadcasters of radio or television

programming and not just receivers or consumers. There are at least two sets of issues involved with writing via the Web. The first has to do with the rhetoric of electronic publication – namely, how to construct effective electronic works, including graphics and links. Here one is likely to see a migration of another key term, with writing coming to refer to the task of construction, assembling disparate elements of a project in a way akin to what we do today in preparing a formal report to a group: preparing visuals, a slideshow, perhaps even a video, as well as a traditional written report. Just as online literacy will entail reading more than traditional texts, writing will entail working with more than words.

The second set of issues has to do with the radical alteration in the notion of publication and hence authorship. In the world of print literacy, for example, a great deal of writing was conceived as mere training for publication and hence preliminary to true authorship. College students took numerous expository or creative courses, writing many works in each course, perhaps with the hope of getting just one of their works published in a campus magazine during their undergraduate years. Within a few years, these same schools will have the ability to have all students mount all their writing directly on the World Wide Web where, with the proper links, it will be immediately available to everyone in the world. It is unclear just what will happen to the notions of apprenticeship, quality, and authority that were essential parts of print authorship, although it must be noted that Internet enthusiasts generally see only good coming out of the breakdown in hierarchical structures and the emergence of full democratic participation. Institutional authorities may still be able to designate certain Web texts as "official publications," perhaps assigning them a graphic imprimatur, but it is unclear just what force such action will have when all texts are equally accessible. There may well not be any test of textual value other than popular appeal to readers. When everyone is an author (in terms of being published), then the value of authorship is liable to migrate (since value requires some notion of scarcity) from being published to being read, with consumers and possibly the marketplace as the final arbiter of value. The issue here is an ancient one, going back at least to Plato's

Republic, and having to do with the role of elites in either restricting or preserving a host of cultural ideals, even that of democracy itself.

## 3. Technology and the Textual Revolution

The inventions of writing and printing constitute the first two revolutions in information recording and transmission. The current development of electronic management of text, somewhat prefigured the typewriter, constitutes a third revolution. The invention of writing enabled humans to record their experience in a form which did not depend on oral transmission. Printing then standardized these texts, since once the type was set, there would be no more copyists' errors which had characterized the dissemination of handwritten manuscripts. Printing also democratized the written word, since it made written text available, in increasingly less expensive forms, to those who could read. It was a catalyst for literacy. But the production was a two-phase process (handwriting, then printing), and it imposed relatively long times from writing to dissemination. There was also an inverse relation between speed and quality. The invention of the typewriter in the nineteenth century – one of the more violent forms of writing, as a correspondent recently observed on the Internet – made a major contribution to combining speed and legibility, or quality of output.

These advantages were further enhanced by the electric typewriter in the mid-twentieth century. But the number of type-bars on the typewriter limited the number and variety of characters, fonts, and styles that could be output. The IBM golfball typewriter, and later daisy-wheel typewriters from many manufacturers, represented a major advance in allowing interchangeable fonts and type-styles. But text, once written, was fixed. The only means of editing were still the traditional scissors and glue (though lift-off correcting ribbon was an invention of greater importance to the development of editable text than is generally realized).

*3.1. Word processing*

The key to the modern textual revolution is the captured key-stroke. The adaptation of computers to the input, editing, and output of text, which acquired the name of "*word processing*," was based on machine-readable texts, which are texts encoded in machine formats and reusable, because the writer's keystrokes have been recorded on disk for later re-use by any writer, editor, or reader with appropriate electronic equipment. Indeed, *Optical Character Readers*, which convert printed paper text directly into machine-readable form without manual re-keying, have even extended this notion of text input. The two-phase process of writing and printing is reduced to a single step, or at most a step where the writer's text is finalized and then sent in that format to an expert printer for formatting and final output, without re-keying. The hot-type operators in enterprises like newspapers have been replaced by on-line systems where journalists compose their text and submit it to the typesetters, who work exclusively on screen-based documents. Because of the ability of the computer to manipulate strings of characters in various screen-based shapes, text becomes malleable. And because of the development of floppy disks, and then electronic networks, text becomes sharable over global distances for both readers and writers, who can thereby become collaborative authors. The cost in equipment terms for a word-processor can be less than a thousand dollars. The cost in human terms is keyboard competence and what is increasingly becoming a standard level of literacy in handling word-processing software. The benefit is an enormous increase in flexibility and power in the writing tools that we use.

There have also been fundamental changes to the relation between writers, texts and their writing tools. These changes are partly ergonomic, and partly conceptual. As one of the pieces of software most commonly used by people lacking specialized computer training, word-processors have had to make major adaptations in order to be "user-friendly," involving the use of screen, mouse, and keys in as transparent a way as possible. These considerations, which began at the functional level of manipulating individual chunks of text,

were later expanded to include higher-level issues of text planning. The model of plan-write-revise created by L.S.Flower and J.R.Hayes (1980) has underpinned much of the thinking behind the word-processor. The whole procedure can be divided into pre-writing, writing and post-writing in the computer environment.

Technology provides a number of tools for pre-writing support, which are sometimes bundled with word-processing packages. These include brainstorming software, which allows the user to develop ideas and make thematic and logical links between sub-themes, often in a hierarchical tree structure. These then lead to outliners, software which guides the writer in creating a shape for the argument into which text is then inserted (P.Wayner 1992). These tools fall into the "planning" category and represent well established components of writing research. They may be bundled with a more widely-based writing environment concept which provides support not only for pre-writing but also for writing and post-writing. *Writer's Helper* from Conduit Software, for instance, provides pre-writing activities like brainstorming and idea-associations, multiple views and connections between ideas, and outliners and organizers at different levels to provide proto-templates to shape the emerging text. Other software of this kind includes *Writer's Workbench* (P.S.Gringrich's "The UNIX Writer's Workbench" 1983), *Writer's Assistant* (M.Sharpies, J.Goodlet and L.Pemberton 1989), and *Writer's Partner* (M.Zellermayer, et al. 1991).

Writing support involves the functions of the mainstream word processor. They cover the insertion and deletion of text, including cut-and-paste editing. Users are able to scroll or browse freely through the text in both directions and to find and/or replace items throughout the text. Word-processing systems also handle multiple fonts, sizes and type styles; formulae, both composite characters and mathematical; color; text formatting; footnotes; page headers and footers; and boxing and shading. Non-textual components can be inserted into the text, including tables and spreadsheets, as well as graphics. The *WYSIWYG* ("what you see is what you get") convention is standard for screen display. In addition, it is

possible to view the text in page layout mode to determine the balance and distribution of text.

The process of writing is supported by a growing number of on-line resources which can typically be consulted from within the word-processing package. Among these are substantial dictionaries and thesauri as well as spelling checkers (these latter usually being seen as post-writing tools: see below). All kinds of statistics and counts of letters, words, lines, and paragraphs can be rapidly carried out, facilitating the kinds of writing, from student assessments to journalism to academic writing, where wordage is specified as part of the required outcome. Even if a specific piece of software is not available from within the word-processing package – if one wants to consult the Oxford English dictionary on CD-ROM, for instance – it is possible to start up another application and complete the lexical searches without having to close the document currently being edited. Reference works of many kinds, including reference grammars and large-scale corpora for on-line consultation, are increasing the power of the computer as a support tool for composition and are helping to turn it into a broadly based integrated writing environment.

The output of word processing is increasingly in the form of a laser- printed document, with definition equal to or greater than 300 dots/inch. This quality is sufficient for *desk-top publishing* (W.Watts 1992), particularly when supported by the layout facilities of software like *PageMaker*. Most writers on word processors lack the skills of professional printers and lay-out artists, so much ill-designed work has been printed. It is also possible to send fully-formatted documents by disk or through electronic networks, either in compressed form or using a format like *"rtf"* (*Rich Text Format*). Major word-processing software is increasingly able to handle documents in other word-processing formats: *Microsoft Word* and *WordPerfect*, for instance, can read documents in each other's formats.

Word processors can handle all the principal types of human writing systems, and can combine left-to-right, right-to-left, and vertical. One of the more dramatic

advances of word processing has been in character-based languages such as Chinese and Japanese. Traditional handwritten methods in these languages are beautiful but slow. Typewriters were cumbersome, with large palettes of type-blocks whose position the user had to memorize. If the user can type in Roman letters (*pinyin* for Chinese, and *romaji* for Japanese), however) the computer can select the most appropriate character. In cases of homonymy, the computer offers a number of possible characters, ranked according to their suitability for the current grammatical and/or semantic context, and the user selects the appropriate character to match the alphabetical input. Such systems require of the software more grammatical and semantic intelligence than standard alphabetically-based word processors. They also require more computing power, but they are incomparably faster than other methods. Various alternatives to non-alphabetical input for character-based languages are also in prototype.

The interface with word-processing systems is carried out by a mixture of keyboard and mouse-control options, the latter using icons, menus, and what are becoming standard mouse conventions for working with text (*click, select, drag*). The ergonomics of the interface and even mouse-use have been sources of some controversy. Expert word-processor users tend to minimize mouse use and prefer equivalent keystroke commands instead because of the ergonomic disadvantages of removing the hand from the keyboard to the mouse and back. On word-processing systems, it is possible to achieve speeds in excess of 150 words/minute, which is close to the speed of speech. Prototypes of voice-driven word-processing systems are also already being tested. Apart from their use as "no-hands" word processors, they will also be particularly valuable to vision-impaired users, since word-processing printers can naturally output Braille as well as conventional text. They belong with a wider agenda of research into *speech recognition*. Here, as in other ways, the word processor promises to democratize language tools and uses by reducing potential disadvantages of access.

The relation between writer and text using a word processor is different from that with conventional writing tools. S.A.Bernhardt's (1993) list of the properties of

screen-based text includes "navigable," "interactive," "graphically rich," "customizable," and "publishable." Probably the key property is that machine-readable text is malleable. Unlike handwritten or typewritten text, machine-readable text is a sequence of captured key-strokes, which means that changes to the text do not require rewriting. Text on disk or from the electronic networks can be copied to the user's computer and, provided that it is not in some specially coded format, reused, perhaps in shared-author mode. The notion of ownership and authorship of text has consequently changed. With so many text archives available on the Internet, on CD-ROM, and on other machine-readable sources, it is possible to take chunks of language from different places and weave them into a text.

The sociology of the text has also changed in the workplace. Secretaries are no longer principally copy typists from handwritten or dictated originals. As keyboard skills become standard for computer use, more authors are typing their own text, often sharing the composition with colleagues and/or secretarial staff. Collaborative writing is possible, especially over networks, in co-author mode, in author-comment mode, and in author-edit mode (where colleagues can insert comments for consideration.

Word processing can be addictive. Writers who have migrated to composing on the screen (as opposed to using the word processor for their final clean version) seldom revert to their former manual ways. In time, the keyboard can become as necessary a component of writing as paper and pencil once were.

Once the basic text exists it needs to be checked. Traditionally this was called "editing." In computer parlance, however, "editing" nowadays also includes the laying down of basic text as well as revision. N.Williams (1990) and others have given the name "*postwriting*" to a range of functions which have become a standard part of many leading word-processing systems. These tools can also be activated during writing.

Some post-writing tools allow the writer to check various aspects of the text. They include checking programs for spelling, grammar, and style (D.Ross and D.Hunter 1994). Most of these tools work in a modular way, each independently

and on very computer-defined principles: A spelling checker checks strings (i.e., sequences of characters) and will accept both *red* and *read,* and *to* and *too,* in *John red/read the book to/too*. "Intelligent" post-writing tools are currently under development. They will allow the post-writing tools to collaborate among each other to improve their performance. *Spelling checkers* will consider grammatical information and so separate red/read and to/too in the example above. *Grammar checkers* will not merely report "this sentence may contain a passive," but will check semantic information to filter out examples like *colorless green ideas sleep furiously*. They will also look at statistical counts and genre analyses so that more passives will be tolerated in scientific prose before a warning is issued to the writer. *Thematic analyzers* will check for thematic and rhetorical structure and development, and will alert the writer to digressions, repetitions, and failure to close off threads of the argument. *Writer's Helper* (Conduit Software) also provides coherence checks, word-frequency summaries, and tests (*inter alia*) for readability and usage.

Making such tools available to the learner or not-fully-competent writer poses obvious problems. It is like giving a black-box calculator to the naïve arithmetician, or a powerful statistical engine to a naive statistician. Black-box tools are excellent if three conditions are met. First, the user must have some broad idea of what the output should be; second, the input and output conditions must be clear; and third, the black box must work perfectly. At the present time, writers and post-writing tools fail to some extent on all three counts. It is necessary to understand, at least to some extent, how the tools work so that the user can interpret just what the tools are reporting back.

While corrections of spelling, grammar, and style can be carried out through software tools, it is still subjectively not easy to achieve a reliable evaluation of text on the screen. Where once the typed version was visually different enough to allow sufficient objectification, now the *WYSIWYG* convention, the increasing quality of output, and the fact that the text is composed and massaged on the screen all make the screen text almost too close to the output for writers to be able to

assess their text properly. One cannot so easily step back for an arm's length view. Accordingly, a large number of word-processor users, including many of the most competent, still print paper copies of text for checking. While the word processor is making great strides to achieve the status of a transparent tool, it is in some respects still a long way from its goal.

### *3.2. Consumption of text*

So far there has been a strong imbalance in the involvement of technology in writing and reading. While the production of text has adopted the technology approach, the consumption of text has overwhelmingly continued to be faithful to paper. Letters, faxes, journals, and books have continued to multiply under the stimulus of technology for creation and printing. But the medium for output has been predominantly paper. Far from creating a paperless office, word-processing technology has vastly promoted the use of paper and the destruction of forests.

The reasons for this conservatism are not hard to find. Paper print is cheap, requiring minimal technology for use. It is conventional: People know how to use it as a result of generations of literacy education. It is portable: To the bus, the beach, the coffee house, whether a computer is available or not. And readers can work and interact directly with paper-print media, marking important passages for later attention. The response of technology has so far been incomplete. The laptop computer is a partial answer, though its relatively short battery life limits its use away from power sources. Portable electronic notebooks like *Apple's Newton* will certainly contribute to the technologization of reading and the decline of paper in the para-work functions of diaries, note-keeping, and record maintenance.

The two technologies that have done the most to promote electronically-based reading are CD-ROM and electronic networks. CD-ROM, because of its very large capacity, and because until recently such media have not been available for multiple rewriting, tends to be a medium favored for the large-volume commercial distribution of established text. Dictionaries and encyclopedias, databases, large text archives, and multimedia with sound, graphics, and text have found CD-ROM

relatively inexpensive and highly reliable. In the longer term, however, the electronic networks will revolutionize access to text for reading. The networks will remove the need physically to transport even CD-ROM disks from place to place. At the present time, the large size of multimedia documents is making the networks, with their high load and limited bandwidth, a relatively slow medium. There are also important issues of copyright and charging to be solved before commercially published material will be available on the networks. But the global library (the "*docuverse*") is already well in plan, and some prototypes are already running (works of G.P.Landow and P.Delany 1993, J.Virbel 1993).

*3.3. Hypertext and hypermedia*

There is one principal direction in which electronic delivery and consumption of text have made a distinct advance. This involves the notion of *hypertext* and *hypermedia*. (The term "*hypertext*" was invented by T.Nelson in 1965). At a simple level, anything that interrupts the temporal or left-to-right (in a language like English) linear march of the printed word is a hyperphenomenon. Footnotes, endnotes, references, and indexes are all modest examples of hypertext—text which can be read as a side-track from the main sequence of the master text itself. Their status as hypertext is especially clear from the viewpoint of the reader, who may choose (or not) to follow these diversions. For the pre-technological author, creating this apparatus can be onerous and is often undertaken in an unstructured way both during and after the composition of the text. But deleting a footnote in modern word-processing packages merely causes the software to renumber all following references automatically. The software can also handle references and indexing. Given reference keys – it is necessary manually to mark the authors and works to be referenced, and the words and topics to be indexed – word-processing programs are able to handle these functions automatically and in a variety of accepted formats.

A conceptually related phenomenon concerns making comments and notes on a text. These notes may be in the margin of the text itself, on a separate sheet of

paper, or on a "stickie" – a yellow stick-on note. Such notes may be simply part of "thinking about the text while reading." They may also be part of the beginning of writing about the text as the reader marshalls ideas and reactions. In any of these cases, they constitute a lateral digression from the text and are therefore also hypertexts. But machine readable texts do not readily lend themselves to such treatment. They have no wide margins for pencilled additions. Adding comments in the body of the text itself, perhaps offset with a line of asterisks, tends to interrupt the reading and so interferes with the integrity of the text itself. For this reason, there have been significant limitations on the usability of machine readable texts until the advent of software like *Notes* (C.M.Neuwirth, et al. 1987) from Carnegie-Mellon University or *Annotext* from *Panda Software*. Other computer analogues include the appearance of "balloon help" explanations as the cursor moves across different items on the screen. On-line help, whether automatically or manually activated, is another kind of hypertext.

This breaking of the traditional dependence on linearity has arguably been the most radical effect of technology on text. The concept was first elaborated by V.Bush (1945) in his discussion of ways in which texts and information could be linked laterally. The concept of cross-indexing is, of course, much older than V.Bush and has long been commonplace in large book indexes, thesauri, and such reference works. Users of reference books regularly operate in hypertext mode by picking up references and pursuing them through other keyword entries. But it was not really until the release of the software application *HyperCard* on *Macintosh* computers that the necessary programming tools became readily available and sufficiently easy to use.

In hypertext, the reader can read in a linear fashion, but s/he can also follow links to other texts and topics. The jumping-off point can be a word, phrase, paragraph, or text. The person who sets up the basic text has to create the lateral links (*hyperlinks*) to the other material, the presence of a link often being signalled by special highlighting features of the text, like color or underlining. The user usually positions the mouse pointer over the highlighted text and clicks the mouse

button. This activates the hyperlink and displays whatever material is linked to the text at the jumping-off point. For instance, a reader may be perusing a text on da Vinci. A link, indicated by highlighted text concerning da Vinci's ideas on machines of war, may connect to another article on the history of the tank, and a further highlighted link may lead to the basics of explosives or Chinese uses of gunpowder.

Links can be anchored on text, or on a graphic symbol, picture, or audio file as well, and links can lead to any of these, either singly or in combinations. There is software which plays Beethoven symphonies and allows the user to follow the score; to interrupt and ask for musicological and historical comment on specific sections of the music; to go forwards or backwards in the composition; to see pictures of Beethoven, his contemporaries, and his surroundings; and so on. Following from this example of multimedia, it is easy to see how hypertext is, in one sense, merely a special case of *hypermedia*.

Users of hypertexts have an urgent need of navigation tools. One can easily become lost and forget where a hypertext exploration began unless the computer can insert electronic markers, like bookmarks, to enable the reader to retrace his/her steps if needed. Navigation tools for hypertext may be in the form of graphics, or as a list of nodes visited during the current session. Even with navigation aids, users can become confused, and there are some undisputed dangers in the use, and particularly the over-use, of hypertext.

Creating hypertext documents can be laborious. It is necessary to assemble all the texts, graphics, audio, and other files which are to be linked together; then software tools must be used to build the links, with anchor points and targets. Software like *HyperCard* is able to do this in a fairly transparent way; so too can HTML ("*hypertext markup language*"), a set of conventions for inserting links between texts, graphics, and other items on the screen. HTML is widely used in the World Wide Web. Documents containing hypertexts are "read" by software called a hypertext browser, of which *Netscape* and *Microsoft Internet Explorer* are

common current examples. Software is also now coming available to convert word-processing document formats directly into hypertext format.

The exploitation of hypertext is fundamental to the operation of information browsing tools and searches on electronic networks. Hypertext links can point not only to other sections of text (graphics, etc.) in the same document, or other documents on the same computer, but also to texts and other files on any computer linked to the Internet. Software like gophers and the World Wide Web are built on hyperlinks and potentially allow global searches through information on millions of computers. Users can, for example, follow a thread starting from da Vinci on a computer in Boston and end up consulting an archive in Hong Kong on Chinese uses of gunpowder. Hypertexts are providing a new way of linking knowledge, like a massive distributed encyclopedia. They are also, in fundamental ways whose importance we are only starting to explore, beginning to build new models of information searching and the construction of knowledge.

### *3.4. Interactive reading*

An entertaining application of hypertext, and one which was in operation before the computer-based implementation of hypertext, is the *interactive text*. Often in forms of novels, interactive texts allow the reader to choose which path to follow at given points in the story. In paper form, this can achieve substantial proportions if all the various paths through the text are to be accommodated. In computer terms, it becomes a matter of generalizing a hypertext link from a digression or help/ auxiliary function to a principal choice in the linear progression of the text. There are significant new conceptual and cognitive issues at work here. Hypertext plays an important role in developing new notions of literacy.

Electronic publication on the Net is now starting to make substantial progress after an uncertain start. *Electronic journals* are starting to multiply as are advanced modes of scholarly discourse on the net. This is evident from the growing use of the Net in academic pre-publishing. With the proliferation of journals, many libraries are unable to keep pace even in targeted fields; as a result, the circulation

of pre-publication papers for comment has become standard in a number of disciplines. Many such papers, in disciplines including physics and linguistics, are archived at the Los Alamos laboratories as the result of an initiative by Paul Ginsparg. There are indications that such dissemination of ideas will become one of the leading-edge media for academic interchange. Other, more formalized pre-publication discussion media can be found in the electronic forum *Psycoloquy*, edited by Stevan Harnad. Harnad's concept, following the example of the paper journal *Behavioral and Brain Sciences*, which he founded, involves interactive dialogue of a kind uniquely well suited to electronic media. Most communications are by e-mail. Papers submitted are sent to a number (often about 20) of reviewers, who e-mail their comments to the editor. The author produces a reply, and the whole package of paper, comments, and rejoinder is published within 6 weeks or so of submission. In this way, an editor-mediated scholarly interchange is combined with the merits of dialogue and rapid turnaround times made possible by the Net.

If the Net has started to encourage formal writing of a scholarly kind, it has also been a powerful catalyst for written communication between a very wide range of users. The typewriter, airmail, and the fax all were catalysts for written communication. (The telephone tended to bypass it.) But now the Net presents a cheap, fast, and convenient resource, one which has the advantage of allowing messages to be collected when next the user logs on. It will even allow interactive written conversation in real time with software like *TALK* or *Internet Relay Chat*. But the principal medium has been electronic mail, which has grown explosively. This channel is fast, typically unedited or lightly edited, immediate, and often very personal. Spelling, grammatical, and semantic errors abound. In spite of books like “Netiquette” (by V.Shea 1994), the Net remains a shapeless, but uniquely vigorous, forum of personal interchange. A whole range of abbreviations has come into use to save time and space, like "*IMHO*" ("in my humble opinion"). And a set of "emoticons" has been invented – faces drawn with alpha-numerical symbols from the keyboard like :-) (a happy face, tilted 90 degrees to the left) – with the

aim of ensuring that messages are not misunderstood in the absence of real-time tone of voice and facial expressions. The Net is, in fact, bringing about a renaissance in epistolarity. It has done something to remedy the depersonalization of communications brought about by the typewriter and the decline of calligraphy as a personal art. Net has helped to create global communities of scholars and people of like mind and like interests, and it will be intriguing to see whether the new wave of letter-writing survives the expansion of the Net to carry real-time audio and video conferencing at reasonable cost.

### *3.5. Automatic production of text*

The next step beyond word-processing is the *automatic production of text*. A common example is the mail merge software available with many word-processing packages. The user creates a form letter with slots specified for the insertion of material (technically, text variables). A second list specifies, for instance, the name and address of the addressee. The computer can then construct as many letters as there are addressees, replacing the variables in the form letter with data from the address list. Mail merge programs are even able to handle simple programming of the style: "*If gender is MASC insert 'he', if gender is FEM insert 'she'*." Similar software can handle output from data-bases in an apparently personalized way, from bank statements to examinations and sales literature. The general concept of reusable text (described by R.A.Buchanan 1992) has other applications, for instance, in the creation of glossaries of chunks of text frequently re-used. The savings in human typing and clerical time are enormous, but so – in the case of sales-promotion literature—is the cost in paper and superfluous information. The emerging problem, indeed, is the glut of information and the urgent need for software tools to filter incoming material.

A famous precursor of one of the ideas behind the mail merge software, and one with much greater intellectual importance, was the *ELIZA* software developed by J.Weizenbaum (1966). *ELIZA* ran a typed dialogue between a patient and what purported to be a psychiatrist:

- Men are all alike
- *In what way?*
- They're always bugging me about something or other.
- *Can you think of a specific example*?
- Well, my boyfriend made me come here.
- *Your boyfriend made you come here?*
- He says I'm depressed much of the time.
- *I am sorry to hear you are depressed.*

The algorithm underlying this superficially convincing dialogue is remarkably simple. The boyfriend episode is simply a repetition, and the computer handles the issue of depression simply by picking up the key word "*depressed*" and reiterating it in a standard syntactic envelope "*I am sorry to hear that you are x*." The same idea of constructing output messages on the basis of form responses and data from a data-base, but with considerably more sophistication, also underlies the ability of a number of expert systems to "converse" with users in more or less natural language. The well-known *Neomycin* system for medical diagnosis, for instance, asks for information on the patient, suggests tests, and collects and analyzes data. It then combines the results in the form of a report in English which is displayed to the user. Going a step further, it is now becoming possible to generate reports and documents more directly from databases. Biographical records of individuals, for example, contain many fields of tagged information, and some of these fields (e.g., "*Achievements*") contain blocks of natural language. A request for the names, dates, and achievements of painters of the Flemish School, for instance, could be presented in the following form:

*There were four main representatives of the Flemish School. The first was x. Born in y in the year z, he ...*

All of these forms of automatic text production, however, are keyed to varying degrees to pre-established text, which is then output in the framework of other, more or less pre-processed, text formats. More demanding is the requirement that software produce a digest on stories and newspaper reports and, in general, the

automatic retrieval of information from texts. Machine translation, where text output is prompted by text input in another language, presents problems of bilingual language management in text production. The real test of automatic text production, however, involves the difficult problems of natural language generation, where computers generate language to express meaning. These issues are currently at the focal point of intense research in artificial intelligence and expert systems.

The investigation of the human writing process is beginning to create a new role for computer models as catalysts for inter-disciplinary linking. Text editing takes us from word processing to literary editing, and the study of literary texts as products of writing connects with the analysis of the language of literature. Stylistic management is also part of corpus linguistics as is computational lexicography and the automatic production of lexicons. The information structure of text is also part of information science, as is indexing, which itself links with hypertext. The persona of the writer and reader in writing are also part of user modelling in text generation. Creative writing forms a bridge between technological text management and the study of cognitive creativity. Text planning in word processing links to text planning in natural language generation. And the ergonomics of interaction with electronic writing tools take us into human-computer interaction (in journals like the *International Journal of Human-Computer Studies*). The electronic production of text is therefore not merely a technological extension of studies of the writing process in applied or educational linguistics; nor is it only a branch of the adaptation of technologies for the purposes of human communication. It is, in a powerful sense, a metaphor and a vehicle for the epistemology of writing in technology.

## 4. Electronic Text Media

In publishing, the traditional media of spoken and printed language have by now become secondary media which the electronic versions are converted into

only when needed. Even texts which have long existed in the traditional print medium are nowadays being transferred into the electronic medium in order to make them susceptible to the methods of electronic processing. Examples are the complete texts of classical Greek and Latin, the complete Shakespeare, and the Encyclopedia Britannica, which are now available on CD-ROM.

Compared to the printed version of a multi-volume edition, the electronic medium has the advantage of compactness, comfort, and speed. The information can usually be stored on a single CD-ROM. Instead of one having to haul down several volumes from the shelf and leaf through hundreds of pages by hand in order to find a particular passage, the use of the CD-ROM merely requires typing in the keywords.

Given suitable software, it is also possible to search using combinations of words, enabling the retrieval of all passages in which, for example, painter, Venice, and 16th century occur within a certain stretch of text. These methods of search can be life-saving, for example, when a textual database is used for diagnosing a rare disease, or for choosing a particular medication.

Another advantage of the electronic medium is the editing, formatting, and copying of text. In the old days, newspaper articles were put together with mechanical typesetting machines. Information coming in from a wire service had to be typeset from the ticker tape letter by letter. To make room for some late-breaking piece of news, the type had to be rearranged by hand.

Today the production of newspapers is done primarily on-line in soft copy. Contributions by wire services are not delivered on paper, but by telephone, whereby a modem converts the signal into the original layout. Form and contents of the on-line newspaper can be freely reformatted, copied, and edited, and any of these versions can be printed as hard copies without additional work.

The form of a newspaper article, like any text, is based on such structural features as title, name of author, date, section headers, sections, paragraphs, etc. In the electronic medium, this textual structure is coded abstractly by means of

control symbols, for example, a newspaper text with control symbols will look like follows:

*<HTML>*

*<HEAD>*

*<TITLE>9/4/95 COVER: Siberia, the Tortured Land</TITLE>*

*</HEAD>*

*<BODY>*

*<!- - # include "header.html" -->*

*<P>TIME Magazine</P>*

*<P>September 4, 1995 Volume 146, No. 10</P>*

*<HR>*

*Return to <A href =" ../../../../.. /time/magazine/domestic/toc/ 950904.toc.html">Contents page</A>*

*<HR>*

*<BR>*

*<!- - end include - ->*

*<H3>COVER STORY</H3>*

*<H2>THE TORTURED LAND</H2>*

*<H3>An epic landscape steeped in tragedy, Siberia suffered grievously under communism. Now the world's capitalists covet its vast riches </H3>*

*<PxEM>BY <A href=" ../../../../.. /time/bios/eugenelinden.html"> EUGENE LINDEN</A>/YAKUTSK</EM>*

*<P>Siberia has come to mean a land of exile, and the place easily fulfills its reputation as a metaphor for death and deprivation. Even at the peak of midsummer, a soul-chilling fog blows in off the Arctic Ocean and across the mossy tundra, muting the midnight sun above the ghostly remains of a slave-labor camp. The mist settles like a shroud over broken grave markers and bits of wooden barracks siding bleached*

*as gray as the bones of the dead that still protrude through the earth in places. Throughout Siberia, more than 20 million perished in Stalin's Gulag. ...*

To be positioned in example above, the text was copied electronically from a publication of TIME magazine available on the Internet. The example contains control symbols of the form <...>, which specify the formatting of the text in print or on the screen. For example, <P>September 4, 1995 Volume 146, No. 10</P> is to be treated in print as a paragraph, and <H2>THE TORTURED LAND</H2> as a header.

At first, different print shops used their own conventions to mark the formatting instructions, for which reason the control symbols had to be readjusted each time a text was moved to another typesetting system. To avoid this needless complication, the International Standards Organization (ISO) developed the SGML standard.

## 5. SGML: Standard Generalized Markup Language

A family of ISO standards for labeling electronic versions of text enables both sender and receiver of the text to identify its structure (e.g., title, author, header, paragraph, etc.) (Dictionary of Computing, p. 416 (ed. by J.Dlingworth et al., 1990)

The SGML language has been adopted officially by the USA, the European Union, and other countries, and has become widely accepted by the users. Texts which use SGML for their markup have the advantage that their formatting instructions can be automatically interpreted by other SGML users. An easier to use subset of SGML is XML, which is oriented towards handling hypertext. In addition to the standardized coding of textual building blocks such as header, subtitle, author, date, table of contents, paragraph, etc., there is the question of how different types of text, such as articles, theater plays, or dictionaries, should best be

constructed from these building blocks. For example, the textual building blocks of a theater play, i.e., the acts, the scenes, the dialog parts of different roles, and the stage descriptions, can all be coded in SGML. Yet the general text structure of a play as compared to a newspaper article or a dictionary entry goes beyond the definition of the individual building blocks.

In order to standardize the structure of different types of texts, the International Standards Organization began in 1987 to develop the TEI-Guidelines. TEI stands for text encoding initiative and defines a DTD (document type definition) for the markup of different types of text in SGML.

SGML and TEI specify the markup at the most abstract level insofar as they define the text structure and its building blocks in terms of their function (e.g., header), and not in terms of how this function is to be represented in print (e.g., bold face, 12 pt.). For this reason, texts conforming to the SGML and TEI standards may be realized in any print style of choice.

An intermediate level of abstraction is represented by the formatting systems developed as programming languages for type-setting only a few years earlier. .Widely used in academic circles are TEX, developed by D. Knuth, and its macro package LATEX. Since they were first introduced in 1984 they have been used by scientists for preparing-camera ready manuscripts of research papers and books.

At the lowest level of abstraction are menu-based text processing systems on PCs, such as *Winword* and *WordPerfect*. They are initially easy to learn, but their control is comparatively limited, and they are incapable of handling longer documents. Also, transferring text from one PC text processing system to another is difficult to impossible.

In summary, SGML and TEI focus on defining the abstract structure of the text, TEX and LATEX focus on control of the print, and PC systems focus on the ease and comfort of the user. Thereby the higher level of abstraction, e.g., SGML, can always be mapped onto a lower level, e.g., LATEX. The inverse direction, on the other hand, is not generally possible because the lower level control symbols have no unambiguous interpretation in terms of text structure.

SGML/TEI and TEX/LATEX have in common that their control symbols are placed into the text's source code by hand; then they are interpreted by a program producing the corresponding print. PC systems, on the other hand, are based on WYSIWYG (what you see is what you get), i.e., the look of the print is manipulated by the user on the screen. In this process, the software automatically floods the text's source code with cryptic control symbols.

For authors, the production of camera-ready manuscripts on the computer has many practical advantages. With this method, called desktop publishing (DTP), the author can shape the form of the publication directly and there are no galley-proofs to be corrected. Also, the time between text production and publication may be shortened, and the publication is much less expensive than with conventional typesetting.

For linguists, on-line texts have the advantage that they can be analyzed electronically. With most current publications originating in the electronic medium it is only a question of access and copyright to obtain arbitrarily large amounts of on-line text such as newspapers, novels, or scientific publications in various domains.

One linguistic task is to select from the vast amounts of electronically stored text a representative and balanced sample of the language at a certain time. Another is to analyze the texts in terms of their lexical, morphological, syntactic, and semantic properties. In either case, linguists are not interested in a text because of its content or layout, but as a genuine instance of natural language at a certain time and place.

There are many possibilities for processing an on-line text for linguistic analysis. For example, using some simple commands one may easily remove all control symbols from the text and then transform it into an alphabetical list of word forms, as it is shown below:

*Alphabetical list of word forms*

| | | | |
|---|---|---|---|
| 10 | 20 | a | a |
| 146 | 4 | a | above |
| 1995 | a | a | across |

and
and
and
and
Arctic
as
as
barracks
grave
grievously
Gulag
has
in
in
in
its
its
LAND
land
landscape
like
LINDEN
Magazine
markets
mean
metaphor
midnight
midsummer
million
mist
more
mossy
muting
No
Now
of
of
of
of
of
off
page
peak
perished
place
places
protrude
remains
reputation
riches
settles
shroud
bits
bleached
blows
bones
broken
camp
capitalists
come
Siberia
Siberia
siding
slave-labour
soul
chilling
steeped
Story
suffered
sun
that
that
The
the
the
the
the
the
the
the
the
the
through
Throughout
to
to
TORTURE
tragedy
tundra
vast
Volume
wooden
world’s
An
as
at
COVER
EUGENE
fulfills
ghostly
in
Ocean
Over
communism
Contents
covet
dead
death
deprivation
earth
easily
Return
September
Siberia
Stalin’s
THE
TIME
under
/Yakutsk
epic
Even
exile
fog
for

In this list, word forms are represented as often as they occur in the text, thus providing the basis for word-form statistics. It would be just as easy, however, to create a unique list in which each word form is listed only once, as for lexical work. Another approach to analyzing an on-line text for linguistic purposes is measuring the co-occurrence of word forms next to each other, based on bigrams and trigrams.

These methods all have in common that they are letter-based. They operate with the abstract, digitally coded signs in the electronic medium, whereby word forms are no more than sequences of letters between spaces. Compared to nonelectronic methods - such as type-writing, typesetting, card indices, search by leafing and/or reading through documents, or building alphabetical word lists by hand, - the electronic computation on the basis of letters is fast, precise, and easy to use.

At the same time the letter-based method is limited, in as much as any grammatical analysis is by definition outside of its domain. Letter-based technology and grammatical analysis may work closely together, however. By combining the already powerful letter-based technology with the concepts and structures of a functional, mathematically efficient, and computationally suitable theory of language, natural language processing may be greatly improved.

## 6. Electronic Media and Speech Technology

The expressions and texts of natural language may be realized in different media. The nonelectronic media comprise the sounds of spoken language, the letters of handwritten or printed language, and the gestures of signed language. Spoken and signed language in its original form has only a fleeting existence. Writing, on the other hand, is the traditional method of storing information more permanently, e.g., on stone, clay, wood, parchment, or paper.

A modern form of storing information is the *electronic medium*. It codes information abstractly in terms of numbers which are represented magnetically. In contrast to the traditional means of storage, the electronic medium has the advantage

of greatest flexibility: the data may be copied, edited, sorted, reformatted, and transferred at will.

The electronic medium may represent language in a realization-dependent or a realization-independent form. The realization-dependent form reproduces accidental properties of tokens in a certain medium, such as a tape recording of spoken language, a bitmap of written language, or a video recording of signed language.

The realization-independent form represents language as abstract types, coded digitally as electronic sign sequences, e.g., in ASCII (American standard code for information interchange). Due to their type character, they may be recognized unambiguously by suitable machines, copied without loss of information, and realized as token surfaces in any imaginable variant in any medium.

In communication, there is a constant transfer between realization-dependent and realization-independent representations. During recognition, the cognitive agent must map realization-dependent representations into the realization-independent ones (d→i transfer). During synthesis, realization-independent representations must be mapped into realization-dependent ones (i →d transfer).

Neither of these directions is trivial to model in computational linguistics, but for different reasons. When building a speaking robot, the challenge with an i →d transfer into spoken language (speech synthesis) is to make it sound natural relative to a free range of utterance situations. The challenge with an d→i transfer from written (optical character recognition) or spoken language (speech recognition) is to correctly interpret tokens from a wide range of different realizations.

The most primitive form of d→i transfer leaves recognition to humans. It consists in typing spoken or written language into the computer. This method is still widely in use, such as dictation in the office, transcription of tape recordings in psychology, or electronic typesetting of books which previously existed only in traditional print.

Automatic d→i transfer from printed language is based on optical character recognition (OCR). Part of an OCR system is a scanner which makes an image of the page as a bitmap - like a camera. Then the OCR software analyzes the image line by

line, letter by letter. By comparing the bitmap outline of each letter with stored patterns, the writing is recognized and stored in a form as if it were typed in.

The input to an OCR system may vary widely in font type, font size, and the form of layout. Even within a given document there are head lines, footnotes, tables, and the foot lines of pictures to deal with. Modern OCR systems handle such challenges by means of an initial learning phase in which the user corrects misclassifications by telling the program whether a certain constellation happens to be, for example, *ii* or *n*.

In addition, OCR systems use large dictionaries on the basis of which they decide which of several possible analyses constitutes a legitimate word form. In this manner a high recognition rate is achieved, sufficient for practical use. Depending on the type of machine a page may take between 50 seconds and a few minutes. The power of scanners and their OCR software has improved considerably since 1980 while prices have fallen. For these reasons the use of scanners in offices has greatly increased.

The speed of today's OCR systems is quite competitive, especially in light of the fact that the machine does not become tired and that the operation of the scanner can be left to unskilled labor. The most important aspect of language transfer in general, however, is the avoidance of errors. In this respect, the human and the mechanical forms of transfer are equal in that both require proof-reading.

Automatic d→i transfer from spoken language turns out to be considerably more difficult than that from written language. Whereas words in print are clearly separated and use uniformly shaped letters, speech recognition must analyze a continuous stream of sound and deal with different dialects, different pitches of voice, as well as background noises.

The possible applications of a good automatic speech recognition are tremendous, however, because there are many users and many circumstances of use for which a computer interaction based on spoken language would be considerably more user friendly than one based on the keyboard and the screen. Therefore automatic speech recognition is subject of an intensive worldwide research effort.

The projects range from a typewriter capable of interpreting dictation to telephone-based automatic information systems (e.g., for train schedules) to *Verbmobil*. *Verbmobil* created in 1993 by is intended as a portable computer into which the users can speak in German or Japanese to obtain a spoken English translation. Its use presupposes that the German and Japanese partners have a passive knowledge of English. In this way, the hearer can understand the output of the system, and the speaker can check whether the system has translated as intended. The system is limited to the domain of scheduling meetings.

The quality of automatic speech recognition should be at least equal to that of average human hearer. This leads to the following desiderata:

- *Speaker independence*

The system should understand the speech of an open range of speakers with varying dialects, pitch, etc. - without the need for an initial learning phase to adapt the system to one particular user.

- *Continuous speech*

The system should handle continuous speech at different speeds - without the need for unnatural pauses between individual word forms.

- *Domain independence*

The system should understand spoken language independently of the subject matter - without the need of telling the system in advance which vocabulary is to be expected and which is not.

- *Realistic vocabulary*

The system should recognize at least as many word forms as an average human.

- *Robustness*

The system should recover gracefully from interruptions, contractions, and slurring of spoken language, and should be able to infer the word forms intended.

Today's continuous speech systems can achieve speaker independence only at the price of domain dependence. The prior restriction to a certain domain - for example, train schedules, or when and where to meet - has the advantage of

drastically reducing the number of hypotheses about the word forms underlying a given sound pattern.

Utilizing domain knowledge is always crucial for inferring the most probable word sequence from the acoustic signal in both, human and artificial speech recognition. The point is that the current domain should not be prespecified by design or have to be preselected by the user. Instead, the system should be domain-independent in the sense that is can determine the current domain automatically.

The vocabulary of speaker-independent continuous speech recognition systems is still limited to no more than 1000 word forms. An average speaker, however, uses about 10 000 words - which in English corresponds to about 40 000 word forms. His or her passive vocabulary is about three to four times as large. Therefore a speech recognition system for English would have to recognize 120000 word forms in order to be in the same class as an average speaker.

Speech recognition will be fully successful only if the technological side is supplied continuously with small bits of highly specific data from large stores of domain and language knowledge. These bits are needed only momentarily and must be provided very fast in order for the system to work in real time.

Therefore, the crucial question for designing truly adequate speech recognition is: How should the domain and language knowledge best be organized? The answer is obvious: *Within a functional theory of language which is mathematically and computationally efficient.*

The better natural communication is modeled on the computer, the more effectively speech recognition can be supplied with the necessary bits of information. Conversely, the better the functioning of speech recognition, the easier the d=M transfer and thus the supply of knowledge needed for understanding during human-computer dialog.

## Conclusion

In spite of the robust growth of technology-based text production, the paperless office is only a little closer than in 1990. The advent of the word processor has resulted in a huge expansion in the volume of text produced (and, as we have seen, some qualitative changes as well); it has also led to a very large growth in the amount of printed text, both public and personal. Technology has fed text production, but it has not yet saved many, if any, forests. The factor which is likely to do more to focus text more strongly on electronic media is electronic networks. As more users over a wider demographic and geographical profile are connected to the Net, as text sources are increasingly published and archived in electronic form, and as libraries move closer to electronic storage and delivery, so will there be greater initiatives for accessing and reading text in electronic format.

Technology has changed the nature of text, including its ontology. Before electronic text processing, text was a linear sequence of letters with a start and a finish. Now it is rather a network of meaning potentials, waiting to be constructed by individual readers and users depending on their contexts and goals. This interpretation fits comfortably with contemporary theories of semiotics about the construction of meaning. It has always, in some sense, been present or at least imminent in text. But now we have a physical and mechanical means of coding many of these networks explicitly in terms of hypertext links. We not only have metaphors but also concrete models which can be used to represent and test ideas about the nature and operation of such hyperlinks.

Other traditional properties of text have also been changed, or at least colored, by technology. Text is now malleable. It can be easily shared and transmitted cheaply over long distances. It can be part of interactive dialogue, in real time or as intellectual interchange. Text can now, in a real sense, be less monologic than it once was. Text will also, in the reasonably near future, be naturally and reliably produced from speech. Software and hardware now becoming commercially available are able to convert speech into machine-readable texts, and either act on them (robotics) or

reproduce them as text (automated text-registration). These systems are still not commercially fully functional, but their advent is a matter of time and money rather than of fundamental shortfalls in research.

The way ahead, then, envisages increasingly intelligent software tools for supporting writing. It incorporates multimedia in text and it includes hypertext, with software tools to facilitate both. The one feature that has not changed through this revolution is the keyboard, the main interface tool between writers and their text-producing technology. Many people do not know that the current keyboard layout was designed to slow typists down, and avoid the jams of type-bars which occurred in early typewriters. Many new designs for text input and editing have also been proposed. Some have argued that direct voice input will make devices like keyboards redundant. All interaction with text will be via voice, and all text will be electronic. There is no doubt that voice-driven text management is coming. In the meantime we have developed a staunch attachment to QWERTY keyboards and have made them very much part of the interface between writers and their texts. This process is ongoing.

# CHAPTER III
# LINGUISTIC CORPORA AND LEXICOGRAPHY

## 1. Corpus Linguistics

Over the past ten to fifteen years, the discipline of lexicography has changed almost beyond recognition. This change is due to the technological revolution which has computerized the lexicographers' working environment to a very high degree and which has permitted a veritable quantum leap in the amount and variety of resources that can be brought to bear on the lexicographical process. The most important of these resources are computerized corpora of real, mostly written, but now increasingly also spoken, running text. When the first entirely *corpus-based dictionary* – COBUILD – came out in 1987, it was on the basis of a corpus of around 20 million words of connected text. Now all major British dictionary publishers use corpora of at least one hundred million words of text. Harrap/Chambers, Longman, and Oxford University Press have built the 100 million word British National Corpus (BNC), HarperCollins has the 200 million-plus word Cobuild Bank of English (BoE), and Cambridge University Press has compiled the 100 million word Cambridge Language Survey corpus (CLS).

Significantly, in 1995 all the major British publishers mentioned above have produced new (or completely revised new editions of) general-purpose monolingual dictionaries. Not surprisingly, they all shared the view that lexicography without computerized corpus data is practically unthinkable nowadays. The differences between them have to do with different views on corpus composition, different emphases and interests in the way corpus data are explored, differences in the way that the results are used in the dictionary-making process, etc.

### *1.1. The Development of Corpus Linguistics*

Before the rise of generative transformational grammar in the sixties, the systematic collection of illustrative data was established linguistic practice.

Traditional grammarians like O.Jespersen and H.Poutsma were constantly on the lookout for illustrative examples of language use. In fact, they would not include a rule of syntax in their grammars until and unless they could adduce a number of excerpts from actual use to support it. Most traditional grammarians were not corpus linguists in the modern sense of the word, though. They did not collect texts for their own sake, but only as sources for appropriate quotations, which were then taken out of context to support some point of grammar. This is similar to how data were collected for major dictionaries ever since Johnson's Dictionary (1755): Citations from authentic sources were painstakingly collected by dictionary compilers and their associates, the most elaborate project of this kind leading to the production of the OED in the last quarter of the nineteenth century and the first quarter of the twentieth. But, again, the process was basically one of looking for examples, with some fairly clear predetermined ideas about what to look for, and then lifting them out of context for inclusion in the dictionary. It was not until the Structuralist era that corpus data began to be used systematically to generate quantitative, statistical evidence. Frequency lists such as *Thorndike and Lorge* (1944) in the United States and *Michael West's General Service List of English Words* (1953) in the U.K. were both derived from manually compiled corpora.

Interestingly, modern computerized *corpus linguistics* began in an era when mainstream linguistics showed no interest in performance data at all. While in early sixties most linguists were engrossed in the theoretical (competence) abstractions of generative transformational grammar, Nelson Francis and Henry Kucera (both at Brown University) began compiling the first computer corpus of American English, the "Brown Corpus", a collection of about one million words in the form of 500 stretches of written text, approximately 2000 words each. By the end of the 1960s, Randolph Quirk had set up a collection of transcribed spoken (British) English texts, the Survey of Modern English Educated Usage, which was later transformed into a computerized corpus—the "London-Lund Corpus." The 1970s saw the compilation and computerization of the Lancaster-Oslo-Bergen (LOB) Corpus as the British

English counterpart of the Brown Corpus. Most of the work on these "first-generation" computer corpora of a million words or less was performed under the aegis of ICAME, the International Computer Archive of Modern English, through projects reported on during its annual conferences. Much of this work, centered in Britain, Holland, and the Scandinavian countries, was largely within what one might call the "mainstream" British linguistic tradition, taking its orientation from grammars such as that of R.Quirk, et al. (1985), hence a framework that is now often labeled "symbolic" and "rule-based."

While compiling a corpus of a million words was felt to be a major feat in the 1960s and 1970s, the enormous expansion of processing and storage capacity, the large number of texts being produced in electronic form, and the development of ways to "capture" texts and convert them to electronic format (scanning in conjunction with OCR, *Optical Character Recognition*) has fostered an exponential growth in corpus size in the past decade and a half. Nowadays, corpora of hundreds of millions of words are the norm rather than the exception. But large corpora bring their own problems: With this enormous expansion of corpus size comes a need for software that can explore and exploit such resources adequately. This has led to the development of probabilistic, statistically-based approaches to supplement or supplant rule-based ones.

### *1.2. The objectives of Corpus Linguistics*

*Corpus Linguistics* can be defined like the study of language on the basis of textual or acoustic corpora, always involving computer at some phase of storage, processing, and analysis of this data.

Textual corpora usually refer to the written aspect. Acoustic corpora refer to the research of spoken language with application to speech technology. Since the computer is involved, Corpus Linguistics is concerned not only with the analysis and interpretation of language, but also with computational techniques and methodology for the analysis of these texts.

The *main task* of Corpus Linguistics is, thus, the creation of machine-readable corpora and involving associative computational techniques as the basis for linguistic investigation. So, the generally accepted name for this science is now Computer Corpus Linguistics (CCL).

CCL focuses its attention on:

1) linguistic performance, rather than competence;
2) quantitative, as well as qualitative models of language;
3) linguistic description rather than linguistic universals;
4) more empiricist, rather than a rationalist view of scientific investigation.

For the foreseeable future, CCL projects will tend to be concerned with analysis and processing vast amounts of textual data, because larger quantities of texts are needed in order to build probabilistic systems for NLP. In 1960s 'large' meant the collection of a million or so words of text. In the future it is likely to mean hundreds and thousands of millions words.

However, the stress on size or quantity for a corpus does not necessarily mean that all types of computer corpora must be large, since there are some genres of texts restricted in scope or size. For instance, the corpus of Old English texts can never be of hundred million words, simply because it is restricted by the set of texts which have survived from the Old English period.

It has been suggested that the guiding principle for calling some collection of machine-readable texts 'a corpus' is that it should be designed or required for a particular *'representative' function*.

A corpus can be designed to serve as a resource for general purposes, or for a more specialized function such as being the resource which is representative of a particular sublanguage (roughly equivalent to a language genre).

B.T.S.Atkins distinguishes 4 types of *text collection* in Corpus Linguistics:

1) *Archive* – a repository of MR electronic texts, not linked in any coordinated way (the Oxford Text Archive).

2) *Electronic Text Library* – a collection of electronic texts in standardized format with certain conventions relating to content, but without rigorous selectional constraints.
3) *Corpus* – a subset of an ETL, built according to explicit criteria for a specific purpose (the Cobuild Corpus, the Oxford Pilot Corpus);
4) *Subcorpus* – a subset of a corpus, a static component of a complex corpus or a dynamic selection from a corpus during on-line analysis.

The *methodology* of CCL can be regarded like quantificational analysis of language that uses corpora as the basis from which the adequate language models may be built.

The term '*language model*' is typically associated with notions like probabilistic part-of-speech taggers and parsers. A tagger assigns syntactic categories to lexical items. Thus, the output of such program can be used to annotate a word-list with part-of-speech labels. A parse tree would represent the subcategorization information.

Such view of language analysis and processing involves a methodology for the derivation of lexical information from corpus processing and storage of this information in a permanent lexical structure, i.e. a suitable lexical database.

The *main functions* of corpus databases are:

1) frequency based account of word-distribution patterns;
2) concordance-driven definition of context and word behavior;
3) extracting and representing word collocations;
4) acquisition of lexical semantics of verbs from sentence frames;
5) derivation of lexicons for machine translation.

*The TEI project.*

The common focus on the use of computer to analyze texts by Computational Lexicography has led to the establishment of *Text Encoding Initiative* (TEI) which

had its origin in 1987 and was found by the Association for Computers in the Humanities and the Association for the Computational Linguistics.

The TEI has its tasks the production of a set of guidelines to achieve the adequate interchange of existing encoded texts and the creation of newly-encoded texts. The guidelines are meant to specify both the types of features that should be encoded, as well as to suggest ways of describing the encoding scheme and its relationship with pre-existing schemes. The development of such text-encoding standards opens up the possibility of encoding extra layers of information – this means entire categories of information that can be searched for automatically.

The main *TEI advantages* are as following:

1) Standardized descriptive-structural markup, by means of the Standard Generalized Markup Language (SGML) offers strategic advantages over procedural document markup by separating text structure and content from textual appearance.
2) Such documentation encoding forms the basis for a wide range document interchange and text processing operations common to publishing, database management, and office automation.
3) TEI/SGML encoding renders textual data accessible both for traditional printing demands and to electronic search and retrieval.
4) TEI encoding supports language-specific text processing within multilingual dimension of research documents and databases.

While the range of information in any lexicon depends on the purpose for which it has been built, the list of lexical information proposed by TEI Guidelines should contain all the possible types of information that can be considered for inclusion in a computational lexicon. The TEI Guidelines contain the base tag set for encoding human-oriented monolingual and polyglot dictionaries.

It should be mentioned that there are other related attempts to identify and promote the reusability of lexical information for machine readable dictionaries (MRDs), lexicons and corpora.

These include the following projects, the first five of which are based in the European Community: ESPRIT ACQUILEX, EUROTRA-7, GENELEX, MULTILEX, The EUROPEAN CORPORA NETWORK, The CONSORTIUM FOR LEXICAL RESEARCH AT NEW MEXICO.

The European Commission has also accepted three projects coordinated by the Institute of Computational Linguistics at the University of Pisa – RELATOR and PAROLE. These projects aim to create, manage, specify standards and distribute such linguistic resources as lexicons.

*1.3. Types of Lexicographic Evidence*

How is lexical (for the formulation of linguistic theory) or lexicographic (for dictionary-making) evidence derived? An obvious way is first to consult and rely on a dictionary. A dictionary, in its turn, bases its evidence on one of three methods of gathering evidence:

1) lexical introspection;
2) casual citation;
3) corpus method.

The method of *lexical introspection* exists since the time the first dictionary was created. If we rely on this method alone, the dictionary is only as good as its lexicographer, because it is based on the subjective introspection of the lexicon. The average person's lexicon is, firstly, finite and, secondly, static and unchanged (i.e. in need of updating). So it is not sufficient to rely upon only one's linguistic intuition. The reliance on such lexical intuitions in making dictionaries is often called 'armchair lexicography'.

The method of *casual citation* is offered when the lexical behavior of the members of society is observed and recorded. This can be done on the examples of the analysis of the family lexicons, lexicons of social groups and professional lexicons.

*Corpus method* was offered in 1995 edition of LDOCE dictionary. The essential difference between corpus and citational data is that, although both are instances of

observed data, only the corpus is systematically gathered for a particular purpose, and is coherently organized for this purpose.

The criterion for distinguishing between an archive and a corpus is also one of systematicity. An archive is a repository of available language materials. A corpus is a systematic selection and collection of material for given purposes. A corpus draws upon the resources of an archive. The corpus, if it possesses a representative function, will indicate the collective intuition of a relevant group of people using the word or linguistic expression under study.

Thus, if one wants to see whether writers of computer manuals tend to use the expression computer error in preference to computer mistake, then one should use a corpus of computer manuals and not a cookbook of recipes as evidence. This corpus is sufficiently contemporary and maximally representative.

If we concern the terms *handphone, mobile phone, cellular phone*, the term *handphone* is regarded like a core English term, ant it means that it is appropriate to both educated British and American speakers. But if we stick to the COBUILD 'Bank of English', it appears that out of 323 million words there are 2479 occurrences of *mobile phone*, 447 occurrences of *cellular phone* and only one occurrence of *handphone*. The sentence with *handphone* was taken from the Australian newspaper component of the Bank of English, which therefore indicates that this word is not transparent to contemporary English.

It is important to note that the reliance on corpus data does not mean a denial of one of the three methods for gathering the lexical or lexicographic evidence mentioned above. It is often necessary to utilize all three methods for treating the evidence adequately.

V.van Ooi (1997) says that any lexical enterprise using corpus data can stick to two main approaches, both of which are equally respective. They are:

1) *corpus-based linguistics*;
2) *corpus-driven linguistics*.

The difference in approaches is represented in the below (the scheme borrowed from V.B.Y. Ooi 1998):

| *Corpus-based linguistics* | *Corpus-driven linguistics* |
|---|---|
| A corpus is used to validate, check and improve linguistic observations that have already been made. | A corpus is of primary importance in bringing out new ideas for the examination of data. |
| The linguist does not question received theoretical positions or established categories. His position to language structure is already formed. | The linguist understands that the kind of evidence emerging from corpora may be difficult to reconcile with the established positions, and he leaves an open space for some changes in linguistic theory in order to cope with the evidence. |
| The corpus is used to extend and improve linguistic description. | The evidence from the corpus is paramount, therefore the linguist makes some assumptions about the nature of theoretical and descriptive categories. |
| An example of a relevant question is "Is the lexical item still used in English, if so, then how?" | An example of a relevant question is "Is the distinction between grammar and lexis possible?" |

The distinction between the corpus-based and corpus-driven approaches corresponds to the '*top-down*' and '*bottom-up'* approaches to the analysis of lexical data. A *top-down approach* begins with some theory, which is then applied to some data for confirmation, extension or rejection. A *bottom-up approach* begins with some data, whose analysis leads to the formulation of the theory. In practice, a mixed 'top-down' and 'bottom-up' approach is often necessary.

*1.4. Corpus as a lexical resource*

The dictionaries can not be regarded like the only viable lexical resources. It was suggested that the corpus should be considered as an alternative for the construction

of the lexicons, especially those suitable for NLP systems. Primarily this is because idiosyncrasies exist in most available MRDs. However, the use of the corpus as the alternative to MRDs is not without its problems. Basically there are two main procedures associated the use of a corpus as a lexical resource: corpus building and corpus utilization.

A corpus built wrongly or inadequately runs the risk of generating not only some faults in the information acquired but not offering any information at all. The use of wrong or inappropriate computational techniques for corpus utilizing runs the risk of generating false or incomplete results. Therefore, the methodology of corpus building depends on how well-representative or well-balanced it is for the language it represents.

*Representativeness* may be defined as the ability of the lexicon to refer to the lexical items reliable for the use with the definitely given purpose. Unless the corpus is representative, it can not be regarded like an adequate means for acquiring lexical knowledge. A true corpus is one which reveals the general core of the language to a broad range of documents types.

A *representative corpus* promotes the generation of reliable frequency statistics. It is commonly known that different corpora will present to the lexicographer different frequencies for words, so there is a need to moderate statistics with common sense.

A related notion to a representative corpus is a *balanced corpus*. B.T.S.Atkins defines it as a corpus so well organized that it offers a model of linguistic material which the corpus builders wish to study.

In 1993 D.Biber sets out a range of principles for achieving representativeness. The criterion of variability for determining variability basically consists of two main parameters of acquiring text: genre/register and text type.

*Genre* is a situationally defined text category (e.g. fiction, sports broadcasts, psychology article). *Text type* is a linguistically defined category (e.g. the distribution of third person pronouns to Present Indefinite tense, '*wh*' relative clauses.

The genre is primary to the text type, since the first is based on criteria external to the corpus which need to be determined on a theoretical ground.

Registers are based on different situations, purposes and functions of the text in a speech community. In contrast, identification of the text type in a language requires a representative corpus of text for analysis. The procedure of compiling texts should take into account:

1) the identification of situational parameters that distinguish text in language and in a culture;
2) the identification of the range of important linguistic features that will be analyzed in the corpus.

While considering the representativeness of a particular corpus, it is helpful to distinguish a *general purpose corpus* from one designed for a more specialized function.

The process of *compiling a representative corpus* is not linear. It seems to function more in cyclical manner, involving the following stages:

1) A pilot corpus should be compiled first, representing a relatively broad range of variations but also representing depth in some registers.
2) Grammatical tagging should be carried out as a basis for empirical investigations.
3) The empirical research should be carried out on this pilot corpus to confirm or modify various design parameters.
4) Parts of this circle should be considered in some continuous manner, with new texts being analyzed as they become available.
5) There should be a set of discrete stages of empirical investigation and revision of the corpus design.

The term '*empirical investigation'* means the use of statistical techniques such as factor and cluster analysis for the analysis of linguistic text variations.

Thus, now that complete yearly issues of newspapers can easily be obtained on CD-ROM, each year containing on average some 33 million words, it would be quite feasible to compile a corpus of over a hundred million words by combining just three

such CD-ROMs. However, that would clearly be a rather skewed corpus, not suitable as the basis for a general-purpose dictionary. D.Summers (1993) explains in detail the various criteria that were used in the compilation of the *Longman Lancaster Corpus*, most of which were taken over in the design of the *British National Corpus (BNC)*. Basically, the aim was to reach a balanced mix of text types such that the result was felt to cover "typical and central aspects of the language, and provide enough occurrences of words and phrases for the lexicographers". Such a corpus involves a variety of broad subject areas (natural and pure science, applied science, world affairs, leisure, etc.) and varies in terms of external factors such as region (British, American, other), date (pre-1950, 1960s and 1970s, post-1970s), level (technical, lay, popular), as well as varying by internal text-type features. The spoken part of the BNC (10 million words) was also carefully designed: Half of it was collected through a demographic survey set up to reach a representative balance in terms of regional and social distribution, age, level of (in)formality, etc. The other half was designed to produce varied data in terms of speech context, including dialogues and monologues in general areas such as education (e.g., lectures, classroom interactions), business (e.g., trade union speeches, sales talks), and leisure (sports commentaries, phone-ins, etc.).

We have to mention here that the computational techniques available for studying machine-readable corpora are at present rather primitive. There is a lack of interactive software which would be able to support the human enterprise for lexical analysis. The main tool that exists now is a concordance program (which is basically a keyword-in-context index with the ability of extending the context) is still very labour-intensive and would work well only if there are no less than a dozen of concordance lines for a word and just two or three main sense divisions.

Human’s mind immediately discovers all the significant patterns, separates lexical groups and rank them in order of importance. So, the most important things at this stage are tagging and skeleton parsing which help the concordances to be more linguistically selective.

As fare as corpus is believed to be representative, there arises a question of *sampling size*. Typically researchers focus on sample size as the most important condition for achieving representativeness. The question stands how many texts must be included in the corpus, and how many words per text sample.

Though, of course, the size is not the only important consideration. A thorough definition of the target population and the choice of sampling methods are also very much essential.

There is also a method of *monitor corpus analysis* which is used to monitor the occurrence of new words as a result of changes in the word senses, as well as extending the general scope of the language. A corpus, like a dictionary, is an account of a language at a certain point of time, and therefore may need to be continually updated reflecting new changes and new patterns of usage.

For such an enterprise, size is the most important consideration. It was observed that even with phrases involving frequent words, each additional word in a phrase requires an order of magnitude raised in the corpus to secure enough instances. Roughly speaking, if 1 million words is sufficient for showing the patterns of an ordinary single word (to fit), then 10 million words will be needed for showing new patterns of selection for the phrasal verb (to fit into), and 100 million words for a three word phrase (fit into place).

A very large corpus is needed for significant phraseological patterns to appear (including very frequent collocations and idiomatic expressions).

An example of a such-like corpus is the COBUILD *'Bank of English'*, which grew from 20 million words in 1987 to 211 million words in 1995. And currently it comprises 323 million words and, thus, is the largest single English database in the world.

The corpus comprises evidences from mainly British (225 million words) and American (65 million words) sources, and also Australian newspapers (33million words). The texts range widely from spoken to written, from newspapers and books to transcribed talk. The Bank of English is organized to provide representative, current spoken and written language by native speakers from around the world.

*Corpus composition*

As corpora have grown in size, the need for structuring and enriching them has naturally grown too. There is little point in compiling a carefully balanced corpus without at the same time creating the means to utilize its internal composition so that the lexicographer can, for instance, restrict searches to specific subcorpora (e.g., subsets of texts such as language learners' texts or popular scientific texts). Most of the corpora used by the large dictionary-makers have therefore been structured using SGML (Standard Generalized Markup Language) or similar coding so that information about the source texts (author, date, publication medium, etc.) is systematically registered and retrievable. In addition, they have been enriched with various types of markups, notably part-of-speech (POS) tagging and, in some cases, forms of syntactic marking showing larger constituents. There have also been some initiatives to provide corpora with semantic information ("sense-tagging"): These latter initiatives will be discussed in the section on current issues.

Essentially, POS tagging of corpus data no longer presents a major problem. Most systems in use nowadays are based on stochastic trigram methods and boast around 95 to 97 percent accuracy; that is about the maximum that one could realistically hope to attain, and it is probably sufficient for most lexicographical applications. The point is that, apart from the fact that sheer size of corpora makes manual POS tagging (with its proneness to human error, inconsistency, and fatigue) virtually impossible, there will always be a margin of indeterminacy since certain words will simply not fit easily into any pre-determined category. POS tagging is important for lexicography since it allows one, inter alia, to distinguish between word forms that can function in different part-of-speech categories – a very common phenomenon in English, as in many other languages. With this information, concordance searches can be restricted or narrowed down to, for example, only verbal or nominal uses of a given word form (matter as a verb vs. matter as a noun for instance). Also, accurate POS tagging enables accurate lemmatization and therefore

allows different forms of the same paradigm to be inspected together, yielding more reliable frequency information, etc.

*Parsing* (i.e., full syntactic analysis) is a rather different matter, however. To date, no system of analysis exists which can automatically provide a truly reliable, complete syntactic parse. Systems like *The Helsinki One* (A.Voutilainen and J.Heikkila 1994) can indicate most major constituent boundaries, but they fall short of providing a full syntactic analysis. Similarly, the system employed in the *Perm Treebank* yields major (categorial rather than functional) constituent boundaries, but the output is again a partial parse with quite a few indeterminacies left in. In any case, for most lexicographical applications to date, fully parsed, as opposed to POS tagged, material is not really essential. This situation, however, may be a matter of expediency: As long as dependable parsing systems are not available, lexicographic exploitation (for instance in the realm of refined subcategorization data, selection restrictions, etc.) will just not be forthcoming.

*Use of corpus data*

Good lexicographical practice has always included assessment of empirical data – real language in the form of citations from various sources deemed authentic and, usually, in the prescriptive tradition prevalent until halfway through the twentieth century, authoritative as well. Clearly, at some stage in the process which leads from data-gathering and data-sifting to the production of the actual dictionary entries, the lexicographer's informed judgment – her/his introspection, linguistic intuition, crafts (wo)manship, etc.—takes over. One could say that the use of computerized corpus data on a massive scale helps lexicographers postpone this moment at which intuitive judgment takes over, not because there is anything wrong with intuition, but because the data can play a more prominent role, allowing the judgment to be as informed and as unbiased as possible. Corpus-based lexicography requires an open mind, a readiness to discover patterns in the data that may go against one's preconceptions. The lexicographer who can accept challenges to his/her introspective expectations may suddenly realize that it is nearly always "bad things" that break out, or that the

phrase a slice of life seems to occur almost exclusively in film reviews. Apart from the more standard uses of interview which introspection might throw up by itself, corpus-data may direct you also to other distinct uses which you might not otherwise have thought of, such as the combination police interview as a euphemistic alternative, perhaps, to the more sinister phrase police interrogation. B.T.S.Atkins and B.Levin (1995) give a vivid description of how detailed inspection, assessment, and comparison of corpus data, combined with sound lexicographical and linguistic considerations, can throw light on the ways in which near-synonyms may subtly differ in meaning from each other.

Lexicographical applications of corpus data include the five following areas:

1. Providing real-life material, for instance, for use as examples: Views vary on the use of examples straight from the corpus. While Cobuild Publications insist on using authentic corpus sentences to serve as examples in their dictionary entries (with only minor alterations to reduce their length, etc.), most other dictionary-makers take a less strict view, often using corpus sentences as inspiration for the example sentences in their dictionaries.

2. Helping lexicographers decide on sense distinctions to be made: Viewing different occurrences of the same word or lemma in the form of on-screen concordance lines (i.e., with a bit of the surrounding contexts and with the possibility of looking up the wider original contexts in the corpus) can help lexicographers confirm or revise their ideas about the senses to be distinguished, suggest new ones, etc.

3. Providing information on grammatical patterns, subcategorization, registers, etc. The kinds of constructions in which an item typically occurs may help the lexicographer to describe its grammatical behavior, while the type of texts in which a word tends to occur may suggest specific characteristics in terms of register, style, (in)formality, etc.

4. Providing frequency information: The frequency of occurrence of a given item, differentiated with regard to senses, part-of-speech status, and

even inflectional form may help lexicographers in deciding in what order to list senses, whether to list an inflected form as a separate entry (e.g., amusing as an adjective as well as the ing-form of amuse), etc.

5. Providing information on new words, new combinations of words, and collocations: Totally new words are rare. Usually they are produced by standard word-formation processes such as derivation and compounding. The ways in which words tend to co-occur in more or less fixed combinations is increasingly felt to be highly important for a proper understanding of how they really function), and in recent years a great deal of effort has gone into attempts to develop software that can trace collocations adequately.

In addition to these uses which are relevant for monolingual and bilingual dictionary-making alike, bilingual lexicography is increasingly making use of machine-readable bodies of text accompanied by their translations. In order to make such "parallel" corpora maximally useful for lexicographical and other exploitation, they need to be "aligned" so that it is clear how the words or phrases in one text correspond to those in the corresponding translation text.

Finally, it should be mentioned that the availability of large and varied corpus resources has also been inspirational in the development of entirely new types of dictionaries, like the *Longman Language Activator*, advertised as "the world's first production dictionary," organized on the basis of approximately a thousand basic concepts.

*The lexicographer's workbench*

Computers were already used in the production of dictionaries before corpus data entered the lexicographical scene. The computer played a central role in the production of the first edition of the Longman Dictionary of Contemporary English (LDOCE1), for instance, to check consistent use of the controlled definition vocabulary of some 2000 words. The LDOCE1 was also one of the first structured machine-readable dictionaries, and it has been used very extensively in lexicological natural language processing research. The "bird's eye" view of the

dictionary, which this kind of research allowed, greatly enhanced linguists' and lexicologists' insight into how dictionaries "hang together" (or fail to do so) with regard to aspects that would otherwise be hard to trace, for instance, their inherent or implied semantic. This development, in turn, has influenced the way new dictionaries have been made.

The basic tool for lexicographic corpus data, now as before, is still the KWIC (key-word-in-context) concordance, which provides centered, on-screen display of occurrences of a given word form (all of them or some selection of them), with a stretch of the original context on either side, often with additional options such as alphabetical sorting on the first word to the right or left of the keyword. The advantage of this device is obvious. The lexicographer can inspect a wide range of attested occurrences of the item s/he is interested in; the bordering words will give some indication as to the typical context in which the item occurs. The concordance may also suggest specific words or phrases that the keyword tends to collocate with, typical subcategorization patterns, grammatical behavior, selection restrictions, and the like. Also, the KWIC concordance display may help the lexicographer confirm or revise presumed sense distinctions or discover new ones.

In addition to the concordancing software, lexicographers can consult or interactively generate various kinds of lists and databases, including the following:

- frequency lists of word forms or lemmas in the corpus (or a selected subcorpus);
- sorted alphabetically or by frequency;
- lists of word combinations sorted by various probability measures;
- lists of items that need special attention for one reason or another, for instance, entries from a previous edition that are felt to need revision.

Naturally these supports mesh in with software that helps lexicographers in the actual construction of the dictionary, allowing them, for instance, to "cut and paste" examples straight from the corpus into the appropriate section of an entry, screen definitions for compatibility with a controlling restricted vocabulary, or cross-check definitions with definitions for related words. In its most sophisticated form, the

lexicographer's workstation is an integrated computer system in which the lexicographer can switch between all the components – the corpus, concordancing, retrieval and statistical software, and the software which guides the gradual compilation of the dictionary – at the click of a mouse button.

A detailed account of the first project in which a dictionary was developed on the basis of computer corpus data (the Cobuild project) can be found in the work *Corpus, Concordance, Collocation* by J.Sinclair (1991), including a description of the tools that the lexicographers had at their disposal.

## 2. Corpus-based investigations of language use.

*The Refinement of Language Statistics*

The use of large, balanced corpora has made it possible for the first time to include reasonably reliable frequency information about individual words in dictionaries. While the frequency indications in COBUILD were still fairly coarse, both LDOCE and COBUILD, building as they do on much larger corpora, now have more detailed frequency data. LDOCE provides more refined information, separating out spoken and written as well as British and American uses, and giving separate frequencies where words can function in different parts of speech and different grammatical patterns; this is done for the 3,000 most frequent words only. COBUILD lacks most of these refinements, but its frequency ranges cover some 14,000 word types.

While the statistics for individual words are reasonably straightforward, those for combinations of words are less so. The notion of "mutual information" of a collocation (MI-score) as a measure of the likelihood of words occurring together, given their individual frequencies in the corpus, appears to be really important for the corpus analysis. For collocations, this probability measure often does not give good results. Cambridge University Press likewise uses a measure, the I-score, which

gives a more efficient weighting to the more frequent combinations, in preference to the MI-score.

*(Semi-)automatic data selection*

Adequate statistics are clearly very important for the question of (semi-) automatic data-selection. With the enormous expansion of corpus size in recent years, the sheer volume of data could become a hindrance. Working through hundreds or thousands of concordance lines for one and the same word or lemma may in fact blunt rather than sharpen the lexicographer's awareness. Research by Alex Collier, based on the Cobuild Bank of English (BoE) and the lexicographers working directly on the Cobuild Bank, shows that a thousand instances per item is about the limit of what human users can cope with, and that, alarmingly, over 93 % of the tokens in the BoE (which now stands at 230 million words) is made up of word types represented by more than a thousand tokens each. What is needed, therefore, is software that will prevent lexicographers from being swamped by the data and that, instead, by sensible pre-selection, focuses their attention on what is significant and worth their attention. This issue has not yet been satisfactorily resolved. There are currently three possible practices for limiting the number of occurrences: One simple, but very simplistic, method would be to look at the first two hundred (or five hundred, or one thousand) occurrences and ignore the rest. Obviously that defeats the purpose of using a large corpus in the first place. Two other methods that have been proposed (and versions of which are actually being used) are to let the system produce a random sample of the total occurrences, or to let it generate a sample by selecting every nth occurrence, in both cases with a pre-set upper limit. However, both of these methods may lead to important cases being left out, again in a way undermining the idea of using a large corpus.

*Sense distinctions and sense-marking*

More ambitious approaches to overcoming the "information overflow" problem in large corpora are based on the assumption that the key to a word's meaning and

function is the company it keeps – the kinds of words and structures that typically surround it. Such approaches therefore attempt to determine (semi-) automatically the different senses of a word and, on the basis of that analysis, present the lexicographer with a representative sample for each putative sense. Since distinguishing and adequately defining different senses is one of the most basic tasks confronting lexicographers, software that would reliably lend itself to data arranged or otherwise marked in terms of different senses would be most welcome. Applying the special framework developed for tracing cohesion in text also helps to produce sensible automatic pre-selections for the word types that are represented by more than a thousand tokens. Collocations (in a fairly loose sense) are very good pointers to a word being used in different senses, illustrating this with the noun bow. Bow will tend to collocate with different words depending on the sense involved: tie, tied, etc. in the sense of "type of knot"; arrow, string in the sense of "weapon"; stern, wave, starboard in the "front part of ship" sense, etc. The option of sorting KWIC output on the basis of the words occurring immediately to the left or the right of the key-word already helps a great deal to show up frequent word combinations. However, collocations may be further apart and show a good deal of flexibility in terms of intervening elements. There was suggested a methodology that matches the contexts of good collocates for a particular sense against that of the key-word as used in the relevant sense. This can lead to the detection of a large range of words associated with the key-word in a particular sense, which can then be used to select just those concordance lines that instantiate that sense. This technique was suggested by J.Clear in his work *The British National Corpus* (1993).

B.T.S.Atkins (1994) describes another approach, developed in the *Hector Project* (sponsored by Digital Electronic Corporation [DEC] and Oxford University Press), which constituted an attempt to arrive at semi-automatic sense-tagging on the basis of prior computer-aided but largely manual marking, again on the basis of KWIC data. P.Procter and Cambridge University Press take a different view. Their long-term aim, starting with the CIDE (Cambridge International Dictionary of English) and the associated Cambridge Language Survey (CLS) corpus, is to develop

automatic sense-tagging on the basis of a weighted assessment of a whole range of indicators such as domain markers, collocations, syntactic patterns, selection restrictions, frequency data, and even punctuation. It is hoped that, as the distribution and the accuracy of the various types of indicators increases, the number of words in the corpus can be cyclically expanded.

The benefits of the various software approaches discussed above may help to systematize and structure the wealth of data, and they may help to postpone the moment of informed introspection. In the final analysis, however, it is still the lexicographer who has to decide whether the information derived makes enough "sense" 'to count as one' rather than 'to indicate important distinctions.'

On a basic level, there are two main areas of study within linguistics: language structure and language use. Language practitioners as well as theoreticians must be concerned with both areas; that is, they need a full understanding of the structural resources available in a language as well as analyses of what speakers and writers actually do with those resources. Investigations of a representative text corpus – a principled collection of texts stored on computer – provide important insights into both of these domains and open new avenues of inquiry.

Language structure is traditionally described using non-empirical methods, relying on the analyst's intuitions. More recently, these descriptions have been complemented by corpus-based investigations of structure which provide authentic examples and identify structures that had previously been disregarded as unimportant or ungrammatical. However, it is in the area of language use that corpus-based techniques have had the most impact. Studies of use are concerned with actual practice, and the extent to which linguistic patterns are common or rare, rather than with potential grammaticality.

Over the past decade, there has been a marked increase in corpus-based studies describing various aspects of language use. Over the last five years, corpus-based investigations have become even more common (e.g., K.Aijmer and B.Altenberg 1991, S.Armstrong 1994, L.Flowerdew and A.Tong 1994, U.Fries, G.Tottie,

C.Johansson and A-B.Stenstrom 1991, N.Oostdijk and P.de Haan 1994, D.Ross and D.Brink 1994, J.Svartvik 1990; 1992).

For example, in descriptive lexicography, which is concerned with the actual use of words, a corpus is an essential resource for documenting the range of meanings for a word; new meanings are discovered only by examining the use of a word in actual discourse contexts. Grammatical structures can also be compared from a use perspective by studying the ways in which seemingly similar structures occur in different contexts and serve different functions. In addition, a use perspective is required to investigate the stylistic preferences of individuals, the differing linguistic preferences of groups of speakers, and the ways in which 'registers' (or 'genres') favor some words and structures over others.

Adequate investigations of language use must be empirical, analyzing the actual distributional patterns in natural texts. Corpus-based analyses are particularly well suited to such investigations. *Essential characteristics* of a corpus-based approach include the following:

- it is based on empirical analysis of a large and principled collection of natural texts, known as a corpus;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative (interpretive) analytical techniques.

One major advantage of a corpus-based approach is that it can provide a scope and reliability of analysis not otherwise feasible. Even more importantly, corpus-based techniques enable investigations of new research questions that were previously disregarded because they were considered intractable. In particular, a corpus-based approach makes it possible to identify and analyze complex 'association patterns' – the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features.

Investigating the variability of a linguistic feature in terms of its association patterns has two major components: 1) non-linguistic associations, and 2) linguistic

associations. Non-linguistic association patterns describe how certain linguistic features are differentially associated with registers, dialects, or historical change. Linguistic association patterns include two main types: lexical associations and grammatical associations. Both individual words and grammatical constructions can be studied with respect to their association patterns.

For a corpus-based study of a word, the lexical associations are the collocations of the target word (other words that the target word frequently co-occurs with). The grammatical associations of the target word describe structural preferences; for example, the research might investigate whether a particular adjective typically occurs with attributive or predicative functions, or whether a particular verb typically occurs with transitive or intransitive functions.

There are two kinds of *association patterns*:

1. Investigating the variability of a linguistic feature (lexical or grammatical)

a) Non-linguistic associations of the feature:

- distribution across registers:
- distribution across dialects;
- distribution across time.

b) Linguistic associations of the feature:

- lexical associations: co-occurrence with particular words;
-grammatical associations: co-occurrence with grammatical features.

2. Investigating the variability among texts (in terms of 'dimensions' — co-occurrence patterns of linguistic features.

Corpus-based studies of a grammatical construction can similarly include both kinds of associations. In this case, the lexical associations are the tendencies for the target grammatical construction to co-occur with particular words. For example, what matrix-clause verbs typically occur with a *that*-clause, and what different set of matrix-clause verbs typically occur with *to*-clauses? Grammatical associations in this case identify contextual factors associated with structural variants. For example, the study might investigate whether *that*-clauses are used in extraposed constructions as often as *to*-clauses.

All of these linguistic association patterns interact with non-linguistic associations. In fact, corpus-based analyses show that linguistic association patterns are generally not valid for the language as a whole. Rather, linguistic and non-linguistic associations interact with one another so that strong linguistic associations in one register often represent only weak associations in other registers.

A final important type of association pattern is studied when the research goal is to describe texts and registers – the ways in which groups of linguistic features commonly co-occur in texts – rather than individual linguistic features. For example, nouns, adjectives, and prepositional phrases commonly co-occur in academic prose texts, working together to provide a dense integration of information. Textual co-occurrence patterns such as these are important in characterizing the salient linguistic characteristics of registers and styles.

The following sections provide example analyses of each type of association pattern: association patterns for individual words are illustrated in the second section; association patterns for grammatical constructions are illustrated in the third section; and register analyses with respect to textual co-occurrence patterns are illustrated in the final section.

*Corpus-based investigation of individual words.*

Over the past decade, lexicographers have come to rely on computer-based text corpora to study the meaning and use of individual words. When coupled with a concordancing program, a corpus provides a wealth of examples for any given word, allowing lexicographers to identify and characterize the range of meanings for the word more accurately. Statistical measurements of word associations can provide an efficient means of further clarifying the senses of words and their patterns of use.

A variety of corpus-based investigations are useful to Applied Linguists studying lexical issues. For example, corpus-based analyses have been used in the following three areas:

- to disambiguate the functions of multifunctional words;
- to investigate the distribution and use of closely related;

- to identify and characterize the use of relatively fixed lexical expressions.

One type of corpus-based investigation that is particularly important for applied linguistics research involves the study of seemingly synonymous words. In any language, dictionaries and thesauruses are based on extensive lists of words with similar meanings. However, corpus-based investigations of association patterns often show that there are important, patterned differences in the ways that native speakers use seemingly synonymous words – differences which are important both for understanding the functions of the words and for teaching their appropriate uses. The association patterns for two pairs of near-synonymous adjectives referring to size *big* versus *large*. and *little* versus *small* are represented below:

| | Conversation | Academic prose |
|---|---|---|
| *big* | ****** | - |
| *large* | - | ****** |
| *little* | ********* | *** |
| *small* | * | ***** |

(Each * marks approximately 100 per million words, - represents less than 50 occurrences per million words).

Corpus research has shown that these adjectives are used in strikingly different ways across registers: *big* and *little* are strongly preferred in conversation, while *large* and *small* are preferred in academic prose.

*Corpus-based investigations of grammatical constructions.*

Corpus-based analyses can also be used to investigate grammatical issues, addressing research questions such as the following:

- How is a grammatical construction used, and how are related constructions
  used differently?
- How rare or common are related constructions?
- Are constructions used more or less frequently in different registers?

- Are there particular words that a grammatical construction commonly co-occurs with?
- What factors in the discourse context are associated with the use of grammatical variants?

Most of the studies use corpora to analyze the influence of contextual factors on the distribution of structural variants. Both lexical and grammatical association patterns have been shown to be important. For example, C.Mair (1990) identifies a number of individual verbs that are particularly common with various infinitival constructions (e.g., the verbs *want, allow, enable, expect, get*, and *like* occurring commonly with subject-to-object raising). P.de Haan (1989) explores the association of relative clauses with head noun phrases having different grammatical roles.

To illustrate association patterns of this type, we briefly describe certain aspects of the grammar of complement clauses in English. The two most common types of complement clause are *that*-clauses and *to*-clauses. In some contexts, these two are similar in meaning. Thus compare:

*I hope that I can go.*

*I hope to go.*

However, corpus-based study shows that the actual use of these two structures is quite different. First, in terms of their overall distribution, *that*-clauses are very common in conversation but not so common in academic prose. In contrast, *to*-clauses are moderately common in both conversation and academic prose.

Here is the overall distribution of *that*-clauses and *to*-clauses in conversation and academic prose (each * represents 500 occurrences per million words)

| | Conversation | Academic prose |
|---|---|---|
| *that*-clauses | ********** | **** |
| *to*-clauses | ******** | ********* |

The difference in overall distribution noted above can be related in part to the differing lexical associations of the two types of complement clause. That is, while a few verbs can control both *that*-clauses and *to*-clauses (e.g., *hope, decide*, and *wish*),

most verbs control only one or the other type of complement clause. For example, the verbs *imagine, mention, suggest, conclude, guess* and *argue* can control a *that*-clause but not a *to*-clause; the verbs *begin, start, like, love, try* and *want* can control a *to*-clause, but not a *that*-clause.

*Corpus-based investigations of ESP and register variation*

Research on discourse and the linguistic characteristics of particular varieties of texts tends to be empirical, based on analysis of some collection of texts. In this regard, most discourse studies can be considered corpus-based. This trend in discourse analysis is seen both in ESP research in *applied linguistics* and in register variation research in sociolinguistics.

Within applied linguistics, numerous discourse studies focusing on specialized varieties have been carried out by researchers in ESP and EAP (*English for Specific/Academic Purposes*). Similarly, within descriptive and sociolinguistics, there is a long tradition of empirical research on 'registers,' 'genres,' and 'styles,' dating from the work of M.A.K.Halliday, J.Leech, D.Crystal, and others in the early 1960s. In recent years, most analysts studying registers have begun to use corpus-based techniques.

In addition to descriptions of a single register, a corpus-based approach enables comparative analyses of register variation. Using computational (semiautomatic techniques to analyze large text corpora, it is possible to investigate variation across a large number of registers with respect to a wide range of relevant linguistic characteristics. Such analyses provide an important foundation for work in ESP in that they characterize particular registers relative to the range of other registers, documenting the extent of linguistic differences across registers.

This approach may be illustrated using D.Biber's *multi-dimensional* (MD) *analytical approach* (1994). Research in this framework analyzes the distribution of linguistic features in a computer corpus to identify text-based association patterns – sets of linguistic features that tend to co-occur in texts. Each grouping of linguistic

features is referred to as a 'dimension.' Studies of this kind have shown that there are systematic patterns of variation among registers; that these patterns can be analyzed in terms of underlying 'dimensions' of variation; and that it is necessary to recognize the existence of a multidimensional space in order to capture the overall relations among registers.

The dimensions are identified from a quantitative analysis of the distribution of linguistic features in a representative corpus (using factor analysis). Six major dimensions are identified and interpreted by D.Biber. Each comprises a distinct set of co-occurring linguistic features; each defines a different set of similarities and differences among spoken and written registers; and each has distinct functional underpinnings.

Registers can be compared with respect to text-based association patterns by computing 'dimension scores'. Such comparisons have been used to investigate a number of linguistic research questions, including the relations among spoken and written registers in English and the historical evolution of written registers. Multi-dimensional comparisons have also been used to describe the linguistic characteristics of ESP/EAP registers.

When all six dimensions are considered, the differences among registers are even more notable. There is no single register that can be identified as 'general English,' and that advanced instruction based on our intuitions about 'general' or 'core' English is not likely to provide adequate exposure to the actual linguistic patterns found in the target registers that advanced students must use on a regular basis.

The *Multi-Dimensional analytical framework* has been applied to other domains as well to identify additional dimensions of variation. In English, a study of elementary student writing and speaking by R.Reppen (1994) has particularly important implications, since it identifies and interprets basic dimensions that characterize elementary students' spoken and written registers. Comparison of the adult and elementary-student multi-dimensional models provides a register perspective on the development of literacy skills. Multidimensional analyses have also been used to study register variation in other languages

## Conclusion

The insights gained from corpus-based investigations are important not simply for understanding language use, but also for designing effective teaching materials and activities. Because corpus-based studies focus on language use, their results provide valuable information on what to teach if we want students to use language appropriately, rather than focusing exclusively on grammatical accuracy. In addition, because corpus-based studies describe the linguistic features associated with different registers and situational contexts, they allow us better to tailor language instruction to students' specific needs so that we expose students to the different types of language that they actually encounter in everyday use. Thus, both for understanding how users exploit the resources of the language and for teaching language based on real communication and appropriate use, corpus-based research promises to be an increasingly valuable and productive resource for Applied Linguists.

# CHAPTER IV
# TECHNOLOGY AND LANGUAGE ANALYSIS

## 1. The Contribution of Lexicography

One of the major resources in the task of building a large-scale lexicon for a natural-language system is the machine-readable dictionary (MRD). Serious flaws (for the user-computer) have already been documented in dictionaries being used as machine-readable dictionaries in natural language processing, including a lack of systematicity in the lexicographers' treatment of linguistic facts; recurrent omission of explicit statements of essential facts; and variations in lexicographical decisions which, together with ambiguities within entries, militate against successful mapping of one dictionary onto another and hence against optimal extraction of linguistic facts.

Large-scale electronic corpora now allow us to evaluate a dictionary entry realistically by comparing it with evidence of how the word is used in the real world. For various lexical items, an attempt is made to compare the view of word meaning that a corpus offers with the way in which this is presented in the definitions of five dictionaries at present available in machine-readable form and being used in natural language processing (NLP) research; corpus evidence is shown to support apparently incompatible semantic descriptions. Suggestions are offered for the construction of a lexical database entry to facilitate the mapping of such apparently incompatible dictionary entries and the consequent maximization of useful facts extracted from these.

Writing a dictionary is a salutary and humbling experience. It makes you very aware of the extent of your ignorance in almost every field of human experience. It fills lexicographer’s working day with a series of monotonous, humdrum, fascinating, exasperating, frustrating, rewarding, and impossible tasks. And when it is all over, the fruits of this labour are enshrined forever in a form that allows other people to take it

(and you) apart, in print, publicly and permanently. Lexicographers should, therefore, be even more enthusiastic than the rest of the linguistic world at the prospect of large-scale lexicons for natural-language systems being built by semi-automatic means.

Machine-readable or not machine-readable, a dictionary is a dictionary. Most machine-readable dictionaries were person-readable dictionaries first. As every lexicographer will confirm, systematicity is high on our list of priorities: but higher still comes user-friendliness. If we had a choice between being completely consistent throughout a 2,000 page (18 million-character) dictionary - were it even possible - and making one line of one entry totally intelligible to the least motivated user, the user would win. Again, consider the time scale: such a dictionary will take at least five years, and can take fifteen to write. No lexicographical task is ever quite the same as the one just completed. There may be twenty, thirty, or forty (or more) lexicographers in the compiling team. However complex the editor's instructions and however conscientious the compilers, the entries in A and B will differ from those in X, Y, and Z by much more than their place in the alphabet. And this is, in human terms, just as it should be. A dictionary is a human artifact, designed to be used by human users. Until the advent of the computer, people took dictionaries in their stride. Their human brains compensated for a lack of systematicity throughout the work. They knew, albeit vaguely sometimes, more or less what words could - and did - do.

In the computer, however, we have the ultimate learner and one with a terrifying capacity for homing in on inconsistencies invisible to the naked eye. Serious flaws (for the user-computer) have already been documented in 'handheld' dictionaries - indeed, in the very dictionaries at present available and being used in machine-readable form. These include the omission of explicit statements of essential linguistic facts, lack of systematicity in the compiling in one single dictionary, ambiguities within entries, and incompatible compiling across dictionaries. However, these are in the main sins of omission rather than commission; they make it more difficult to extract information from the MRD but ultimately detract very little from the value of the information extracted.

The question at issue now is more fundamental: how much semantic information accurate enough to be useful in a computational lexicon is contained in a dictionary definition written for the human user, who often unconsciously supplements and corrects what is being read? Is it indeed possible to write dictionary definitions that encapsulate the essential facts about the senses of a word? Can the meaning of a word be divided into discrete senses without distorting it beyond reason? Large text corpora allow a detailed study of how a word is used, thus enabling us to evaluate the accuracy of dictionary entries much more objectively than before. Lexicographers worked with such corpora, and examined hundreds of individual citations minutely in an attempt to find objective evidence for the existence of dictionary senses.

The most widely known MRDs are: Collins English Dictionary (1986) (CED); Webster's New World Dictionary (1988) (WNWD); Oxford Advanced Learner's Dictionary (1989) (OALD); Longman Dictionary of Contemporary English (1987) (LDOCE); and Collins Cobuild English Language Dictionary (1987) (CCELD).

## 2. The Contribution of Linguistics

The *lexicon* has come to occupy an increasingly central place in a variety of current linguistic theories, and it is equally important to work in natural language processing. The lexicon - the repository of information about words - has often proved to be a bottleneck in the design of large-scale natural language systems, given the tremendous number of words in the English language, coupled with the constant coinage of new words and shifts in the meanings of existing words. For this reason, there has been growing interest recently in building large-scale lexical knowledge bases automatically, or even semi-automatically, taking various on-line resources such as machine readable dictionaries (MRDs) and text corpora as a starting point.

Although in principle on-line resources such as MRDs and text corpora would seem to provide a wealth of valuable linguistic information that could serve as a foundation for developing a lexical knowledge base, in practice it is often difficult to take full advantage of the information these existing resources contain. Dictionaries,

for example, might seem particularly well-suited as a basis for automatic lexicon construction, since the information they provide is structured within the entry, and it would seem possible to extract certain information, for example, part of speech, fairly trivially. However, this is only a fraction of the information available in a dictionary. Dictionaries are designed for human users by humans. Human users are native speakers of language who know at least implicitly how the lexicon of their language is structured, and lexicographers exploit the lexical knowledge of potential users in writing dictionary entries. Consequently, dictionary entries only need to say enough about a word to allow native speakers of a language to tap into their general knowledge. Thus entries often leave much implicit or unsaid, something that would be unacceptable in a lexical knowledge base for a natural language system. The missing information must be filled in from somewhere, and linguistic studies into lexical organization can contribute to this task. Even learner's dictionaries, which are intended for learners of a language, take advantage of general properties of language, although typically they do provide fuller information than dictionaries intended for native speakers of that language about syntactic properties, as well as a range of example sentences illustrating word use.

These considerations aside, the value of using dictionaries as a starting point for building a lexical knowledge base is diminished by the limitations of dictionary-making itself. Dictionaries are written by lexicographers, who are themselves humans working within rigorous time and space constraints. Consequently, not all words receive the attention they deserve. Even the best dictionaries have flaws; for instance, they are often incomplete and inconsistent. For instance, words that pattern in the same way are often not given parallel treatment in dictionaries, due either to time and space limitations or to the failure of the lexicographer to recognize the pattern. The results of linguistic research into lexical organization have implications for the design of a lexical knowledge base: they both suggest the overall structure of the knowledge base and delineate the type of information that must be available in this knowledge base. This framework in turn should facilitate the extraction of as much information as possible from on-line resources. Specifically, efforts to build lexical knowledge

bases automatically or semi-automatically could use template entries for verbs of particular semantic types motivated by linguistic research to guide attempts to extract information about specific verbs from existing on-line resources such as dictionaries and corpora.

'On-line dictionaries are unlikely to serve as a lexical knowledge base, even if, as suggested by some researchers, several dictionaries were merged on the assumption that the result will be more complete than any single dictionary. The process of merging dictionary entries faces many obstacles. Furthermore, there is no guarantee that the result of merging the entries for a given word would be an entry that is substantially better than the entries of individual dictionaries; such an entry is unlikely to approximate a linguistically motivated lexical knowledge base entry for that word.

Although dictionaries are a rich source of information about words, the information needed in dealing with problems of the type described here is often not explicitly signaled, if it is included at all. Most dictionaries indicate whether verbs have a transitive use, an intransitive use, or both, but relationships between transitive and intransitive uses of verbs such as *eat* and *dress* are not as a rule explicitly indicated. However, such relationships are often encoded using a variety of cues in the dictionary entry that involve the grammatical codes, the wording of the definitions, and properties of the example sentences. Thus although the relevant information can sometimes be found in a dictionary, it is not trivially accessible, but will require queries formulated in terms of the specific cues in dictionary entries, a problem complicated by the fact that the same cues are not used consistently across the entries of verbs that pattern in the same way.

As the *eat* / *dress* example illustrates, some verbs may express their arguments in more than one way, sometimes with slightly different semantic interpretations. Any natural language system that aims at substantial coverage of English must be able to handle correctly not only these but the entire range of possible relationships between alternate expressions of the arguments of verbs. The understanding of the lexical

organization of English verbs of the type that emerges from linguistic investigations can contribute to the realization of this goal.

Although the lexicon has been considered the domain of the idiosyncratic, there is much evidence that the relationship between the meaning of verbs and their syntactic behaviour is governed by quite general principles, with evidence coming from studies in both lexical semantics and syntax. The *eat/dress* example shows that certain verbs have both transitive and intransitive uses, and that the relationship between the uses is not uniform across all verbs.

However, such a relationship is not merely an idiosyncratic property of a verb; rather it is to a large extent predictable from the verb's meaning. Interchanges parallel to the one described for *eat* are possible with a wide range of verbs, including *type, sew, sweep,* and *read.* These verbs are all activity verbs; most of them describe typical occupations. Another set of verbs including *bathe, change, shave, shower*, and *wash* - all verbs of grooming or bodily care - behave like *dress*.

Linguists have extensively studied a wide range of linguistic phenomena involving the expression of the arguments of verbs, such as the alternations in transitivity exhibited by the verbs *eat* and *dress*. These studies reveal that English verbs are organized into classes on the basis of shared components of meaning. The members of these classes have in common a range of properties, specifically properties concerning the possible expression and interpretation of their arguments, as well as the extended meanings that they can manifest.

The long-term goal of much current linguistic research is explaining what a native speaker of a language knows about the lexical properties of verbs, focusing on those aspects of lexical knowledge related to argument structures, the semantic and syntactic properties of verbs tied to their status as argument-taking lexical items. A central concern of linguistic research on the lexicon is the study of the meanings of verbs and the elaboration of a theory of the representation of lexical entries in which the meaning of a verb is properly associated with the syntactic expressions of its arguments. Ideal lexical entries of verbs should embody the full range of linguistic knowledge possessed by an English speaker in relation to those verbs. At the same

time, however, any given entry should supply the minimum amount of information necessary to account for the native speaker's linguistic knowledge of it. This dual requirement naturally leads to the investigation of those aspects of the linguistic behavior of lexical items that are determined by general principles of grammar.

Currently, an important part of this research is the rigorous study of diathesis alternations, alternations in the expression of the arguments of verbs. As the discussion of the verbs *eat* and *dress* illustrates, since diathesis alternations reflect the interaction between a representation of the meaning of a verb and the principles that determine the syntactic realization of its arguments, they can be used to probe into both the lexical representation of meaning and the relationship between syntax and semantics. As the distinctive behaviour of verbs with respect to diathesis alternations arises from their lexical properties, specifically their meaning, the exploration of the ways in which diathesis alternations distinguish among verbs should reveal semantically coherent verb classes. Once identified, these classes can be examined to isolate the components of meaning common to verbs participating in particular alternations. These components of meaning would be expected to figure prominently in the lexical representation of the meaning of these verbs. Attempts to formulate the principles according to which these elements of meaning determine the syntactic behaviour of verbs then become possible.

For these reasons, the study of diathesis alternations can make a significant contribution to the elucidation of the lexical representation of meaning. These studies have established a range of diathesis alternations relevant to the lexical organization of English and have identified a number of essential semantically coherent classes of verbs, as well as the central properties characterizing verbs of each type. Nevertheless, much basic research remains to be done in this area.

## 3. The Contribution of Computational Lexicography

*Computational lexicography* is emerging now as a discipline in its own right. In the context of one of its primary goals - facilitation of (semi-)automatic construction

of lexical knowledge bases (e.g. computational lexicons) by extracting lexical data from on-line dictionaries - the concerns of dictionary analysis are related to those of lexical semantics.

The notion of *structured dictionary representation* is exemplified by looking at the wide range of functions encoded, both explicitly and implicitly, in the notations for dictionary entries. This allows the formulation of a framework for exploiting the lexical content of dictionary structure, in part encoded configurationally, for the purpose of streamlining the process of lexical acquisition.

A methodology for populating a lexical knowledge base with knowledge derived from existing lexical resources should not be in isolation from a theory of lexical semantics. Without a theoretical framework the traditional methods of computational lexicography can hardly go further than highlighting the inadequacies of current dictionaries. By reference to a theory that assumes a formal and rich model of the lexicon, dictionaries can be made to reveal - through guided analysis of highly structured isomorphs - a number of lexical semantic relations of relevance to natural language processing, which are only encoded implicitly and are distributed across the entire source.

One approach to scaling up the lexical components of natural language systems prototypes to enable them to handle realistic texts has been to turn to existing machine-readable forms of published dictionaries. On the assumption that they not only represent (trivially) a convenient source of words, but also contain (in a less obvious, and more interesting way) a significant amount of lexical data, recent research efforts have shown that automated procedures can be developed for extracting and formalizing explicitly available, as well as implicitly encoded, information - phonological, syntactic, and semantic - from machine-readable dictionaries (MRDs).

The appeal of using on-line dictionaries in the construction of formal computational lexicons is intuitively obvious: dictionaries contain information about words, and lexicons need such information. If automated procedures could be developed for extracting and formalizing lexical data, on a large scale, from existing on-line

resources, natural language processing (NLP) systems would have ways of capitalizing on much of the lexicographic effort embodied in the production of reference materials for human consumption. On the other hand, there are at least two classes of disadvantages to the use of MRDs in natural language processing. First, because these are produced with the human user in mind, there is a strong assumption about the nature of understanding and interpretation required to make use of a dictionary entry; second, due to the very nature of the process of (human) lexicography, present-day dictionaries are far from complete, consistent, and coherent, certainly with respect to virtually any of the numerous kinds of lexical data they choose to represent and encode. An important question then becomes: where is the line between useful and relevant data to be extracted from existing machine-readable sources, on the one hand, and the level of 'noise' (inconsistencies, mis-representations, omissions) inherent in such sources and detrimental to the enterprise of deriving computational lexicons by (semi-) automatic means, on the other?

A number of arguments have been put forward in support of a claim that, in effect, a dictionary is only as good as its worst (or least experienced) lexicographer - and by that token, it is not much good for developing systematic procedures for extraction of lexical data. For instance, in the process of giving a descriptive introduction to the discipline of computational lexicography, B.T.Atkins (1994) not only summarizes the process of building a large-scale lexicon as "trawling" a machine-readable dictionary in search for lexical facts, but points out an imbalance between the kinds of syntactic and semantic information that can be identified by "minutely examining" existing dictionaries: "the useful semantic information which may be extracted at present is more restricted in scope, and virtually limited to the construction of semantic taxonomies".

From the perspective of building formal systems capable of processing natural language texts, there is (currently) a much better understanding of the nature of the syntactic information required for implementing such systems than of its semantic counterpart. In other words, the state of the art of (applied) computational linguistics is such that syntactic analyzers are much better understood than semantic interpreters;

consequently, there is a fairly concrete notion of what would constitute necessary, useful, and formalizable syntactic information of general linguistic nature. Consequently, given the well-defined lexical requirements at syntactic level, there is that much more leverage in searching for (and finding) specific data to populate a lexicon at the syntactic level.

Most of the investigations aimed at recovery of lexical data from dictionaries fall in the category of 'localist' approaches. The notion is that if our goal is to construct an entry for a given word, then all (and the only) relevant information as far as the lexical properties of this word are concerned is to be found, locally, in the source dictionary entry for that word. This observation explains why constructing taxonomic networks on the basis of the general genus-differentiae model of dictionary definitions is essentially the extent to which identification of semantic information has been developed. It also underlies the pessimism concerning the useful semantic information extractable from a dictionary. Most dictionary entries are, indeed, impoverished when viewed in isolation; therefore, the lexical structures derived from them would be similarly under-representative.

It is important to take into account the relationship between the expressive power of on-line dictionary models and the scope of lexical information available via the access methods such models support. In particular, mounting a dictionary on-line only partially (as when leaving out certain fields and segments of entries) and/or ignoring components of an entry whose function is apparently only of typographical or aesthetic nature (such as typesetter control codes) tends to impose certain limitations on the kinds of lexical relationships that can be observed and recovered from a dictionary. Although, in principle, computational lexicography is concerned not only with developing techniques and methods for extraction of lexical data but also with building tools for making lexical resources available to such techniques and methods, in reality often the on-line dictionary model is not an adequate representation of lexical information on a large scale.

Finally, there is an alternative view emerging concerning a more 'realistic' definition of computational lexicography. Hoping to derive, by fully automatic

means, a computational lexicon — from one, or several, dictionary sources – is overly optimistic, and provably unrealistic. On the other hand, discarding the potential utility of such sources on the grounds that they have not yielded enough consistent and comprehensive information is unduly pessimistic. Between these two extremes there is an opinion that the potential of on-line dictionaries is in using them to facilitate and assist in the construction of large-scale lexicons. The image is not that of 'cranking the handle' and getting a lexicon overnight, but that of carefully designing a lexicon and then, for each aspect of lexical data deemed to be relevant for (semantic) processing of language, using the dictionary sources - in their entirety - to find instances of, and evidence for, such data. This paradigm relies on directed search for a number of specific lexical properties, and requires a much stronger notion of a theory of lexical semantics than assumed by computational lexicography to date.

## 4. Structure and Analysis of Machine-Readable Dictionaries

Prior to seeking interesting and meaningful generalizations concerning lexical information, repositories of such information – and more specifically, machine-readable dictionaries - should be suitably analyzed and converted to *lexical databases* (LDBs). The term "*lexical database*" refers to a highly structured isomorph of a published dictionary, which, by virtue of having both its data and structure made fully explicit, lends itself to flexible querying. Only such a general scheme for dictionary utilization would make it possible to make maximal use of the information contained in an MRD.

Dictionary sources are typically made available in the form of publishers' type-setting tapes. A tape carries a flat character stream where lexical data proper is heavily interspersed with special (control) characters. The particular denotation of typesetter control characters as font changes and other notational conventions used in the printed form of the dictionary is typically highly idiosyncratic and usually regarded as 'noise' when it comes to mounting a typesetting tape on-line for the purposes of computational lexicography.

None of the lexical database creation efforts to date addresses, explicitly, the question of fully utilizing the structural information in a dictionary, encoded in the control characters at source. Consequently, little attention has been paid to developing a general framework for processing the wide range of dictionary resources available in machine-readable form.

In situations where the conversion of an MRD into an LDB is carried out by a 'one-off program, typesetter information is treated mostly as 'noise' and consequently discarded. More modular (and, by design, customizable) MRD-to-LDB conversion schemes consisting of a parser and a grammar appear to retain this information; however, they assign only minimal interpretation to the 'semantics' of control codes. As a result, such efforts so far have not delivered the structurally rich and explicit LDB ideally required for easy and unconstrained access to the source data, as they have been driven by processing demands of a different nature from ours.

The computerization of the OED had, as its primary goal, setting up a dictionary database to be used by lexicographers in the process of (recompiling a dictionary for human, and human only, use. As a particular consequence, mapping from database representation to visual form of dictionary entries was a central concern of the design; so was efficiency in access. Another consequence of the same design was a highly idiosyncratic query language, making the kind of structure analysis discussed below difficult and unintuitive.

The majority of computational lexicography projects to date fall in the first of the above categories, in that they typically concentrate on the conversion of a single dictionary into an LDB. Even work based on more than one dictionary tends to use specialized programs for each dictionary source. In addition, not an uncommon property of existing LDBs is their completeness with respect to the original source: there is a tendency to extract, in a pre-processing phase, only some fragments (e.g., part of speech information or definition fields) while ignoring others (e.g., etymology, pronunciation, or usage notes).

This reflects a particular paradigm for deriving computational lexicons from MRDs: on the assumption that only a limited number of fields in a dictionary entry

are relevant to the contents of the target lexicon, these fields are extracted by arbitrary means; the original source is then discarded, and with it the lexical relationships implicit in the overall structure of an entry are lost. Such a strategy may be justified in some cases; in particular, it saves time and effort when a very precise notion exists of what information is sought from a dictionary and from where and how this is to be identified and extracted. In the general case, however, when a dictionary is to be regarded as a representative 'snapshot' of a language, containing a substantial amount of explicit and implicit information about words, selective analysis and partial load inevitably loses information. Although this process of 'pre-locating' lexical data in the complete raw source is occasionally referred to as "parsing" a typesetting tape, it is substantially different from the use of the same term below, where a parser is essentially a converter of the flat character stream into an arbitrarily complex structured representation, and parsing is both constrained never to discard any of the source content, and augmented with interpretations of the typesetter control codes in context.

Partial LDBs may be justified by the narrower, short-term requirements of specific projects; however, they are ultimately incapable of offering insights into the complex nature of lexical relations. The same is true of computerized dictionaries, which are available on-line, but only via a very limited, narrow bandwidth interface (typically allowing access exclusively by orthography). Even a functionally complete system for accessing an analyzed dictionary rapidly becomes unintuitive and cumbersome, if it is not based on fine-grained structural analysis of the source. For instance, the query processor designed to interact with the fully parsed version of the OED and capable of supporting a fairly comprehensive set of lexical queries, still faces problems of formulation and expressive power when it comes to asking questions concerning complex structural relationships between fields and components of dictionary entries.

An example of the functionality require for converting to a common LDB format a range of MRDs exhibiting a range of phenomena, is provided by the general mechanism embodied in the design of a *Dictionary Entry Parser* (DEP). A specific

implementation, described in detail by M.Neff and B.Boguraev, (1989, 1990) has been applied to the analysis of several different dictionaries.

DEP functions as a stand-alone parsing engine, capable of interpreting a dictionary tape character stream with respect to a grammar of that particular dictionary, and building an explicit parse tree of the entries in the MRD. In particular, rather than just tagging the data in the dictionary to indicate its structural characteristics, the grammar explicitly controls the construction of rather elaborate tree representations denoting deeper configurational relationships between individual records and fields within an entry. Two processes are crucial for 'unfolding', or making explicit, the structure of an MRD: identification of the structural markers, and their interpretation in context resulting in detailed parse trees for entries.

One final point needs to be made here. Dictionaries are incomplete and unreliable, as well as not fully consistent in form and content of definitions. This is an uncontroversial statement, and has been argued for (and against) quite extensively. For instance, in the case of looking for the default predicates naturally composable with "*book*", the most common - and by that token the most relevant - ones, namely, "*read*" and "*write*", are not part of any of the answers.

One of the concerns of computational lexicography is to remain aware of this fact, and consequently to develop techniques and methods for ensuring that the computational lexicons derived from machine-readable sources are more consistent, as well as fully representative with respect to the various lexical phenomena encoded in them.

The question is: how can we make absolutely certain that complete lists of collocationally appropriate forms, ranked by relevance, can be derived systematically for any kind of input? The answer to this question comes from a separate line of research, becoming integral to the study of word meaning and already beginning to extend the definition of "computational lexicography" given in the beginning of this chapter. Machine-readable dictionaries are not the only type of large-scale lexical resource available; equally important, and arguably richer and more representative of real language use, are the text corpora. Traditionally the basis for inducing stochastic

models of language, text corpora more recently have been used for extraction of a variety of lexical data.

## 5. The Contribution of Computational Linguistics

*The objectives of Computational Linguistics*

Work in *computational linguistics* began very soon after the development of the first computers, yet in the intervening four decades there has been a pervasive feeling that progress in computer understanding of natural language has not been commensurate with progress in other computer applications. Recently, a number of prominent researchers in natural language processing met to assess the state of the discipline and discuss future directions (M.Bates and R.M.Weischedel 1993). The consensus of those meetings was that increased attention to large amounts of lexical and domain knowledge was essential for significant progress, and current research efforts in the field reflect this point of view.

*The traditional approach in computational linguistics* included a prominent concentration on the formal mechanisms available for processing language, especially as these applied to syntactic processing and, somewhat less so, to semantic interpretation. In recent efforts, work in these areas continues, but there has been a marked trend toward enhancing these core resources with statistical knowledge-acquisition techniques. There is considerable research aimed at using online resources for assembling large knowledge bases, drawing on both natural language corpora and dictionaries and other structured resources. Recent research in lexical semantics reflects an interest in the proper structuring of this information to support linguistic processing. Furthermore, the availability of large amounts of machine-readable text naturally supports continued work in analysis of connected discourse. In other trends, statistical techniques are being used as part of the parsing process for automatic part-of-speech assignment and for word-sense disambiguation.

An indication of the development of natural language processing systems is that they are increasingly being used in support of other computer programs. This trend is

particularly noticeable with regard to information management applications. Natural language processing provides a means of gaining access to the information inherent in the large amount of text made available through the Internet.

In computational linguistics, *the lexicon* supplies paradigmatic information about words, including part-of-speech labels, irregular plurals, and subcategorization information for verbs. Traditionally, lexicons were quite small and were constructed largely by hand. There is a growing realization that effective natural language processing requires increased amounts of lexical (especially semantic) information. A recent trend has been the use of automatic techniques applied to large corpora to acquire lexical information from text. Statistical techniques are an important aspect of mining lexical information automatically. C.D.Manning (1993) uses such techniques to gather subcategorization information for verbs. M.R.Brent (1993) also discovers subcategorization information; in addition, he attempts to discover verbs in the text automatically.

### *5.1. The objectives of Computational Linguistics*

The *objectives of modern computational linguistics* are outlined below:

*Automatic tagging*

Automatically disambiguating part-of-speech labels in text is an important research area since such ambiguity is particularly prevalent in English. Programs resolving part-of-speech labels (often called automatic taggers) typically produce results that are approximately 95% accurate. Taggers can serve as preprocessors for syntactic parsers and contribute significantly to efficiency. There have been two main approaches to automatic tagging: probabilistic and rule-based. B.Merialdo (1994) and E.Dermatas and G.Kokkinakis (1995) review several approaches to probabilistic tagging. Typically, probabilistic taggers are trained on disambiguated text and vary as to how much training text is needed and how much human effort is required in the training process. Further variation concerns knowing what to do about unknown words and the ability to deal with large numbers of tags.

One drawback to stochastic taggers is that they are very large programs requiring considerable computational resources. E.Brill (1992) describes a rule-based tagger which is as accurate as stochastic taggers, but with a much smaller program. The program is slower than stochastic taggers, however. Building on Brill's approach, another rule-based, finite-state tagger which is much smaller and faster than stochastic implementations was proposed. Accuracy and other characteristics remain comparable.

*Parsing*

The traditional approach to natural language processing takes as its basic assumption that a system must assign a complete constituent analysis to every sentence it encounters. The methods used to attempt this are drawn from mathematics, with context-free grammars playing a large role in assigning syntactic constituent structure.

In continuing research in this tradition, context-free grammars have been extended in various ways. The so-called "mildly context sensitive grammars," such as tree adjoining grammars, have had considerable influence on recent work concerned with the formal aspects of parsing natural language.

Computational linguistics uses parsers for the automatic analysis of language. The term '*parser*' is derived from the Latin word pars meaning ‘part’, as in 'part of speech. Parsing in its most basic form consists in:

1. the automatic decomposition of a complex sign into its elementary components;
2. the automatic classification of the components via lexical lookup;
3. the automatic composition of the classified components via syntactic rules in order to arrive at an overall grammatical analysis of the complex sign.

Methodologically, the implementation of natural language grammars as parsers is important because it allows one to test the descriptive adequacy of formal rule systems automatically and objectively on real data. This new method of verification

is as characteristic for computational linguistics as the method of repeatable experiments is for natural science. Practically, the parsing of natural language may be used in different applications.

There exist nontraditional approaches to syntactic analysis. One such technique is partial, or underspecified, analysis. For many applications such an analysis is entirely sufficient and can often be more reliably produced than a fully specified structure. Statistical methods combined with a finite state mechanism are employed to impose an analysis which consists only of noun phrase boundaries; the analysis does not specify the complete internal structure or the exact place of the NP in a complete tree structure.

A recent innovation in syntactic processing has been investigations into the use of statistical techniques. In probabilistic parsing, probabilities are extracted from a parsed corpus for the purpose of choosing the most likely rule when more than one rule can apply during the course of a parse. In another application of probabilistic parsing, the goal is to choose the (semantically) best analysis from a number of syntactically correct analyses for a given input.

A more ambitious application of statistical methodologies to the parsing process is grammar induction where the rules themselves are automatically inferred from a bracketed text; however, results in the general case are still preliminary.

*Word-sense disambiguation*

Automatic word-sense disambiguation depends on the linguistic context encountered during processing. S.W.McRoy (1992) appeals to a variety of cues while parsing, including morphology, collocations, semantic context, and discourse. Her approach is not based on statistical methods, but rather is symbolic and knowledge intensive. Statistical methods exploit the distributional characteristics of words in large texts and require training, which can come from several sources, including human intervention.

*Formal Semantics*

*Formal semantics* is rooted in the philosophy of language and has as its goal a complete and rigorous description of the meaning of sentences in natural language. It concentrates on the structural aspects of meaning. G.Chierchia and S.McConnell-Ginet (1990) provide a good introduction to formal semantics. Various aspects of the use of formal semantics in computational linguistics are based on Montague grammar (R.Montague 1974).

*Lexical semantics* (D.A.Cruse 1986) has recently become increasingly important in natural language processing. This approach to semantics is concerned with psychological facts associated with the meanings of words. A very interesting application of lexical semantics is *Word-Net* (G.Miller 1990), which is a lexical database that attempts to model cognitive processes.

*Cognitive grammar* represents another approach to language analysis based on psychological considerations (R.W.Langacker 1988).

*Discourse analysis*

*Discourse analysis* is concerned with coherent processing of text segments larger than the sentence and assumes that this processing requires more than just the interpretation of the individual sentences. B.J.Grosz, A.K.Joshi and S.Weinstein (1995) provide a broad-based discussion of the nature of discourse, clarifying what is involved beyond the sentence level, and how the syntax and semantics of the sentences support the structure of the discourse. In their analysis, discourse contains linguistic structure (syntax, semantics), attentional structure (focus of attention), and intentional structure (plan of participants) and is structured into coherent segments. During discourse processing one important task for the hearer is to identify the referents of noun phrases. Inferencing is required for this identification. A coherent discourse lessens the amount of inferencing required of the hearer for comprehension. Throughout a discourse, the particular way that the speaker maintains "focus of attention" or "centering" through choice of linguistic structures for referring expressions is particularly relevant to discourse coherence.

Other work in computational approaches to discourse analysis has focused on particular aspects of processing coherent text. E.Hajicova, H.Skoumalova and P.Segall (1995) distinguish topic (old information) from focus (new information) within a sentence: Information of this sort is relevant to tracking focus of attention.

Several recent papers investigate those aspects of discourse processing having to do with the psychological state of the participants in a discourse, including goals, intentions, and beliefs. N.Asher and A.Lascarides (1994) investigate a formal model for representing the intentions of the participants in a discourse and the interaction of such intentions with discourse structure and semantic content. J.M.Wiebe (1994) investigates psychological point of view in third person narrative and provides an insightful algorithm for tracking this phenomenon in text. In Wiebe's study, the point of view of each sentence is either that of the narrator or any one of the characters in the narrative. J.M.Wiebe examines the importance of determining point of view for a complete under standing of a text and discusses how this concept interacts with other aspects of discourse structure.

*5.2. Practical tasks of computational linguistics*

1. *Indexing and retrieval in textual databases*

Textual databases electronically store texts such as publications of daily newspapers, medical journals, and court decisions. The user of such a database should be able to find exactly those documents and passages with comfort and speed which are relevant for the specific task in question. The World Wide Web (WWW) may also be viewed as a large, unstructured textual database, which daily demonstrates to a growing number of users the difficulties of successfully finding the information desired.

*2. Machine translation*

Especially in the European Union, currently with eleven different languages, the potential utility of automatic or even semi-automatic translation systems is tremendous.

*3. Automatic text production*

Large companies which continually bring out new products such as engines, video recorders, farming equipment, etc., must constantly modify the associated product descriptions and maintenance manuals. A similar situation holds for lawyers, tax accountants, personnel officers, etc., who must deal with large amounts of correspondence in which most of the letters differ only in a few, well-defined places. Here techniques of automatic text production can help, ranging from simple templates to highly flexible and interactive systems using sophisticated linguistic knowledge.

*4. Automatic text checking*

Applications in this area range from simple spelling checkers (based on word form lists) via word form recognition (based on a morphological parser) to syntax checkers based on syntactic parsers which can find errors in word order, agreement, etc.

*5. Automatic content analysis*

The printed information on this planet is said to double every 10 years. Even in specialized fields such as natural science, law, or economics, the constant stream of relevant new literature is so large that researchers and professionals do not nearly have enough time to read it all. A reliable automatic content analysis in the form of brief summaries would be very useful. Automatic content analysis is also a pre-condition for concept-based indexing, needed for accurate retrieval from textual databases, as well as for adequate machine translation.

*6. Automatic tutoring*

There are numerous areas of teaching in which much time is spent on drill exercises such as the more or less mechanical practicing of regular and irregular paradigms in foreign languages. These may be done just as well on the computer, providing the students with more fun (if they are presented as a game, for example) and the teacher with additional time for other, more sophisticated activities such as conversation. Furthermore, these systems may produce automatic protocols detailing the most frequent errors and the amount of time needed for various phases of the

exercise. This constitutes a valuable heuristics for improving the automatic tutoring system ergonometrically. It has led to a new field of research in which the 'electronic text book' of old is replaced by new teaching programs utilizing the special possibilities of the electronic medium to facilitate learning in ways never explored before.

*7. Automatic dialog and information systems*

These applications range from automatic information services for train schedules via queries and storage in medical databases to automatic tax consulting.

This list is by no means complete, however, because the possible applications of computational linguistics include all areas in which humans communicate with computers and other machines of this level, today or in the future.

In summary, traditional language sciences may contribute substantially to improving automatic language processing in computational applications. Computers, on the other hand, are an essential tool for improving empirical analysis in linguistics - not only in certain details, but as a complete and efficiently functioning theory of language which is realized concretely in terms of unrestricted natural human-computer communication.

## *5.3. Applications of Computational Linguistics Research*

As natural language processing technology matures, it is increasingly being used to support other computer applications. Such use naturally falls into two areas: One involves linguistic analysis and serves as an interface to the primary program; the second involves natural language considerations that are central to the application.

Natural language processing interfaces to *data base management systems* (DMBS) translate users' input into a request in a formal data base query language, and the program then proceeds as it would without the use of natural language processing techniques. It is normally the case that the domain is constrained and the language of the input consists of comparatively short sentences with a constrained set of syntactic structures.

The design of *question-answering systems* is similar to that for interfaces to data base management systems. One difference, however, is that the knowledge base supporting the question-answering system does not have the structure of a data base. The *underlying knowledge base* may function as an on-line encyclopedia. Processing in this system not only requires a linguistic description for users' requests but also a representation for the encyclopedia itself. As with the interface to a DBMS, the requests are likely to be short and have a constrained syntactic structure.

In *message-understanding systems*, a fairly complete linguistic analysis may be required, but the messages are relatively short and the domain is often limited.

In three closely related applications (*information filtering*, *text categorization*, *and automatic abstracting*), no constraints on the linguistic structure of the documents being processed can be assumed. One mitigating factor, however, is that effective processing may not require a complete analysis. For all of these applications, there are also statistically-based systems that operate on frequency distributions of words. These systems work fairly well, but most people feel that for further improvements, and for extensions, some sort of understanding of the texts, such as that provided by linguistic analysis, is required.

*Information filtering* and *text categorization* are concerned with comparing one document to another. In both applications, natural language processing imposes a linguistic representation on each document being considered. In information filtering, documents satisfying some criterion are singled out from a collection. In text categorization, a collection of documents is inspected and all documents are grouped into several categories based on the characteristics of the linguistic representations of the documents.

In *automatic abstracting*, a summary of each document is sought, rather than a classification of a collection. The underlying technology is similar to that used for information filtering and text categorization: Use is made of some sort of linguistic representation of the documents. Of the two major approaches, one puts more emphasis on semantic analysis for this representation, and the other places less emphasis on semantic analysis.

*Information retrieval systems* typically allow a user to retrieve documents from a large bibliographic database. During the information retrieval process, a user expresses an information need through a query. The system then attempts to match this query to those documents in the database which satisfy the user's information need. In systems which use natural language processing, both query and documents are transformed into some sort of a linguistic structure, and this forms the basis of the matching. Several recent information retrieval systems employ varying levels of linguistic representation for this purpose.

## 6. Technology and Grammar

A purely technological approach to natural language processing may be enhanced by using bits of linguistic knowledge. However, without a comprehensive theory of natural language communication, the resulting improvements turn out to be quite limited.

### *6.1. Indexing and retrieval in textual databases*

A textual database is an arbitrary collection of electronically stored texts. In contrast to a classic, record-based database, no structural restrictions apply to a textual database. Thus, the individual texts may be arranged, for example, in the temporal order of their arrival, according to their subject matter, the name of their author(s), their length, or according no principle at all.

The search for a certain text or text passage is based on the standard, letter-based indexing of the textual database.

The indexing of a textual database is based on a table which specifies for each letter all the positions (addresses) where it occurs in the storage medium of the database.

The electronic index of a textual database functions in many ways like a traditional library catalog of alphabetically ordered file cards.

Each file card contains a keyword, e.g., the name of the author, and the associated addresses, e.g., the shelf where the book of the author may be found. While the file cards are ordered alphabetically according to their respective keywords, the choice of the addresses is free. Once a given book has been assigned a certain address and this address has been noted in the catalog, however, it is bound to this address.

In an unordered library without a catalog, the search for a certain book requires looking through the shelves (linear search). In the worst case, the book in question happens to be on the last of them. A library catalog speeds up such searching because it replaces a linear search by specifying the exact address(es) of the physical location. Thus, a book may be found using the alphabetic order of the file cards, irrespective of how the actual locations of the books are arranged.

The electronic index of a textual database uses the letters of the alphabet like the keywords of a library catalog, specifying for each letter all its positions (addresses) in the storage medium. The occurrences of a certain word form, e.g., sale, is then computed from the intersection of the position sets of s, a, I, and e. The electronic index is built up automatically when the texts are read into the database, whereby the size of the index is roughly the same as that of the textual database itself.

The search for relevant texts or passages in the database is guided by the user on the basis of words (s)he considers characteristic of the subject matter at hand. Consider for example a lawyer interested in legal decisions dealing with warranties in used car sales. After accessing an electronic database in which all federal court decisions since 1960 are stored, (s)he specifies the words warranty, sale, and used car. After a few seconds the database returns a list of all the texts in which these words occur. When the user clicks on a title in the list the corresponding text appears on the screen.

The user might well find that not all texts in the query result are actually relevant for the purpose at hand. It is much easier, however, to look through the texts of the query result than to look through the entire database.

Also, the database might still contain texts which happen to be relevant to the subject matter, yet are not included in the query result. Such texts, however, would be those which deal with the subject matter without mentioning the query words.

The use of an electronic index has the following advantages over a card index:

- *Power of search*

Because the electronic index of a textual database uses the letters of the alphabet as its keys, the database may be searched for any sequence of letters, whereas the keys of a conventional catalogue are limited to certain kinds of words, such as the name of the author.

- *Flexibility*

  - General specification of patterns

  An electronic index makes it possible to search for patterns. For example, the pattern *in.*i..tion* matches all word forms of which the first two letters are *in*, the seventh letter from the end is *i* and the last four letters are *tion*, as in *inhibition* and *inclination*.

  - Combination of patterns

  The electronic index makes it possible to search for the combination of several word forms, whereby a maximal distance for their co-occurrence may be specified.

  Though it is theoretically possible to create a conventional card index for the positions of each letter of the books in a library, this would not be practical. For this reason, searching with patterns or the combination of keywords and/or patterns is not technically feasible with a conventional card index.

- *Automatic creation of the index structure*

The electronic index of a textual database is generated automatically during the reading-in of texts into the database. In a conventional card index, on the other hand, each new keyword requires making a new card by hand.

- *Ease, speed, and reliability*

While an electronic search is done automatically in milliseconds, error free, and complete, a conventional search using a card index requires human labor, is susceptible to errors, and may take anywhere from minutes to hours or days. The advantages of electronic search apply to both the query (input of the search words) and the retrieval (output of the corresponding texts or passages).

- *Query*

An electronic database is queried by typing the search patterns on the computer, while the use of a card index requires picking out the relevant cards by hand.

- *Retrieval*

In an electronic database, the retrieved texts or passages are displayed on the screen automatically, while use of a conventional card index requires going to the library shelves to get the books.

The quality of a query result is measured in terms of recall and precision.

*Recall* measures the percentage of relevant texts retrieved as compared to the total of relevant texts contained in the database.

For example: a database of several million pieces of text happens to contain 100 texts which are relevant to a given question. If the query returns 75 texts, 50 of which are relevant to the user and 25 irrelevant, then the recall is 50 : 100 = 50%.

*Precision* measures the percentage of relevant texts contained in the result of a query.

For example: a query results in 75 texts of which 50 turn out to be relevant to the user. Then the precision is 50 : 75 = 66.6%.

Experience has shown that recall and precision are not independent of each other, but inversely proportional: a highly specific query will result in low recall with high precision, while a loosely formulated query will result in high recall with low precision.

High recall has the advantage of retrieving a large percentage of the relevant texts from the database. Because of the concomitant low precision, however, the user has to work through a huge amount of material most of which turns out to be irrelevant.

High precision, on the other hand, produces a return most of which is relevant for the user. Because of the concomitant low recall, however, the user has to accept the likelihood that a large percentage of relevant texts remain undiscovered.

Measuring recall is difficult in large databases. It presupposes exact knowledge of all the texts or passages which happen to be relevant for any given query. To obtain this knowledge, one would have to search the entire database manually in order to objectively determine the complete set of documents relevant to the user's question and to compare it with the automatic query result.

Measuring precision, on the other hand, is easy, because the number of documents returned by the system in response to a query is small compared to the overall database. The user need only look through the documents in the query result in order to find out which of them are relevant.

In a famous and controversial study, J.Blair & D.Maron in 1985 attempted to measure the average recall of a leading commercial database system called STAIRS (STAIRS is an acronym for Storage and Information Retrieval System, a software product developed and distributed by IBM). For this purpose they cooperated with a large law firm whose electronic data comprised 40 000 documents, amounting to a total of 350 000 pages. Because of this substantial, but at the same time manageable, size of the data it was possible to roughly determine the real number of relevant texts for 51 queries with the assistance of the employees.

Prior to the study, the employees subjectively estimated an electronic recall of 75%. The nonelectronic verification, however, determined an average recall of only 20%, with a standard deviation of 15.9%, and an average precision of 79.0%, with a standard deviation of 22.2%.

### *6.2. Using grammatical knowledge*

The reason for the surprisingly low recall of only 20% on average is that STAIRS uses only technological, i.e., letter-based, methods. Using grammatical knowledge in addition, recall could be improved considerably. Textual phenomena

which resist a technological treatment, but are suitable for a linguistic solution, are listed below under the heading of the associated grammatical component.

To the linguistic phenomena requiring new solutions and reinterpretation belong:

*Morphology*

A letter-based search does not recognize words. For example, the search for *sell* will overlook relevant forms like *sold*.

A possible remedy would be a program for word form recognition which automatically assigns to each word form the corresponding base form. By systematically associating each word form with its base form, all variants of a search word in the database can be found. A program of automatic word form recognition would be superior to the customary method of truncation - especially in languages with a morphology richer than that of English.

*Lexicon*

A letter-based search does not take semantic relations between words into account. For example, the search for *car* would ignore relevant occurrences such as *convertible, pickup truck, station wagon*, etc.

A lexical structure which automatically specifies for each word the set of equivalent terms (synonyms), of the superclass (hypernyms), and of the set of instantiations (hyponyms) can help to overcome this weakness, especially when the domain is taken into account.

*Syntax*

A letter-based search does not take syntactic structures into account. Thus, the system does not distinguish between, for example, *teenagers sold used cars* and *teenagers were sold used cars*.

A possible remedy would be a syntactic parser which recognizes different grammatical relations between, for example, the subject and the object. Such a parser, which presupposes automatic word form recognition, would be superior to the currently used search for words within specified maximal distances.

*Semantics*

A letter-based search does not recognize semantic relations such as negation. For example, the system would not be able to distinguish between *selling cars* and *selling no cars*. Also, equivalent descriptions of the same facts, such as *A sold x to B* and *B bought x from A*, could not be recognized.

Based on a syntactic parser and a suitable lexicon, the semantic interpretation of a textual database could analyze these distinctions and relations, helping to improve recall and precision.

*Pragmatics*

According to Blair & Maron 1985, a major reason for the poor recall was the frequent use of context-dependent formulations such as *concerning our last letter, following our recent discussion*, as well as nonspecific words such as *problem, situation, or occurrence*.

The treatment of these frequent phenomena requires a complete theoretical understanding of natural language pragmatics. For example, the system will have to be able to infer that, for example, *seventeen-year-old bought battered convertible is relevant to the query used car sales to teenagers.*

In order to improve recall and precision, linguistic knowledge may be applied in various different places in the database structure. The main alternatives are whether improvements in the search should be based on preprocessing the query, refining the index, and/or postprocessing the result. Further alternatives are an automatic or an interactive refinement of the query and/or the result, as described below.

*Linguistic methods of optimization:*

*A. Preprocessing the query*

- Automatic query expansion

1. The search words in the query are automatically 'exploded' into their full inflectional paradigm and the inflectional forms are added to the query.

2. Via a thesaurus the search words are related to all synonyms, hypernyms, and hyponyms. These are included in the query - possibly with all their inflectional variants.

3. The syntactic structure of the query, e.g., *A sold x to B*, is transformed automatically into equivalent versions, e.g., *B was sold x by A*, *x was sold to B by A*, etc., to be used in the query.

- Interactive query improvement

The automatic expansion of the query may result in an uneconomic widening of the search and considerably lower precision. Therefore, prior to the search, the result of a query expansion is presented to the user to eliminate useless aspects of the automatic expansion and to improve the formulation of the query.

*B. Improving the indexing*

- Letter-based indexing

This is the basic technology of search, allowing one to retrieve the positions of each letter and each letter sequence in the database.

- Morphologically-based indexing

A morphological analyzer is applied during the reading-in of texts, relating each word form to its base form. This information is coded into an index, which for any given word (base form) allows one to find the associated word forms in the text.

- Syntactically-based indexing

A syntactic parser is applied during the reading-in of texts, eliminating morphological ambiguities and categorizing phrases. This information is coded into an index on the basis of which all occurrences of a given syntactic construction may be found.

- Concept-based indexing

The texts are analyzed semantically and pragmatically, whereby the software eliminates syntactic and semantic ambiguities and infers special uses characteristic of the domain. This information is coded into an index on the basis which all occurrences of a given concept may be found.

*C. Postquery processing*

- The low precision resulting from a nonspecific formulation of the query may be countered by an automatic processing of the data retrieved. Because there is only a small amount of raw data retrieved, as compared to the database as a whole, they may

be parsed after the query and checked for their content. Then only those texts are displayed which are relevant according to this postquery analysis.

The ultimate goal of indexing textual databases is a concept-based indexing founded on a complete morphological, syntactic, semantic, and pragmatic analysis of the texts.

*Smart versus solid solutions*

Which of the alternatives mentioned above is actually chosen in the design of a textual database depends on the amount of data to be handled, the available memory and speed of the hardware, the users' requirements regarding recall, precision, and speed of the search, and the designer's preferences and abilities. At the same time, the alternatives are not independent from each other.

For example, if an improvement of recall and precision is to be achieved via an automatic processing of the query, one can use a simple indexing. More specifically, if the processing of the query explodes the search words into their full inflectional paradigm for use in the search, a morphological index of the database would be superfluous. Conversely, if there is a morphological index, there would be no need for exploding the search words.

Similarly, the automatic expansion of queries may be relatively carefree if it is to be scrutinized by the user prior to search. If no interactive fine-tuning of queries is provided, on the other hand, the automatic expansion should be handled restrictively in order to avoid a drastic lowering of precision.

Finally, the indexing of texts can be comparatively simple if the results of each query are automatically analyzed and reduced to the most relevant cases before output to the user. Conversely, a very powerful index method, such as concept-based indexing, would produce results with such high precision that there would be no need for an automatic postprocessing of results.

The different degrees of using linguistic theory for handling the retrieval from textual databases illustrate a general alternative in the design of computational applications, namely the choice between *smart* versus *solid* solutions.

*Smart solutions* avoid difficult, costly, or theoretically unsolved aspects of natural communication, as in

1) Weizenbaum's Eliza program, which appears to understand natural language, but doesn't;
2) direct and transfer approaches in machine translation, which avoid understanding the source text ;
3) finite state technology and statistics for tagging and probabilistic parsing.

These methods may seem impressive because of the vast number of toys and tools assembled in the course of many decades. But they do not provide an answer to the question of how natural language communication works. Imagine that the Martians came to earth and modelled cars statistically: they would never run.

Initially, smart solutions seem cheaper and quicker, but they are costly to maintain and their accuracy cannot be substantially improved. The alternative is solid solutions.

*Solid solutions* aim at a complete theoretical and practical understanding of natural language communication. Applications are based on ready-made off-the-shelf components such as

- on-line lexica;
- rule-based grammars for the syntactic-semantic analysis of word forms and sentences;
- parsers and generators for running the grammars in the analysis and production of free text;
- reference and monitor corpora for different domains, which provide a systematic, standardized account of the current state of the language.

Solid solution components are an application-independent long term investment. Due to their systematic theoretical structure they are easy to maintain, can be improved continuously, and may be used again and again in different applications.

Whether a given task is suitable for a smart or a solid solution depends to a great extent on whether the application requires a perfect result or whether a partial answer

is sufficient. For example, a user working with a giant textual database will be greatly helped by a recall of 70%, while a machine translation system with 70% accuracy will be of little practical use.

The two tasks differ in that a 70% recall in a giant database is much more than a user could ever hope to achieve with human effort alone. Also, the user never knows which texts the system did not retrieve.

In translation, on the other hand, the deficits of an automatic system with 70% accuracy are painfully obvious to the user. Furthermore there is an alternative available, namely professional human translators. Because of the costly and time-consuming human correction required by today's machine translation, the user is faced daily with the question of whether or not the machine translation system should be thrown out altogether in order to rely on human work completely.

Another, more practical factor in the choice between a smart and a solid solution in computational linguistics is the off-the-shelf availability of grammatical components for the natural language in question. Such components of grammar, e.g., automatic word form recognition, syntactic parsing, etc., must be developed independently of any specific applications as part of basic research - solely in accordance with the general criteria of (1) their functional role as components in the mechanism of natural communication, (2) completeness of data coverage, and (3) efficiency.

Modular subsystems fulfilling these criteria can be used in practical applications without any need for modification, using their standard interfaces. The more such modules become available as ready-made, well-documented, portable, off-the-shelf products for different languages, the less costly will the strategy of solid solutions be in practical applications.

The main reason for the long term superiority of solid solutions, however, is quality. This is because a 70% smart solution is typically very difficult or even impossible to improve to 71%.

## 7. Machine Translation and Other Translation Technologies

Computers have become much more widely used in translation since the early 1980s but in unexpected ways. From the beginnings of machine translation research in the 1950s until recently, it was expected that computers would be in direct competition with human translators for the same work. Instead, it has turned out that most translation done by computers fills latent needs that do not reduce the amount of work available to professional translators. Computers have also turned out to be useful as productivity tools for human translators who still perform the central translation task. Today, the relationship between computers and human translators is often seen as synergistic rather than competitive.

*Machine translation* may be defined as a computer application in which the machine analyzes complete sentences into constituents, selects target-language words (usually to correspond to source-language words), and generates target-language sentences. Other types of translation technology are classified as *translator tools*. For example, consider a computer system that retrieves the translation of a complete sentence from a stored database of pairs of sentences. Such a system would not qualify as machine translation since the target-language sentences were previously translated rather than being generated by the system. The origin of the sentence pairs could have been human or machine translation. The retrieval system is a translator tool in either case.

### *7.1. Approaches to Machine Translation*

Translation in general requires understanding a text or utterance in a certain language (interpretation) and reconstructing it in another language (production).

On the one hand, translation goes beyond the general repertoire of natural communication. Superficially, it may seem related to bilingual communication. Bilingualism, however, is merely the ability to switch between languages, whereby only one language is used at any given time - in contradistinction to translation.

On the other hand, translation offers the facilitating circumstance that a coherent source text is given in advance - in contrast to automatic language production, which has to grapple with the problems of 'what to say' and 'how to say it.' A given source text can be utilized to avoid the really difficult problems of interpretation (e.g., a language independent modeling or understanding) and production (e.g. the selection of content, the serialization, the lexical selection) in order to automatically translate large amounts of nonliterary text, usually into several different languages at once.

The administration of the European Union, for example, must publish every report, protocol, decree, law, etc., in the 11 different languages of the member states (as of 2001). For example, a decree formulated in French under a French EU presidency would have to be translated into the following 10 languages:

French – English

French – German

French – Italian

French – Dutch

French – Swedish

French – Spanish

French – Portuguese

French - Greek

French - Danish

French - Finnish

Under a Danish EU presidency, on the other hand, a document might first be formulated in Danish. Then it would have to be translated into the remaining EU languages, resulting in another set of language pairs.

The total number of language pairs for a set of different languages is determined by the following formula:

$n \times (n — 1)$, where n = number of different languages

For example, an EU with 11 different languages has to deal with a total of $11 \times 10 = 110$ language pairs.

In a language pair, the source language (SL) and the target language (TL) are distinguished. For example, 'French→Danish' and 'Danish→French' are different language pairs. The source language poses the task of correctly *understanding* the meaning, taking into account the domain and the context of utterance, whereas the target language poses the task of *formulating* the meaning in a rhetorically correct way.

The first attempts at machine translation tried to get as far as possible with the new computer technology, avoiding linguistic theory as much as possible. This resulted in the smart solution of 'direct translation,' which was dominant in the 1950's and -60's.

Direct translation systems assign to each word form in the source language a corresponding form of the target language. In this way the designers of these systems hoped to avoid a meaning analysis of the source text, while arriving at translations which are syntactically acceptable and express the meaning correctly. The schema of direct translation may be viewed in fig. 4.1.

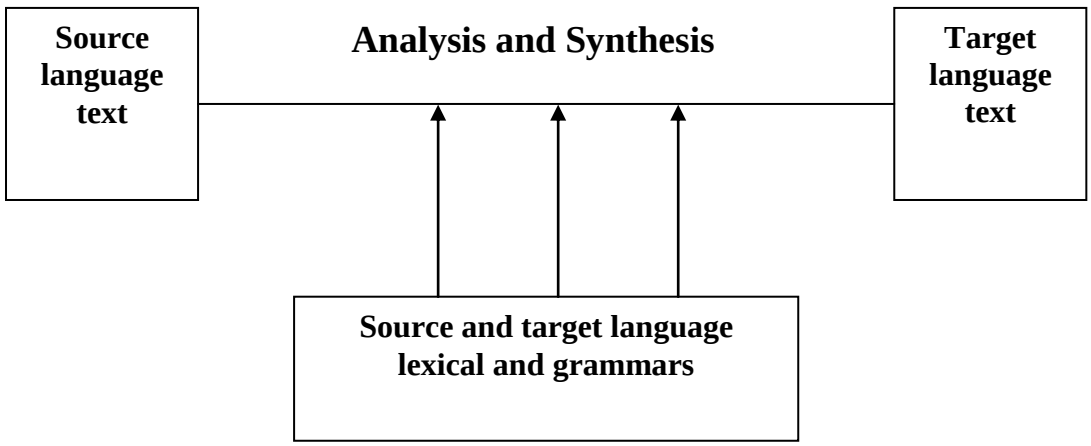



Figure 4.1. Schema of direct translation

Each language pair requires the programming of its own direct translation system. Direct translation is based mainly on a differentiated dictionary, distinguishing many special cases for a correct assignment of word forms in the target language. In the source language, grammatical analysis is limited to resolving ambiguities as much as possible; in the target language, it is limited to adjusting the word order.

The methodological weakness of direct translation systems is that they do not systematically separate source language analysis and target language synthesis. Consequently one is forced with each new text to add new special cases and exceptions. In this way the little systematic structure which was present initially is quickly swept away by a tidal wave of exceptions and special cases.

Even though in the 1950's, representatives of the direct approach repeatedly asserted that the goal of machine translation, namely FULLY AUTOMATIC HIGH QUALITY TRANSLATION (FAHQT) was just around the corner, their hopes were not fulfilled. Hutchins 1986 provides the following examples to illustrate the striking shortcomings of early translation systems:

*Out of sight, out of mind. => Invisible idiot.*

*The spirit is willing, but the flesh is weak.=> The whiskey is all right, but the meat is rotten.*

These two examples are apocryphal, described as the result of an automatic translation from English into Russian and back into English.

An attempt to avoid the weaknesses of direct translation is the transfer approach, which is viewed in fig. 4.2.:

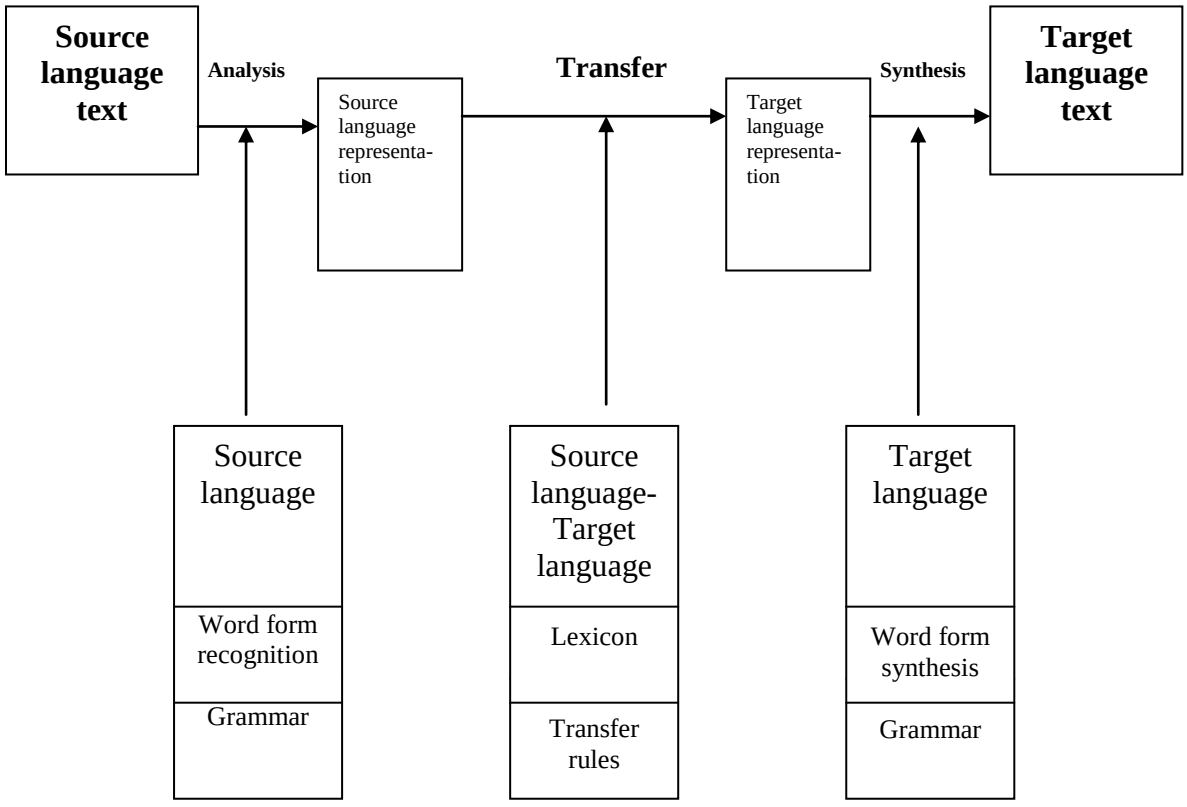


Figure 4.2. Schema of Transfer approach

The transfer approach is characterized by a modular separation of:

1) source language analysis and target language synthesis;

2) linguistic data and processing procedures;

3) the lexica for source language analysis, target language transfer and target language synthesis.

This results in a clearer structure as compared to the direct approach, facilitating debugging and upscaling of transfer systems. Implementing the different modules independently of each other and separating the computational algorithm from the language-specific data also makes it possible to reuse parts of the software.

For example, given a transfer system for the language pair A-B, adding the new language pair A-C will require writing new transfer and synthesis modules for language C, but the analysis module of the source language A may be reused. Furthermore, if the language-specific aspects of the new transfer and synthesis modules are written within a prespecified software framework suitable for different languages, the new language pair A-C should be operational from the beginning.

The three phases of the transfer approach are illustrated below with a word form (English-German transfer):

1. *Source language analysis:*

Unanalyzed surface: *knew*

Morphological and

lexical analysis: (*knew* (N A V) *know*)

The source language analysis produces the syntactic category (N A V) of the inflectional form (categorization) and the base form *know* (lemmatization).

2. *Source-target language transfer:*

Using the base form resulting from the source language analysis, a source-target language dictionary provides the corresponding base forms in the target language.

*know → wissen*

*kennen*

3. *Target language synthesis*

Using the source language category (resulting from analysis) and the target language base forms (resulting from transfer), the desired target language word forms are generated based on target language morphology.

The transfer of a syntactic structure functions similarly to the transfer of word forms. First, the syntactic structure of the source language sentence is analyzed. Second, a corresponding syntactic structure of the target language is determined (transfer). Third, the target language structure is supplied with the target language word forms (synthesis), whereby correct handling of agreement, domain-specific lexical selection, correct positioning of pronouns, rhetorically suitable word order, and other issues of this kind must be resolved.

Due to similarities between the direct and the transfer method, they have the following shortcomings in common:

- Each language pair requires a special source-target component.
- Analysis and synthesis are limited to single sentences.
- Semantic and pragmatic analyses are avoided, attempting automatic translation without understanding the source language.

Thus, the advantage of the transfer approach over the direct approach is limited to the reusability of certain components, specifically the source language analysis and the target language synthesis for additional language pairs.

*7.2. Linguistic aspects of Machine Translation*

Most machine translation systems are based on mainstream linguistic theory. Mainstream linguistics identifies several levels of representation, including a morphological level, a syntactic level, and a semantic level. These levels will each be discussed in turn.

1. *Morphology*

Machine translation systems typically identify the words of a source-language sentence and look them up in a dictionary. For English, whose morphology is relatively simple, it is feasible to store all the inflected forms of a word in the index of a dictionary. However, for more highly inflected languages, computer routines are written to analyze a word into its base form (or, in some cases, into its possible base forms). For example, the French verb form *prenons* would be analyzed into the base form *prendre* and some codes to indicate that the base form is related to the inflected form by the features "first person, plural, present, indicative." This morphological analysis permits the dictionary builder to put just the base form in the dictionary index. The dictionary may also contain information under the base form to guide the process of morphological analysis. Of course, dictionaries intended for use by humans are also usually organized so that only base forms are used as headwords. Humans somehow perform the equivalent of morphological analysis when they see an inflected form in a text and look it up in a dictionary. Humans do this rapidly and unconsciously, and they also perform the equivalent of the reverse process: morphological generation. Morphological generation starts with a base form and a set of features and produces an inflected form. Applied computational morphology – that is, technology that performs morphological analysis and generation – is an element of natural language processing which is essential to machine translation. Computational morphology is also useful in other translation technologies.

2. *Syntax*

One reason for looking up words in a machine-translation dictionary is to retrieve syntactic information about them. The types of syntactic information are taken from linguistic theory. For example, the dictionary entry for a base form will usually specify what grammatical category or categories it can fill and, in the case of a verb, how many and what kinds of arguments it takes. The grammatical information is used to perform a syntactic analysis of a sentence. This process of syntactic analysis is called "parsing," and the computer application that performs it is called a "parser." The result of parsing is a syntactic representation. This syntactic representation may look like a D-structure tree in Government and Binding theory,

like an F-structure in Lexical Functional Grammar, or it may resemble some eclectic representation.

3. *Semantics*

Another reason for looking up words in a machine-translation dictionary, apart from the retrieval of morphological and syntactic information, is the retrieval of semantic information. One approach to semantics, from the Chomskyan framework called Principles and Parameters, is to derive a semantic representation, called *a logical form*, from the syntactic representation. However, in machine translation practice this is seldom done. More common is a practice to retrieve semantic features that assist the parser in producing a reasonable syntactic representation. Another type of semantic information that must somehow be retrieved in a machine-translation system is what target-language words to use. In the case of a specialized term in a document from a specific domain, there may well be a straightforward one-to-one correspondence between source and target terms. However, there is not always a one-to-one correspondence. General vocabulary words may be translated in many different ways, and it is not at all straightforward to program a computer to select an appropriate target-language word. The process of selecting an appropriate target-language word during machine translation is called *lexical transfer*. In machine translation, lexical transfer is an unavoidable and therefore highly important aspect of semantics.

*Lexical Transfer*

The problem with lexical transfer is that, in a sense, it takes machine translation outside the realm of linguistics. This apparently strange statement is based on the fact that mainstream linguistic theory does not include lexical transfer in the way it includes morphology and syntax. *Computational morphology* is not trivial, but at least the word forms that are supposed to go in and come out are well-defined in mainstream linguistic theory. *Computational syntax* is far from trivial, but at least the syntactic representations that are supposed to be derived from a given sentence are well-defined by a particular syntactic theory (within the bounds of disagreement among syntacticians over details). However, mainstream linguistic theory is

monolingual (e.g., Universal Grammar deals with one language at a time rather than with cross-language equivalence), and thus there can be no component of mainstream theory on which lexical transfer, a necessarily bilingual process, is based.

The field of translation studies, which is outside of mainstream linguistics, also offers no relief to our dilemma. From the point of view of modern translation studies, the standard approach to machine translation is based on a now-discarded view of what humans do when they translate. The standard *analysis-transfer-generation* approach to machine translation is based on the assumption that high-quality translations can be obtained by analyzing the syntax of a source text sentence by sentence, substituting target-language words for source-language words, and adjusting the syntactic structure as needed to conform to the grammar of the target language. However, results of work in the field of translation studies, which is primarily concerned with studying how humans translate, would suggest that humans do not use an analysis-transfer-generation approach. The rub is that translation study does not currently provide a formal model of how to translate. Even worse, it condemns the very idea that a word of general vocabulary can be translated as an isolated unit. Only texts can be translated. Mainstream monolingual semantics does not come to the rescue either, since it does not deal directly with the problems of translation. Thus, it would seem that lexical transfer is not part of linguistics because there is no formal theory of source-target word substitution on which to base lexical transfer.

In the absence of a solid theoretical basis for the analysis-transfer-generation approach to machine translation, it is not surprising that the results of this type of machine translation are not very impressive when compared directly with the results of a professional human translator. A ten-year-long machine-translation project of the European Commission, called *Eurotra*, recently ended. It was heavily funded, yet it did not result in a large-scale industrial prototype as was originally expected by the funders. At the same time, nevertheless, machine translation is being used more and more, although not in direct competition with human translation.

*7.3. Real-world uses of Machine Translation*

The nature of the machine translation depends on whether high-quality output is needed or whether low-quality output will be sufficient.

1. *High-quality output*

If high-quality output – output that is comparable to the work of a professional human translator – is important, then at least three conditions must be met before machine translation is an option:

- The texts to be translated must be restricted to a well-defined domain of knowledge shared between source and target languages.
- The source texts must be carefully controlled to conform to a formal syntax
  and semantics. Such texts are said to be in a "controlled language."
- The machine-translation system must be tailored to the domain and the controlled language.

In practice, these conditions mean that less than five percent of what is translated is high-quality machine translation.

Human translators are often uninterested in doing work that meets the first two conditions because such work is so repetitive and boring. Thus, there is little competition between human and machine translation when high-quality output is needed. There is also an option for human-post-edited machine-translation output. This hybrid is only cost-effective when the raw machine-translation output is already very close to being useable without major revision or when the quality of the output is set to a lower standard.

2. *Low-quality output*

Low-quality machine translation output is full of errors of various kinds, errors which reflect the fact that current machine-translation systems do not understand what they are translating but are blindly manipulating words according to pre-programmed rules. Such output is surprisingly useful. Its usefulness is based on the fact that humans are very adaptable and can make sense of sentences which are ungrammatical and would sport an asterisk in the generative grammar

tradition. Sometimes, a human will accept a low-quality machine translation that is available quickly and inexpensively. Low-quality translation, also called "indicative" translation, is primarily intended for individual use and is never used for publication purposes. There are two growing markets for low-quality machine translation because it gives an indication of the content of the source-language text.

One use of low-quality output is the professional market. Here a government, academic, or business individual needs to gain a general understanding of the content of a document in a language that the individual does not read. Often, an indicative machine translation is adequate to decide whether the document is of genuine interest. If it is of sufficient relevance, then a human translation can be requested. This methodology has been used with success for many years in the United States Air Force, and indicative translation is the fastest growing type of translation done at the European. It is interesting to note that, although the use of indicative translation has dramatically increased at the Commission over the past few years, the demand for traditional human translation has not diminished. This supports the claim that much indicative translation addresses a latent, unfulfilled need rather than being in competition with human translation.

The other growing market for indicative translation is in personal communication, particularly by electronic mail. The on-line service *CompuServe* recently began offering a service which allows users to participate in a discussion forum even though they do not all speak the same language. Every three minutes, new messages are machine-translated into the other languages of the forum. The translations are of low quality, but users still pay for the service.

*Statistics Based Machine-Translation*

Another type of machine translation that has been recently attempted is a statistics-based approach. This alternative, a major departure from the analysis-transfer-generation approach, is based on a large, aligned, bilingual corpus.

A bilingual corpus is a corpus of texts which includes text pairs such that one member of the pair is in language A and the other member is in language B. Normally, one member of the pair is a careful human translation of the other. An aligned bilingual corpus includes links between the two members of a pair of texts. The links show explicitly which units of text correspond to each other. Typically, the basic unit chosen for alignment is the paragraph. So long as each paragraph of the source text corresponds to exactly one paragraph of the target text, the process of alignment is relatively straightforward and can be automated with quite accurate results. However, there are always difficulties that complicate the alignment process, such as one paragraph that is broken into two paragraphs in the translation. An aligned bilingual corpus is sometimes called a "bitext." Clearly, if discourse strategies differ greatly between two languages, and the order of corresponding elements differs dramatically between source and target texts, automatic alignment may not be feasible.

A statistics-based translation begins with a large bitext. A huge amount of computation follows. Alignment of the bitext is performed not just to the paragraph level but on down to the word and phrase level. Surprisingly, although consistent with the name of the approach, this low-level alignment is not done using a bilingual dictionary. Instead, it is done using purely statistical techniques that formulate many hypotheses about how words correspond and that test those hypotheses on the bitext. Significantly, the statistics-based approach does not limit itself to a one-to-one correspondence between words. For example, the French expression *avoir faim* may be correctly linked to the English expression *be hungry*. Such sophisticated processing allows a statistics-based approach to produce results superior to what can be achieved in a mechanical word-for-word translation. However, it is now generally agreed that the statistics-based approach has reached its limits and that an integration of statistical linguistics with mainstream morphology and syntax is the next logical step.

*Artificial-Intelligence-Based Machine Translation*

Another approach to machine translation is to use techniques of artificial intelligence (AI). However, AI is not usually considered to be part of linguistics, and therefore such systems are outside the realm of applied linguistics. Although none of the commercial machine translation systems currently on the market are primarily based on AI techniques (marketing claims notwithstanding), there are serious, potentially commercial Al-based systems being developed at Carnegie-Mellon University and at New Mexico State University (Computing Research Laboratory).

*7.4. How linguistic theory is applied to translation tools*

The first use of computers by translators was the *word processor*. In the early 1980s, the most popular ways of getting a translation onto paper were: (1) dictation and subsequent transcription by the secretary and (2) typing on a typewriter. Now there has been a major shift toward using word processing software on personal computers. The driving forces were naturally cost and speed. With the introduction of commercial computerized voice-to-text systems, some translators are shifting back to dictation but without a secretary – even though such systems are error-prone and require an unnatural pause between every pair of words.

Some of the most widely-used translator tools are fax machines, electronic mail and other telecommunications, word-count software, and terminology management. By accepting a source text that arrives on a fax machine and producing a translation in a word-processing file that is then transmitted electronically back to the requester, a translator five thousand miles away can compete effectively with a local translator. This near elimination of distance as a factor has made it easier to match translator and text appropriately. Word-count software allows translators to bill more quickly and accurately. Terminology management allows increased quality through greater consistency in the use of specialized terms.

*Terminology management*, however, is the least used of the tools mentioned so far and reflects a period of transition in its application. Terminology management is gradually being separated from the translation task itself. As requesters better

understand that translation is not a black box which produces the one and only correct translation when presented with a naked source text, the requesters of translation are becoming more involved in supplying organization-specific terminology with the text to be translated. Translators also now discuss up front with the requester how terminology will be managed and who will own it and pay for it.

There is another set of translator tools which are coming into use. However, they require that the source text be available in machine-readable form. A machine-readable source text is typically a word-processing file where each character is represented as a code in a character set rather than as a pattern of dots as in a fax. Such tools are sometimes called *Level-Two tools*, as opposed to the *Level-One tools* already mentioned (which can be used when the source text is supplied on paper or by fax). Some of the most important Level-Two tools are terminology-research, automatic-lookup, translation-memory, and quality-check tools. All but automatic lookup require a bitext (an aligned bilingual corpus).

Terminology research allows one to retrieve all occurrences of a term and how it has been translated by presenting to the human for analysis the source-target pairs in which the requested term appears in the source element of the pair. Translation memory is especially useful in translating a revision of a previously translated text. It allows the translator to focus on just the segments that have been modified in the revision and facilitates the re-use of unmodified segments of the translation. Quality-check software, which is not yet in commercial form, looks for common translation errors and verifies that standardized equivalents are used for crucial terms. Automatic lookup displays the terms from the current segment of source text that are found in a bilingual file of standardized terms and allows easy insertion of the standard equivalent without typing it in.

What is remarkable about the technology used on translator tools is that it is mostly restricted to morphology. Syntax and semantics do not currently play a large role in translator tools. The human performs the central translation task. The computer retrieves information much more quickly than a human could, but the human interprets the information and decides how to use it.

*7.5. Machine translation today*

The importance of language understanding for adequate translation is illustrated by the following examples:

1. *Julia flew and crashed the airplane.*

   Julia (flew and crashed the airplane)

   (Julia flew) and (crashed the airplane)

2. *Susanne observed the yacht with a telescope.*

   *Susanne observed the man with a beard.*

3. *The mixture gives off dangerous cyanide and chlorine fumes.*

(dangerous cyanide) and (chlorine fumes)

dangerous (cyanide and chlorine) fumes

The first example is ambiguous between using the verb *fly* transitively (*someone flies an airplane*) or intransitively (*someone/-thing flies*). The second example provides a choice between an adnominal and an adverbial interpretation of the prepositional phrase. The third example exhibits a scope ambiguity regarding dangerous. A human translator recognizes these structural ambiguities, determines the intended reading, and recreates the proper meaning in the target language.

A second type of problem for translation without understanding the source language arises from lexical differences between source and target language. Compare the examples:

1. *The men killed the women. Three days later they were caught.*
   *The men killed the women. Three days later they were buried.*
2. *The watch included two new recruits that night.*

When translating example 1 into French, it must be decided whether they should be mapped into *ils* or *elles* - an easy task for someone understanding the source language. Example 2 shows a language-specific lexical homonymy. For translation, it must be decided whether *watch* should be treated as a variant of *clock* or of *guard* in the target language.

A third type of problem arises from syntactic differences between the source and the target language:

- German:

*Auf dem Hot sahen wir einen kleinen Jungen, der einem Ferkel nachlief.*

*Dem Jungen folgte ein großer Hund.*

- English:

*In the courtyard we saw a small boy running after a piglet.*

*a. A large dog followed the boy.*

*b. The boy was followed by a large dog.*

German with its free word order can front the dative *dem Jungen* in the second sentence, providing textual cohesion by continuing with the topic. This cannot be precisely mirrored by the English translation because of its fixed word order. Instead one can either keep the active verb construction of the source language in the translation (a), losing the textual cohesion, or one can take the liberty of changing the construction into passive (b). Rhetorically the second choice would be preferable in this case. A fourth type of problem is caused by the fact that sequences of words may become more or less stable in a language, depending on the context of use. These fixed sequences range from frequently used 'proverbial' phrases to collocations and idioms.

In light of these difficulties, many practically oriented researchers have turned away from the goal of fully automatic high quality translation (FAHQT) to work instead on partial solutions which promise quick help in high volume translation.

The following partial solutions for practical machine translation may be suggested:

1. *Machine-aided translation (MAT)* supports human translators with comfortable tools such as on-line dictionaries, text processing, morphological analysis, etc.
2. *Rough translation* - as provided by an automatic transfer system – arguably reduces the translators' work to correcting the automatic output.

3. *Restricted language* provides a fully automatic translation, but only for texts which fulfill canonical restrictions on lexical items and syntactic structures.

Systems of restricted language constitute a positive example of a smart solution. They utilize the fact that the texts to be translated fast and routinely into numerous different languages, such as maintenance manuals, are typically of a highly schematic nature. By combining aspects of automatic text generation and machine translation, the structural restrictions of the translation texts can be exploited in a twofold manner.

First, an on-line text processing system helps the authors of the original text with highly structured schemata which only need to be filled in (text production). Second, the on-line text system accepts only words and syntactic constructions for which correct translations into the various target languages have been carefully prepared and implemented (machine translation).

The use of restricted language may be compared to the use of a car. To take advantage of motorized transportation, one has to stay on the road. In this way one may travel much longer distances than one could on foot. However, there are always places a car cannot go. There one can leave the car and continue on foot.

Similarly, due to their automatic input restrictions, systems of restricted language provide reliable machine translation which is sufficiently correct in terms of form and content. If the text to be translated does not conform to the restricted language, however, one may switch off the automatic translation system and look for a human translator.

Besides these smart partial solutions, the solid goal of fully automatic high quality translation (FAHQT) for nonrestricted language has not been abandoned. Today's theoretical research concentrates especially on the Interlingua approach, including knowledge-based systems of artificial intelligence. In contrast to the direct and the transfer approach, the Interlingua approach does not attempt to avoid semantic and pragmatic interpretation from the outset.

The Interlingua approach uses a general, language-independent level called the Interlingua. It is designed to represent contents derived from different source languages in a uniform format. From this representation, the surfaces of different target languages are generated.

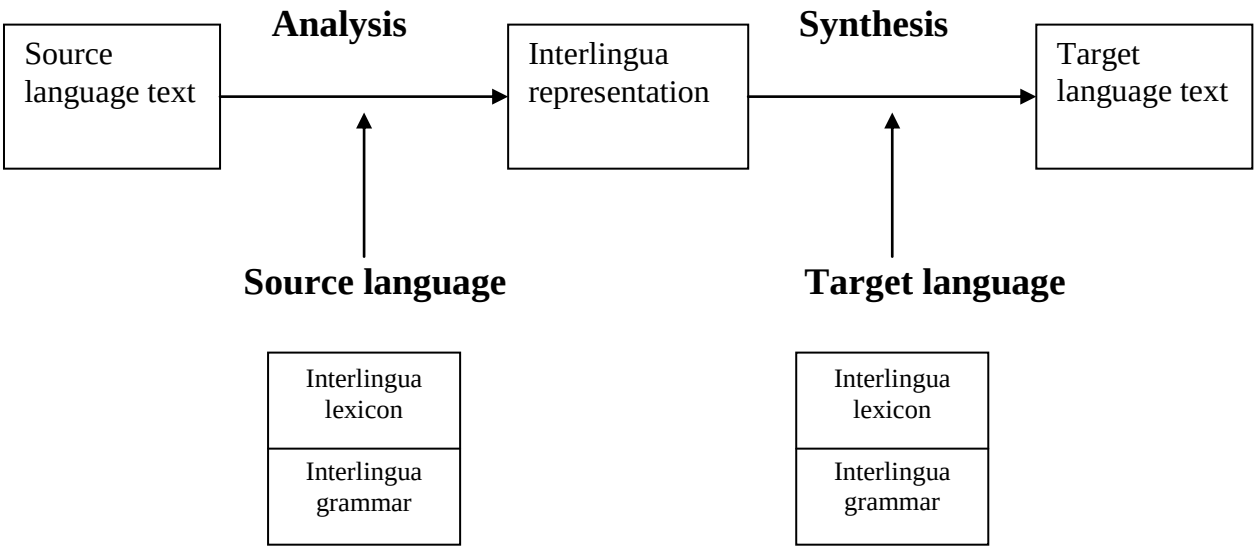



Figure 4.3. Schema of the Interlingua approach

An Interlingua system (viewed schematically in fig. 4.3.) handles translation in two independent steps. The first step translates the source language text into the Interlingua representation (analysis). The second step maps the Interlingua representation into the target language (synthesis).

It follows from the basic structure of the interlingua approach that for n(n — 1) language pairs only In interlingua components are needed (namely n analysis and n synthesis modules), in contrast to the direct and the transfer approach which require n(n — 1) components. Thus, as soon as more than three languages (n > 3) are involved, the Interlingua approach has a substantial advantage over the other two.

The crucial question, however, is the exact nature of the Interlingua. The following interlinguas have been proposed:

- an artificial logical language;

- a semi-natural language like Esperanto which is man-made, but functions like a natural language;
- a set of semantic primitives common to both the source and the target language, serving as a kind of universal vocabulary.

Closer inspection shows, however, that these proposals have not yet resulted in theoretically and practically acceptable results. Existing interlingua systems are highly experimental, usually illustrating theoretical principles by translating tiny amounts of data by means of huge systems.

The experimental character of these attempts is not surprising because a general solution to Interlingua translation may almost be equated with modeling the mechanism of natural language communication. After all, Interlingua translation requires (1) a language-independent representation of cognitive content, (2) the automatic translation of the natural source language into the language-independent Interlingua, and (3) the automatic generation of the natural target language from the Interlingua.

Conversely, as soon as natural communication has been modeled on the computer in a general way, fully automatic high quality translation (FAHQT) is within reach. At the same time, all the other applications of computational linguistics, such as human-computer communication in natural language, a concept-based indexing of textual databases with maximal recall and precision, etc., can be provided with solid solutions, on the basis of available off-the-shelf modules.

## Conclusion

Some lessons to be learned from machine translation and translator tools are that the human mind is extremely flexible and that our linguistic theories are sorely limited. What is commonly called machine translation of natural language may not be that at all. Machine translation produces output most like human translation when processing 'controlled language.' But controlled language may be closer to formal

language than to natural language. Perhaps there is only a superficial resemblance between the two. Machine translation of controlled language may appear to be similar to human translation, but that does not mean that humans and machines are using the same process when translating. Indeed, the fact that machine-translation techniques that work well on controlled language break down when applied to more general and dynamic texts indicates that the process is not the same. Humans are willing to adapt to the formal language recognized by a machine when useful results are forthcoming. We control our VCRs and microwave ovens and automobiles by giving commands in an extremely limited and controlled formal "language." This does not necessarily mean that these devices understand natural language.

Machine translation reminds us that *Applied Linguistics* is applied and that there must be something to apply. The current use of machine translation allows for high-quality translation of controlled language but only indicative translation of natural language. This level of progress suggests that current mainstream linguistics may not have captured the essence of natural language beyond the realms of morphology and syntax.

Perhaps machine translation is an example of Applied Linguistics performing a significant service to Theoretical Linguistics by pointing out where theory is lacking. In the past, it was expected that machines would replace human translators. Not only are human translators benefiting from machines, human linguists can benefit from a demonstration that linguistics is not a completed task. Earlier in this century, it was thought that physics was nearly complete and that theoretical physicists would no longer be needed for research. Since then relativity theory and quantum mechanics have brought in a flowering of research in physics. Let us hope that substantial breakthroughs are also on the horizon for linguistics.

# CHAPTER V.
# NATURAL LANGUAGE PROCESSING

## 1. Challenges for Natural Language Processing

Although natural language processing (NLP) has come very far in the last twenty years, the technology has not yet achieved a revolutionary impact on society. These are the problems that most need additional research and most deserve the attention of scholars:

1. *Knowledge acquisition* from natural language (NL) texts of various kinds, from interactions with human beings, and from other sources. Language processing requires lexical, grammatical, semantic, and pragmatic knowledge. Current knowledge acquisition techniques are too slow and too difficult to use on a wide scale or on large problems. Knowledge bases should be many times the size of current ones.

2 *Interaction with multiple underlying systems* to give NL systems the utility and flexibility demanded by people using them. Single application systems are limited in both usefulness and the language that is necessary to communicate with them.

3. *Partial understanding* gleaned from multi-sentence language, or from fragments of language. Approaches to language understanding that require perfect input or that try to produce perfect output seem doomed to failure because novel language, incomplete language, and errorful language are the norm, not the exception.

The *limitations* of today's practical language processing technology may be summarized as follows:

1. *Domains must be narrow enough* so that the constraints on the relevant semantic concepts and relations can be expressed using current knowledge representation techniques, i.e., primarily in terms of types and sorts.

Processing may be viewed abstractly as the application of recursive tree rewriting rules, including filtering out trees not matching a certain pattern.

2. *Handcrafting is necessary*, particularly in the grammatical components of systems (the component technology that exhibits least dependence on the application domain). Lexicons and axiomatizations of critical facts must be developed for each domain, and these remain time-consuming tasks.
3. *The user must still adapt to the machine*, but, as the products testify, the user can do so effectively.
4. *Current systems have limited discourse capabilities* that are almost exclusively handcrafted. Thus current systems are limited to viewing interaction, translation, and writing and reading text as processing a sequence of either isolated sentences or loosely related paragraphs. Consequently, the user must adapt to such limited discourse.

It is traditional to divide natural language phenomena (and components of systems designed to deal with them) into *three classes*:

1. *Syntactic phenomena* - those that pertain to the structure of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning.
2. *Semantic phenomena* - those that pertain to the meaning of a sentence relatively independent of the context in which that language occurs.
3. *Pragmatic phenomena* - those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue), or nonlinguistic (such as knowledge about the person who produced the language, about the goals of the communication, about the objects in the current visual field, etc.).

*Syntax* is without doubt the most mature field of study in both computational linguistics and the closely related field of linguistics. The most thorough computational accounts of natural language phenomena exist for syntax; and grammars

with very large coverage of English have existed since the early 1980s. Formalisms for describing syntactic phenomena, mathematical analyses of the expressive power of those formalisms, and computational properties of processors for those formalisms have existed for more than twenty-five years, since the definition of the N.Chomsky hierarchy (finite state languages, context-free languages, context-sensitive languages, and the recursively enumerable languages).

During the 1970s and most of the 1980s, the dominant NLP formalism for writing grammars of natural language was the augmented transmission network (ATN) (W.A.Woods 1970), a procedural language that allowed compact statements of not only the context-free aspects of language but also the apparently context-sensitive aspects as well. In the late 1980s, a shift began from the ATN and its procedural aspects toward a declarative formalism. What the new dominant formalism will be is not yet clear, but a likely candidate is a class of grammar formalisms that combine context-free rules with unification as a way of compactly capturing both context-free and context-sensitive aspects of language. The declarative formalisms offer the promise of exploring alternative parsing algorithms and pose minimum constraints on parallel algorithms for parsing. This shift in the research community from procedural specifications of syntax, such as ATN grammars, to declarative specifications, such as unification grammars, parallels the similar shift in interest in programming language research away from procedural languages and toward newer functional programming languages and declarative representations.

Because syntax is by far the most mature area in natural language processing, it is difficult to foresee that further developments in syntax will have as great an impact on utility as would emphasis on research and development on other, less developed areas of technology.

In *semantics*, much recent progress has been made by focusing on limited application domains. For database access, the semantics of the system can be confined to that of the individual entities, classes of entities, relationships among the entities, attributes of the entities, and the typical operations that are performed in database retrieval. This simplifies the problem of semantics in at least the following ways:

first, the meaning of individual words and of the phrases they compose can be restricted to the domain-specific meanings actually modelled in the database. Instead of needing to come up with a very general semantics for each word, a very literal semantics providing the mapping from the words to the entities modelled in the database is all that is required, for the database could not provide additional information even if a more general semantics were available. Second, problems of semantic ambiguity regarding alternative senses for a word are reduced, for only those word senses corresponding to entities in the database will contribute to the search space of possible alternatives.

For the task of database updating from messages, a key simplifying condition is that the information sought can be characterized ahead of time. Suppose the goal is to update automatically a database regarding takeover bids. Suppose further that the information desired is the date of the bid, the bidder, the target, the percentage of stock sought, the value of the offer, and whether it is a friendly or hostile bid. A first approximation is that the remaining information in the article can be ignored. The assumption is that although other concepts may be mentioned in the message or news wire item, they normally do not impact the values of the fields to be updated in the database. If that assumption applies in the proposed message processing application, then one can model the literal semantics of those words and phrases that have a correlate in the data being sought. For cases where the proposed update would be in error because the unanalyzed text does impact the update, human review of the proposed update can minimize erroneous entries into the database. Semi-automatic update with a human in the loop may be more cost effective and timely than fully manual update and may be more desirable than not having the data at all.

No uniform semantics representation language has emerged, although three general classes of semantic representations are employed: those that allow one to state anything that arises in a prepositional logic, those that allow expressions equivalent to a first-order logic, and those that allow expressions not representable in a first-order logic. Most framed-base representations are equivalent to a prepositional logic, since they do not allow expression of quantifiers. Most systems for database

retrieval use a first-order logic. Many research systems employ extensions beyond first-order logic, such as the modal intensional logic defined by R.Montague (1970). Encoding the semantics of all the words and phrases for a particular application domain is one of the most significant costs in bringing up a natural language system in a new application domain. Knowledge acquisition procedures that would reduce this cost would therefore have great impact on the applicability of the technology.

In *pragmatics* the modelling of context and using context in understanding language is the most difficult, and therefore the least well-understood, area of natural language processing. Unlike programming languages where one can define contextual influence in a limited and controlled way, context is all-pervasive and very powerful in natural language communication.

Context is fundamental to communicating substantial information with few words. For instance, if one says, *How about Evans?*, those three words may suggest a lot. If the context had been that the immediately previous request was *List the salary, highest degree, race, and marital status of Jones*, then *How about Evans?* means *List the salary, highest degree, race, and marital status of Evans*. If the context has been your boss saying *I need someone to go to Phoenix next week without jeopardizing meeting the XYZ deadline*, then *How about Evans?* means *Consider Evans for going to Phoenix next week without jeopardizing meeting the XYZ deadline*.

The single phenomenon that has received the most attention in pragmatics is pronominal or other referring expressions. Progress has been substantial enough that pronouns *it, they, those,* etc.) and definite reference (*those submarines, the first three men*, etc.) can be used rather freely in today's systems.

## 2. Knowledge acquisition for language processing

### *2.1. Types of knowledge*

Typically, porting a NLP system to a new domain requires acquiring knowledge for the domain-dependent modules, which often include:

- *Domain model*. The major classes of entities in the domain and the relations among them must be specified. In a Navy command and control domain, example concepts are *Naval unit, vessel, surface vessel, submarine, carrier, combat readiness ratings, and equipment classes*. Class-subclass relationships must be specified, e.g., every carrier is a surface vessel, and every surface vessel is a vessel and a Naval unit. Other important relationships among concepts must be specified. For instance, each vessel has a single overall combat readiness rating, and each Navy unit has an equipment loadout (a list of equipment classes).
- *Lexical syntax*. Syntactic information about each word of the domain includes its part of speech (e.g., noun, verb, adjective, adverb, proper noun), its related forms (e.g., the plural of ship is regular ships, but the plural of sheep and child are irregular sheep and children), and its grammatical properties (e.g., the verb sleep is intransitive). Lexical semantics. For each word, its semantics must be specified as a concept in the domain model, a relation in the domain model, or some formula made up of concepts and relations of the domain model.
- *Mappings to the target application*. Transformations specify how to map each concept or relation of the domain model into an appropriate piece of code for the underlying application system. For example, to find out whether a given vessel is equipped with helicopters, one might have to check whether there is a "Y" in the HELO field of the VES table of the database.

Currently, domain-independent knowledge is usually hand-built and is not re-acquired when moving to a new domain, although it may be necessary to "tweak" rules and extend this knowledge, again, often by hand. It includes:

- *Grammar rules*. Most rules of English grammar are domain independent, but almost every domain encountered in practice either turns up instances of general tries that had not been encountered in previous domains, or requires that some domain-specific additions be made to the grammar.

- *General semantic interpretation rules*. Some semantic rules may be considered to be domain independent, such as the general entity/property relationship that is often expressed with the general verb "have" or the general preposition "of." To the extent that such general rules can be found and embedded in a system, they do not have to be redone for every new domain.

The success of all current NLP systems depends on the so-called *the Limited Domain Assumptio*n, which may be stated as follows: one does not have to acquire domain-dependent information about words that do not denote some concept or relation in the domain. Another way of looking at this assumption is that it says understanding can be confined to a limited domain. The Limited Domain Assumption simplifies the problem of NLP in three ways: (1) formal modelling of the concepts and relationships of the domain is feasible, (2) enumeration of critical non-linguistic knowledge is possible, and (3) both lexical and semantic ambiguity is limited. Reducing lexical ambiguity reduces the search space and improves effectiveness of most NL systems.

Those three facts have the combined effect of making it more tractable to determine what the user meant by a given input, among a welter of possibilities. But whether one tries to loosen the domain restrictions or is willing to live within them; it seems obvious (although we will examine this assumption later) that the more knowledge that is available to the system, the better its chances of understanding its input.

### *2.2. Types of knowledge acquisition*

Just as there are many kinds of knowledge, there are a number of different ways of acquiring that knowledge. M.Bates and R.M.Weischedel (1993) suggest the following types:

- *Knowing by being pre-programmed* - this includes such things as hand-built grammars and semantic interpretation rules.

- *Knowing by being told* - this includes things that a human can "tell" the system using various user-interface tools, such as semantic interpretation rules that can be automatically built from examples, selectional restrictions, and various lexical and morphological features.
- *Knowing by looking it up* - this means using references such as an online dictionary, where one can find exactly the information that is being sought.
- *Knowing by using source material* - this means using references such as an encyclopedia or a corpus of domain-relevant material, from which one might be able to find or infer the information being sought; it may also mean using large volumes of material as the source of probabilistic knowledge (e.g., "bank" is more likely to mean a financial institution than the side of a river).
- *Knowing by figuring it out* - this means using heuristics and the input itself (such as the part of speech of words surrounding an unknown word).
- *Knowing by using a combination of the above techniques* - this may or may not involve human intervention.

The ways of learning do not necessarily correspond to the types of knowledge in any direct way. Certainly all of the types of knowledge can be pre-programmed into an NLP system; indeed that is how most of the current systems were created. It is a fairly simple step from that to learning by being told - usually all that is needed is a nice user interface for creating the same structures that can be pre-programmed. It is not until we reach the level of knowing by looking it up that it seems right to use the word "learning" to describe what is going on.

The two areas of particular interest here are learning from sources, and learning by figuring it out, or some combination of these with learning by being told by a human being reserved for situations that cannot be covered by the other means.

*Learning by looking it up*

It is hard to learn by looking it up or from sources, but it is going to get easier. On-line dictionaries and other reference books have been available for many years, as have bodies of text such as news wires and technical abstracts, but they have not enjoyed wide usage. Why not? It is not entirely a matter of cost, or speed of access. We believe there are four fundamental reasons why computational linguists have been avoiding these sources:

1. The required information is often not there.
2. Information is hard to extract from the sources.
3. Once extracted, the information is hard to use.
4. The information is often incomplete and/or incorrect.

Most domains use common English words with specialized meanings. For example, most dictionaries contain definitions of the words "virus" and "worm," but not with the meanings that are current in the computer industry. Even if a word is found with its appropriate meaning, the dictionary entry may lack information that is critical to the NL system (e.g., selectional restrictions). And if the word is found in a corpus of source material, how is the meaning to be inferred? As a larger volume of domain-specific material becomes available for many domains, this problem may be reduced, but it will always be with us.

Extracting detailed information about words or concepts from the kind of text found in dictionaries and encyclopedias is an enticing prospect, but it presents a chicken-and-egg problem. A system cannot read a dictionary or encyclopedia entry unless it knows all the words in the definition (and usually a great deal more). Since language of this type is often beyond the capabilities of NL systems (particularly those built on the premise that the input and output must be complete), NLP systems typically cannot read the reference material. One solution to this problem is to pre-process the reference material, as is being done by Mitch Marcus at the University of Pennsylvania in an effort to produce text roughly annotated with part of speech and syntactic structure.

Another solution is to relax the constraints on input and output of NLP systems, and to develop *partial understanders* that can glean some information from sources

and, using that information, can re-read the sources to increase their understanding by bootstrapping. Recent work by Will Crowther (1989) has taken this approach quite successfully.

*Learning from sources*

Does knowing more mean that understanding is easier, or harder? Suppose we solve the problem of extracting information about words and other things from reference books. Will that automatically mean that our NLP systems will perform better? There is strong evidence that this is not the case – because the increased lexical, syntactic, and semantic alternatives that are introduced by knowing, for example, all the parts of speech and all the possible meanings of all the words in a sentence can easily swamp a NL processor with an explosion of possible alternatives to explore, and irresolvable ambiguities may arise when exploring even just a few.

The last major reason for avoiding source material is that such sources, massive as they are, are inevitably incomplete and incorrect. Nearly all NLP systems deal with specific limited domains, generally rather technical domains (weather forecasts, Navy messages, banking, etc.) in which ordinary English words are given special or restricted meanings. Thus general sources such as dictionaries give meanings that are misleading or actually wrong, but the NLP system has no way of knowing this. It would be far better for the sources to have no information than to have the wrong information, but that is not realistic or even remotely possible.

The conclusion is that dictionary and other source information will not be useful unless we learn how to focus NL processing, order meanings and partially understood phrases, and interact with other knowledge sources (including humans) when necessary. Fortunately, there are several ways of achieving these goals, including:

1. Representing ambiguity at many levels of processing in a computationally tractable way.
2. Using statistical probabilities at many levels to order choices and cut off low likelihood paths.

*Learning by being told*

Some situations will always call for learning by being told. To illustrate this, consider the following sentence:

*Sebastian compensated his Glock.*

Do you know what that means? What can you figure out, and how? Presumably you know that Sebastian is a male's name, although if you did not know that, you might find it out by consulting a good dictionary with a list of names. You already know the verb "*compensate*" (or can look it up), with meanings roughly comparable to "*pay*" and "*make up for*"; the latter meaning is unlikely since it requires a for-clause. The word "*Glock*" is a stumper. You are unlikely to find it in any dictionary or encyclopedia you have handy. It seems to be a proper noun, judging from the capitalization. You might guess that it is a person's name, although the use of the possessive pronoun with a proper name is quite unusual, and would probably carry some special meaning that cannot be figured out from the sentence itself. Perhaps you have some other hypothesis about the word "*Glock*". The point is, without help from a human being knowledgeable about the subject area (or an extremely specialized dictionary), you are unlikely to figure out what that sentence means, even with considerable effort.

Adding context is not necessarily a help. Suppose the sentence had come to you as part of a message which said, in its entirety:

*Henry and Sebastian were rivals, each preparing for the upcoming competition in his own way. In order to improve his chances, Henry practiced hard. Sebastian compensated his Glock. Lyn didn't think this would help, and advocated more practice instead, but Sebastian pursued his plan single-mindedly.*

There is quite a lot of information in that paragraph, but nothing that is very helpful in figuring out about compensating *Glocks*.

But if you are told that *Glock* is a firearms manufacturer (and therefore can be used to refer generically to any firearm they produce, as is the case with *Colt*), and that certain guns can have a device called a compensator installed to reduce the recoil when they are fired, then you can probably figure out that *Sebastian compensated his*

*Glock* means that Sebastian had a compensator added to his *Glock* pistol. There is no good alternative to being told this information.

The hard part is not developing rules to infer the meaning of *XXXed* from *XXXor*; such rules have been known for a long time. The hard part is to know when to apply those rules, and how to keep hundreds of those rules from interacting to produce more fog than clarity.

### *2.3. Linguistic analysis of large bodies of text*

A breakthrough in the effectiveness and applicability of knowledge acquisition procedures may be possible within the next five years. In this section the following two research approaches are identified:

1. Employing large, growing knowledge bases acquired from reference texts such as dictionaries. This contributes to robustness by facilitating acquisition of knowledge for semantic and pragmatic components.
2. Acquisition of syntactic, semantic, and discourse facts from annotated bodies of language. This contributes to robustness of syntactic, semantic, and discourse components and allows semi-automatic learning of syntactic and semantic knowledge.

NLP research has been hampered by a lack of sufficient linguistic data to derive statistically significant patterns. Volumes of text are available on-line; the problem has been how to derive linguistic facts from unanalyzed text. Corpora of annotated text will be available to other research sites. The annotations will include parts of speech and phrasal structure, e.g., syntactic structure. This syntactically annotated corpus should make two new developments feasible:

1. Development of acquisition procedures to learn new grammar rules for expressions never seen before by the NLP.
2. Collection of statistics regarding constructions and their probability of occurrence in context.

Automatic acquisition will reduce the need for handcrafting of both grammars and lexicons (the formal model of dictionary information for an NLP).

Original text:

*Collection of statistics regarding constructions and their probability of occurrence in context.*

Part of Speech Tagging:

Collection $_{Noun}$ of$_{prep}$ statistics$_{Noun}$ regarding $_{prep}$ constructions$_{Noun}$ and$_{conjunction}$ their $_{pro}$ probability $_{Noun}$ of$_{prep}$ occurrence $_{Noun}$ inp$_{rep}$ context$_{Noun}$

Structure Tagging:

[Collection [of [statistics [regarding [ [constmctionslNP and [their probability [of [occurrence [in [context]**NP** ]**PP**]**NP**]**PP**]**NP**]**NP**]**PP**]**NP**]**PP**]NP

Any rules implicit in the annotation but not present in the current grammar are candidates to be automatically added to the grammar.

The annotations also allow acquisition of lexical information; for words not in the system dictionary, the annotations state part of speech and the syntactic context in which they occur. Suppose *regarding* were not known to the system before it encountered the annotated example above; this word could be added to the system lexicon as a preposition through processing the annotated example. Thus, annotations provide data that can be used to create systems that adapt by acquiring grammar rules and information about new words.

## 3. Interaction with Multiple Underlying Systems (MUS)

Most current NL systems, whether accepting spoken or typed input, are designed to interface to a single homogeneous underlying system; they have a component geared to producing code for that single class of application systems, such as a relational database (D.Stallard. 1987; *Parlance User Manual*, *Learner User Manual*). These systems take advantage of the .simplicity of the semantics and the availability of a formal language (relational calculus and relational algebra) for the system's output.

The challenge is to recreate a systematic, tractable procedure to translate from the logical expression of the user's input to systems that are not fully relational, such

as expert system functions, object-oriented and numerical simulation systems, calculation programs, and so on. Implicit in that challenge is the need to generate code for non-homogeneous software applications - those that have more than one application system.

The norm in the next generation of user environments will be distributed, networked applications. A seamless, multi-modal, NL interface will make use of a heterogeneous environment feasible for users and, if done well, transparent. Otherwise, the user will be limited by the complexity, idiosyncrasy, and diversity of the computing environment.

Such interfaces will be seamless in at least two senses:

1. The user can state information needs without specifying how to decompose those needs into a program calling the various underlying systems required to meet those needs. Therefore, no seams between the underlying systems will be visible.

2. The interface will use multiple input/output modalities (graphics, menus, tables, pointing, and natural language). Therefore, there should be no seams between input/output modalities.

In military uses, the need to access several heterogeneous application systems will arise as the norm in command and control, in logistics, and in contract management. Because of the need to include previously existing application software, each having its own assumptions regarding operating systems, heterogeneous software environments will arise. Because of the relative performance-cost trade-offs in workstations, mainframes, and parallel hardware, the hardware equipment will be heterogeneous as well.

For example, in *DARPA's Fleet Command Center Battle Management Program* (FCCBMP), several applications (call them *underlying systems*) are involved, including a relational database (IDB), two expert systems (CASES and FRESH), and a decision support system (OSGP). The hardware platforms include workstations, conventional time-sharing machines, and parallel mainframes. Suppose the user asks *Which of those submarines has the greatest probability of locating A within 10 hours?* Answering that question involves sub problems from several underlying

applications: the display facility (to determine both what those submarines means and to display those which fulfill the user's request); FRESH to calculate how long it would take each submarine to get to the area A; CASES, for an intensive numerical calculation estimating the probabilities; and the display facility again, to present the response.

Although acoustic and linguistic processing can determine what the user wants, the problem of translating that desire into an effective program to achieve the user's objective is a challenging, but solvable problem.

In order to deal with multiple underlying systems, not only must our NL interface be able to represent the meaning of the user's request, but it must also be capable of organizing the various application programs at its disposal, choosing which combination of resources to use, and supervising the transfer of data among them. It is called *the Multiple Underlying Systems (MUS) problem*.

In order to access the multiple underlying systems, the user's request, whatever its modality, is translated into an internal representation of the meaning of what the user needs. A first-order logic is applied for this purpose. An intensional logic may also be used to investigate whether intensional logic offers more appropriate representations for applications more complex than databases, e.g., simulations and other calculations in hypothetical situations. From the statement of what the user needs, the system next derives a statement of how to fulfill that is needed – an executable plan composed of abstract commands. The executable plan is in essence an abstract data-flow program on a virtual machine that includes the capabilities of all of the application systems. At the level of that virtual machine, specific commands to specific underlying systems are dispatched, results from those application systems are composed, and decisions are made regarding the appropriate presentation of information to the user. Thus, the Multiple Underlying Systems (MUS) problem is a mapping,

MUS: Semantic representation → Program

That is, a mapping from what the user wants to a program to fulfill those needs, using the heterogeneous application programs' functionality.

Although the statement of the problem as phrased above may at first suggest an extremely difficult and long-range program of research in automatic programming (C.Rich and R.C.Waters 1988), there are several ways one can narrow the scope of the problem to make utility achievable. Substantially restricting the input language is certainly one way to narrow the problem to one that is tractable.

A way to paraphrase the effect of assuming acyclic data-flow graphs as the output of the component is that the programs generated will be assumed to include:

- Functions available in the underlying applications systems,
- Routines pre-programmed by the application system staff,
- Operators on those elements such as: functional composition, if-then-else, operators from the relational algebra.

Therefore, the system need not derive programs for terms that it does not already know. Contrast that with the general automatic programming problem. Suppose that someone says to the system *Find the square root of the sum of the squares of the residuals*, so that the input can be correctly translated into a logical form, but that the underlying applications do not provide a square-root function. Then the interface will not be expected to derive a square-root program from arithmetic functions. Rather, this system will be expected to respond *I don't know how to compute square root*. Furthermore, if all the quantifiers are assumed to be restricted to finite sets with a generator function, then the quantifiers can be converted to simple loops over the elements of sets, such as the mapping operators of Lisp, rather than having to undertake synthesis of arbitrary program loops.

Even with these simplifying assumptions, there are interesting problems remaining, and the work offers highly desirable utility. The utility arises from two dimensions:

1. It frees the user from having to identify for each term (word) pieces of program that would carry out their meaning, for the application system programmers would do that for some appropriate set of terms. '

2. It provides good software engineering of the interface, so that table input/output functionality, for instance, is insulated from the details of the underlying application or applications as they evolve.

The problem of multiple systems may be decomposed into the following sub-problems:

*Representation:* It is necessary to represent underlying system capabilities in a uniform way, and to represent the user request in a form independent of any particular underlying system. The input/output constraints for each function of each underlying system must be specified, thus defining the services available.

*Formulation*: One must choose a combination of underlying system services that satisfies the user request. Where more than one alternative exists, it is preferable to select a solution with low execution costs and low passing of information between systems.

*Execution:* Actual calls to the underlying systems must be accomplished, information must be passed among the systems as required, and an appropriate response must be generated.

The example of MUS is *Janus* created by R.M.Wieschedel (R.M.Weischedel 1988). Since the meaning of an utterance in *Janus* is represented as an expression in WML (*World Model Language* (E.W.Hinrichs et al., 1987], an intensional logic, the input to the MUS component is in WML. The choice of WML was based on two grounds: first and foremost, although we found first-order representations adequate (and desirable) for NL interfaces to relational databases, as a richer semantic representation was important for future applications. The following classes of representation challenges appeared to be essential:

- explicit representations of time and world, for instance, to support object-oriented simulation systems and expert systems involving hypothetical worlds;
- distributive/collective readings; generics, and mass terms;
- propositional attitudes, such as statements of user preference and belief.

The motivation for choosing intensional logic was the necessity to capitalize on the advantages of applying intensional logic to natural language processing (NLP), such as the potential simplicity and compositionality of mapping from syntactic form to semantic representation and the many studies in linguistic semantics that assume some form of intensional logic.

For a sentence such as *Display the destroyers within 500 miles of Vinson*, the WML is as follows:

```
(bring about
  ((intension
    (exists ?a display
      (object-of ?a
        (iota ?b  (power destroyer)
          (exists ?c
            (lambda (/d) interval
              (& (starts-interval ?d VINSON)
                (less-than
                    (iota ?e length-measure
                        (interval-length ?d ?e))
                    (iota ?f length-measure
                        (& (measure-unit ?f miles)
                            ( measure-quantity ?f 500))))))
                (ends-interval ?c ?b))))))
  TIME WORLD))
```

To represent the functional capabilities of underlying systems, such notions as *services* and *servers* are used. A *server* is a functional module typically corresponding to an underlying system or a major part of an underlying system. Each server offers a number of *services*: objects describing a particular piece of functionality provided by a server. Specifying a service in MUS provides for the mapping from fragments of logical form to fragments of underlying system code. For instance, the following is a list of services in a naval application. Each service has

associated with it the server it is part of, the input variables, the output variables, the conjuncts computed, and an estimate of the relative cost in applying it.

```
Land-avoidance-distance:
    owner: Expert System 1
    inputs: (x y)
    locals: (z w)
    pattern:
      ((in-class x vessel)
       (in-class y vessel)
       (in-class z interval)
       (in-class w length-measure)
       (starts-interval z x)
       (ends-interval z y)
       (interval-length z w))
  outputs: (w)
  method: ((route-distance (location-of x) (location-of y)))
  cost: 5
Great-circle-distance:'
    owner: Expert System 1
    inputs: (x y)
    locals: (z w)
    pattern:
      ((in-class x vessel)
       (in-class y vessel)
       (in-class 7. interval)
       (in-class w length-measure)
       (starts-interval z x)
       (ends-interval z y)
       (interval-length z w))
  outputs: (w)
```

method: ((gc-distance (location-of x) (location-of y)))

cost: I

In the example above, there are two competing services for computing distance between two ships: *Great-circle-distance*, which simply computes a great circle route between two points, and *Land-avoidance-distance*, which computes the distance of an actual path avoiding land and sticking to shipping lanes.

Usually, the applicability of a service is contingent on several facts, and therefore several propositions must all be true for the service to apply. To facilitate matching the requirements of a given service against the needs expressed in an utterance, the expressions in WML are converted to a *disjunction normal form* (DNF), i.e., a disjunction of conjunctions where quantifiers and higher level operators have been removed. The advantages of DNF are:

1) In the simplest case, an expression in disjunctive normal form is simply a conjunction of clauses, a particularly easy logical form with which to cope.

2) Even when there are disjuncts, each can be individually handled as a conjunction of clauses, and the results then combined together via union.

3) In a disjunctive normal form, each disjunct effectively carries all the information necessary for a distinct subquery.

If one takes the input request to be a conjunction of requirements, finding the services to fulfill the request may be viewed as a form of covering problem: one seeks a plan of execution that satisfies all requirements at minimal cost.

A search is required both to find collections of services that fulfill the request, and to find a low cost solution. A beam search is used.

Inherent in the collection of services covering a DNF expression is the data flow that combines the services into a program to fulfill the DNF request. The next step in the formulation process is data-flow analysis to extract the data-flow graph corresponding to an abstract program fulfilling the request.

The data-flow graph for *Display the destroyers within 500 miles of Vinson* is presented in fig.5.1:

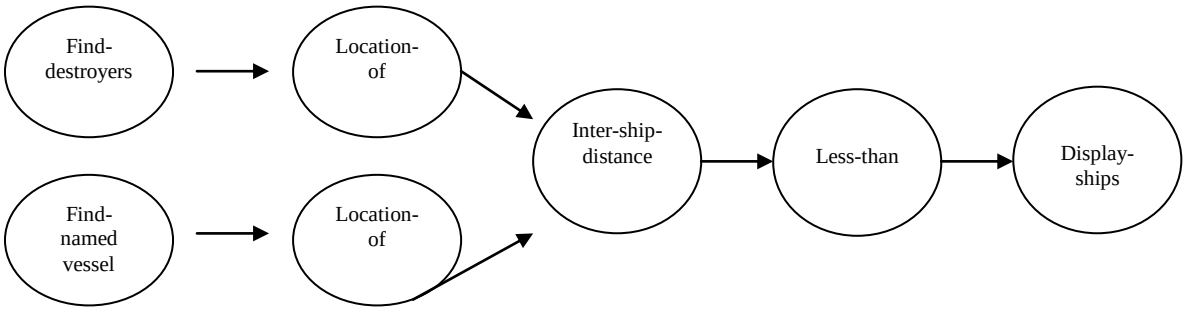



Figure 5.1. The data-flow graph for *Display the destroyers within 500 miles of Vinson*

The execution phase for the represented multiple underlying system has two purposes:

1. Walk through the data-flow graph, calling operators in the underlying systems corresponding to the nodes of the graphs

2. Supply functions for data combination not available in any of the underlying systems. In the example above, a general function for comparing two measures, performing the appropriate unit conversions was assumed.

The MUS component has been applied in the domain of the Fleet Command Center Battle Management Program (FCCBMP). Integrated Database (IDB) – a relational database – as one underlying resource, and a set of LISP functions as another. The system includes more than 800 services.

An earlier version of the system was also applied to provide natural language access to data Intellicorp's KEE knowledge base system, to objects representing hypothetical world-states in an object-oriented simulation, and to LISP functions capable of manipulating this data.

The MUS components are now being integrated with BBN’s Spoken Language System HARCS (Hear And Respond to Contmuous Speech).

## 4. Partial understanding of fragments, novel language, and errorful language

Modern works in the sphere of Natural Language Processing tend to move away from dependence on the sentence as the fundamental unit of language. Historically, input to NL systems has often had to consist of complete, well-formed sentences. The systems would take those sentences one at a time and process them. But language does not always naturally occur in precise sentence-sized chunks. Multi-sentence input is the norm for many systems that must deal with newspaper articles or similar chunks of text. Subsentence fragments are often produced naturally in spoken language and may occur as the output of some text processing. Even when a sentence is complete, it may not be perfectly formed; errors of all kinds, and new words, occur with great frequency in all applications.

*Multi-sentence input.*

Historically, computational linguistics has been conducted under the assumption that the input to a NL system is complete sentences (or, in the case of speech, full utterances) and that the output should be a complete representation of the meaning of the input. This means that NL systems have traditionally been unable to deal well with unknown words, natural speech, language containing noise or errors, very long sentences (say, over 100 words), and certain kinds of constructions such as complex conjunctions.

One of the problems is that advocates of local processing have tended to ignore syntactic and other constraints, while advocates of top down processing have tended to ignore coherent fragments unless they fit properly in the overall scheme.

The solution is to move away from thinking that language comes in sentences and that the goal of understanding is a complete representation of meaning. Users must move toward processing bits and pieces of language, whether the input to our

NL systems comes that way or not, and toward creating structures that, like the fractals found in nature, have a kind of coherency that can be viewed at many levels.

Some semantic distinctions have no selectional import (e.g., quantifiers, and some adjuncts), while others have considerable selectional import.

One of the ideas whose time has passed is the notion of prepositional phrase attachment. Although in many cases it is not harmful to think of a PP attaching to a particular constituent, sometimes it is more useful to think of a single PP attaching simultaneously at several different points (for example, "*I kicked the shell on the beach*"), or relating two different constituents in a sentence (for example, "*The average concentration of aluminum in breccias*"). When fixed constituent structure pinches too much, language should not be forced into it.

The "right" representation for text may depend on the type of text and its purpose. For example, commands may be represented very differently from questions. It may also depend on the purpose of the user: for example, question answering versus controlling a process versus storing information for later retrieval.

Currently, most systems that attempt to understand multi-sentence text create a frame as output (or some other structure that is similar in function). Generally, the names of the slots of the frame consist of the type of information and relationships that were to be gleaned from the text, and the fillers describe the entities that were found. Thus it is difficult to represent unexpected information.

*Errorful language, including new words.*

Handling novel, incomplete, or errorful forms is still an area of research. In current interactive systems, new words are often handled by simply asking the user to define them. However, novel phrases or novel syntactic/semantic constructions are also an area of research. Simple errors, such as spelling or typographical errors resulting in a form not in the dictionary, are handled in the state-of-the-art technology, but far more classes of errors require further research.

The state-of-the-art technology in message understanding systems is illustrative. It is impossible to build in all words and expressions ahead of time. As a

consequence, approaches that try for full understanding appear brittle when encountering novel forms or errorful expressions.

The state of the art in spoken language understanding is similarly limited. New words, novel language, incomplete utterances, and errorful expressions are not generally handled. Including them poses a major roadblock, for they will decrease the constraint on the input set, increase the perplexity (a measure of the average number of words that may appear next at any point in the input) of the language model, and therefore decrease reliability in speech recognition.

There is ample evidence that the ability to deal with novel, incomplete, or errorful forms is fundamental to improving the performance users can expect from NLP systems. Statistical studies for written database access (C.M.Eastman and D.S.McLean, 1981; B.H.Thompson, 1980) show that novel, errorful, or incomplete language comprises as much as 25-30% of type input; such phenomena (V.A.Fromkin, 1973) probably arise even more frequently in spoken language than in written language. In addition, we believe that interpreting incomplete input is particularly important for the following reasons:

Fragments occur frequently in military messages, such as Navy CASREPs, Navy OPREPs, Army SITREPs, and Army Operations Orders.

2. Incomplete input arises in spoken language not only because we speak in fragments but also because acoustic processing at times can detect only fragments with high confidence.

3. Fragments result when processing an incomplete, novel, or errorful input, since a complete interpretation cannot be produced.

In current technology, almost all systems employ a search space of the possible ways of combining the meanings of words into meaningful phrases and a meaningful whole in context. In artificial intelligence terms, the search is a constraint satisfaction problem: find one or more interpretations such that no applicable constraint is violated. Formal models of grammar, semantics, and discourse state constraints on language in an all-or-nothing fashion, as if we always spoke and wrote in complete

thoughts, saying exactly what we mean without vagueness, inaccuracy, error, or novelty in expression.

In constraint satisfaction problems, if a search fails to find a solution where all constraints are satisfied, many search alternatives will have been tried without leading to ultimate success. The problem is to come up with a partial solution (in the case of language processing, a partial interpretation), an explanation of why no solution is found (e.g., why no interpretation can be found), or a way to relax a constraint to produce with a complete solution (a complete interpretation). Which of the partial solutions, if any, is the most likely path to lead to success if a constraint is relaxed? Which partial path(s) in the search space is a good basis for explaining why no solution can be found?

All previous work suffers from this problem mentioned above, unless the application domain is very limited or the types of errorful/novel forms allowed are very few. This is because too many alternatives for what was meant are possible; an NLP system does not even have a foolproof way of knowing whether the user's input is errorful or whether the input represents a novel form. What is not known is how to rank the many alternative interpretations that arise. The lack of a reliable scoring mechanism has been a technological roadblock.

Real language may be *absolutely ill-formed* (a native speaker would judge it to be something to be edited, an error, not what was intended, or otherwise "bad"), or *relatively ill-formed* (ill-formed with respect to a NL system's well-formedness constraints, even though a native speaker may judge it well-formed).

Some examples of absolutely ill-formed language that are peculiar to written language are:

1. *Typographical errors*, e.g., *oter*, instead of *other*. Typos may also result in recognizable words, such as *an* instead of *and*.
2. *Spelling errors*, e.g., *Ralf* instead of *Ralph*.
3. *Punctuation errors*, e.g., inserting or omitting commas incorrectly, misplacement or omission of apostrophes in possessives, etc.

4. *Homonym errors*, e.g., *to* instead of *too*, or confusing *there*, *their*, and *they're*.

Similarly, there are classes of absolute ill-formedness peculiar to spoken language.

5. *Mispronunciations*, e.g., saying that word as if it were spelled *mispronounciations*, or stressing the wrong syllable. V.A.Fromkin (1973) has provided a taxonomy of human speech production errors that appear rule-based, as opposed to ungoverned or random occurrences.

6. *Spoonerisms*, e.g., saying *fouler waucet* instead of *water faucet*.

Each of the classes above are human performance errors, resulting in absolute ill-formedness. However, the overwhelming variety of ill-formedness problems arises in both the spoken and written modality; examples of *absolute ill-formedness* include:

1. *Misreference*, as in describing a *purple* object as *the blue one*.

2. *Word order switching*, as in saying *the terminal of the screen* when one meant *the screen of the terminal*. (V.A.Fromkin (1973) has recorded these errors.)

3. *Negation errors*, e.g., *All doors will not open* when the train conductor meant *Not all doors will open*.

4. *Omitting words*, as in *Send file printer* rather than the full form *Send the file to the printer*. (Although this may seem to occur only in typed language, such omissions occur in spoken language either. Further, consider how many times, when struggling for the appropriate word, you start the utterance over, or someone supplies an appropriate word for you.)

5. *Subject-verb disagreement*, as in *A particularly important and challenging collection of problems are relatively ill-formed and arise in both spoken and written language* or in *One of the overwhelming number of troubles that befell them are . . .*

6. *Resumptive pronouns and resumptive noun phrases*, as in *The people that he told them about it*, where *them* is intended to be coreferential with *people*.

7. *Run-together sentences*, as if the person forgot how the sentence was started. An example is: *She couldn't expect to get a high standard salary and plus being so young*.

8. *Restarted sentences*, as in *Some people many try to improve society*, which was also collected in a written corpus.

9. *Pronominal case errors*, as in *between you and I*.

10. *Word order errors*, as non-native speakers can make, e.g., *I wonder where is the problem*.

Some particularly important and challenging problems are relatively ill-formed and arise in both spoken and written language. They are:

1. Words unknown to the hearer or reader, but part of the language.

2. Novel or unknown word senses, although the word itself is known. For instance, Navy jargon includes phrases such as *What is Stark's readiness?* Although that sublanguage does not include *preparedness* as a synonym for *readiness*, it would be useful for a system to be able to infer what a user means by the input *What is Stark's preparedness?*

3. Novel (non-frozen) figures of speech, e.g., metaphor, metonymy, and synecdoche.

4. Novel nominal compounds, as in *window aisle seat*, which was used by a flight attendant on a wide-body jet.

5. Violated presuppositions, as in *Did John fail to go?* when John did not try to go.

The above lists are not intended to be exhaustive. More thorough taxonomies of ill-formedness exist. Statistical studies of frequency of occurrence for various classes of ill-formedness have been conducted for written database access; those studies suggest that as much as 25-30% of typed input may be absolutely or relatively ill-formed.

From the definitions and examples, it is clear that:

1. Ill-formed input need not be ungrammatical; there may be no interpretation due to semantic or pragmatic problems.

2. The NL system will probably not know whether the input contains an error or whether its models are too limited to process the input.

3. Since there is no interpretation for the input, then one or more of the constraints of the NL system are violated; understanding ill-formed input therefore is a constraint satisfaction problem.

4. Since one or more of the constraints are violated, relaxing constraints in order to find an interpretation will mean opening up the search space for an interpretation substantially.

One new approach is to use *probabilistic language models* based on statistics derived from a chosen corpus, and utilizing those statistics together with the knowledge bases acquired from the corpus. The probabilistic model will rank partial interpretations for incomplete, errorful, or novel expressions. This will enable ranking of alternative interpretations when the input is complete, incomplete, or errorful.

The large annotated corpora described in the previous section will offer significant data to estimate such probabilities. For instance, the frequency of occurrence of types of phrases (e.g., NP and PP in the earlier annotated example) and statistics on relative frequency of grammar rules can be computed. Such statistics can be used to find the most predictive statistical language models for NLP systems.

The probabilistic language models in speech recognition (described in the next chapter) are probably not directly applicable. Typically probabilities of two- or three-word sequences are computed from a corpus of utterances and are used in assigning weights to each alternative rendering of the speech wave into sequences of words. The limitation in those models is that only local information is used, whereas it is well known in linguistics that there are long distance dependencies well beyond three-word sequences.

Scoring techniques based on large annotated corpora may provide the missing link for progress in understanding fragmentary language, in processing errorful language, in determining what was meant in novel expressions, and in processing incomplete forms.

In the last ten years, it has often been suggested that ignoring constraints, or bottom-up parsing, or a semantics-first strategy might be used to deal with ill-formed input, but in each case, although particular examples could be made to work, the

approach generated too many possibilities to be used in a truly general way. However, there seems to be a clear distinction between those classes of problems for which reasonably good syntactic and semantic strategies exist, and classes of ill-formedness that seem particularly intractable without a strong model of pragmatic knowledge for proper understanding. Examples of the latter include spelling/typographical errors that result in a known word, run-together sentences, pragmatic overshoot, contextual ellipsis requiring considerable reasoning to resolve, and inferring the meaning of unknown words.

## 5. Linguistic research opportunities

The following areas of opportunity for near-term research to make significant breakthroughs that will move NLP may be outlined:

*Acquisition of corpora, grammars, and lexicons*.

The development of useful systems requires observation of the behavior of potential users of interactive systems under realistic circumstances, and the collection of corpora of typical data for text analysis and machine translation systems. Although it is unlikely that full grammars and lexicons can be induced completely automatically in the near future, useful results may be obtained soon from induction and acquisition techniques based on annotated corpora and machine-readable dictionaries. It is also likely that statistical measures useful for biasing algorithms can be extracted from a handcrafted grammar and a corpus. Approaches that appear promising are (1) the learning of grammatical structures where the input has already been annotated by part of speech and/or phrase structure, and (2) the learning of lexical syntax/semantics from examples and/or queries to the user given some pre-coded domain knowledge.

*Increasing expressive power of semantic representation languages*.

Moving beyond database query systems will require increasing the expressive power of the languages used to express meaning, to include at least modal and higher-order constructs. Reasoning tools for modal logics and for second-order logics already exist, but appear intractable for language processing tasks. Approaches that seem promising include encoding modal constructs in first-order logic, hybrid approaches to representation and reasoning, and approaches to resource-limited and/or shallow reasoning, such as adding weights to formulae and sub-formulae.

*Reasoning about plans.*

Recent work on plan recognition - the inference of the beliefs and intentions of agents in context - has provided formal definitions of the problem and some new algorithms. These have not yet been used as part of a discourse component to help resolve reference, quantification, and modification ambiguities or to formulate an appropriate response. The interaction between plans, discourse structure, and focus of attention must also be investigated. Promising approaches include incorporation of beliefs of the discourse participants, integrating existing models into discourse processing under simplifying conditions, and exploring prosodic/linguistic cues to dialogue.

*Combination of partial information.*

The standard control structure by which various sources of information are combined in language interpretation seems to limit what NL systems can do. Several proposals for more flexible control structures have been made recently, each covering a subset of the knowledge sources available. More comprehensive schemes need to be developed. Two promising approaches are generalization of unification to NL architectures, and use of global, weighted control strategies, such as in evidential reasoning.

*Improving robustness.*

Published studies suggest that as much as 25-30% of typed input contains errors, is incomplete, uses novel language, or otherwise involves challenging phenomena that are not well handled theoretically. The frequency of occurrence for these classes is even higher in spoken language than in written language. The text of some messages, such as Navy RAINFORM and CASREP messages and bank telexes, is highly telegraphic. It should be possible to develop a domain-independent theory that allows at least partial understanding of some of these novel and errorful uses, and test it in narrowly defined domains. Promising approaches are to employ unification strategies, plan recognition, and/or weighted control strategies to determine the most likely interpretation and the most appropriate response/action.

*Relating interpretation and action.*

The problem of how to relate interpretations expressed in a meaning representation language and calls to application systems (databases, summarizing algorithms, etc.) has not been fully resolved, nor in fact precisely stated. This is crucial to the systematic separation of the natural language part of the system from the application part. Any approach should deal with applications beyond databases (beyond the semantics of tables) and should avoid the challenges of automatic programming.

*Finding the relationship between prosody, syntactic ambiguity, and discourse structure.*

Syntactic and discourse boundaries are one of the main sources of interpretation ambiguity. Recently discovered evidence shows that prosodic information is a good indicator of these boundaries. Automatic extraction of prosodic information would revolutionize the interpretation of spoken language. Further, generation systems could add prosodic information to signal syntactic structure and discourse structure.

*Measuring progress.*

The means of measuring progress is still an active area of discussion among NL scientists. Measures of correctness can be relatively simply stated for database query systems without dialogue capabilities (e.g., without sequence-related queries or clarifications), or for text analysis systems for database entry. They are much more difficult to state when stylistic matters need to be considered (as in machine translation systems) or when system responses affect subsequent user utterances. They probably cannot be usefully stated in a domain- or task-independent way. Measures of task difficulty, or of ambiguity of the language model, analogous to speech recognition's perplexity, are much more difficult to state. The recent DARPA program in spoken language understanding is developing formalisms for evaluating spoken language systems (M.Bates, S.Boisen, J.Makhoul 1991).

Measurement of NL systems requires three distinct types of comparisons:

1. Longitudinal: It is critical to be able to measure the performance of a system over time, so that progress can be tracked.

2. Cross-System: It should be possible to compare the overall performance of two systems in explicit terms. This focus on whole-system performance will help localize the strengths and weaknesses of complete systems and will identify topics for research and development efforts.

3. Component: It should be possible to evaluate and compare parts of systems and evaluate coverage of unknown phenomena. This focus on components will help point out areas of relative strength in different systems and will provide priorities and goals for specific research.

Both the longitudinal and cross-system measures should include not merely the percentage of inputs banded correctly but also estimates of productivity improvements for the end user.

## Conclusion

The most visible results in NLP in the last ten years are several commercially available systems for database question-answering. These systems, the result of transferring technology developed in the 1970s and early 1980s, have been successfully used to improve productivity by replacing fourth-generation database query languages. The success of these systems has depended on the fact that sufficient coverage of the language is possible with relatively simple semantic and discourse models. The semantics are bounded by the semantics of the relations used in databases and by the fact that words have a restricted number of meanings in one domain. The discourse model for a query is usually limited to the previous answer (usually numeric, simple strings, or a table) and the noun phrases mentioned in the last few queries.

It goes without saying that any NLP system must know a fair amount about words, language, and some subject area before being able to understand language. Currently, virtually all NLP systems operate using fairly laboriously hand-built knowledge bases. The knowledge bases may include both linguistic knowledge (morphological, lexical, syntactic, semantic, and discourse) and nonlinguistic knowledge (semantic world knowledge, pragmatic, planning, inference), and the knowledge in them may be absolute or probabilistic. Not all of these knowledge bases are necessary for every NLP system.

# CHAPTER VI.
# SPEECH TECHNOLOGY

## 1. The Value of Speech in Human-Machine Communication.

As information technology continues to make more impact on many aspects of our daily lives, the problems of communication between human beings and information-processing machines become increasingly important. Up to now such communication has been almost entirely by means of keyboards and screens, but there are substantial disadvantages of this method for many applications. Speech, which is by far the most widely used and natural means of communication between people, is, at first sight, an obvious substitute. However, this deceptively simple means of exchanging information is, in fact, extremely complicated. Although the application of speech in the man-machine interface is growing rapidly, in their present forms machine capabilities for generating and interpreting speech are still a travesty of what a young child can achieve with ease.

Advances in electronic and computer technology are causing an explosive growth in the use of machines for processing information. In most cases this information originates from a human being, and is ultimately to be used by a human being. There is thus a need for effective ways of transferring information between people and machines, in both directions. One very convenient way in many cases is in the form of speech, because speech is the communication method most widely used between humans; it is therefore extremely natural and requires no special training.

There are, of course, many circumstances where speech is not the best method for communicating with machines. For example, large amounts of text are much more easily received by reading from a screen, and positional control of features in a computer-aided design system is easier by direct manual manipulation. However, for interactive dialogue and for input of large amounts of text or numeric data speech

offers great advantages. Where the machine is only accessible from a standard telephone instrument there is no practicable alternative.

To appreciate how communication with machines can use speech effectively, it is important to understand the basic facts of how humans use speech to communicate with each other. The normal aim of human speech is to communicate ideas, and the words and sentences we use are not usually important as such. However, development of intellectual activity and language acquisition in human beings proceed in parallel during early childhood, and the ability of language to code ideas in a convenient form for mental processing and retrieval means that to a large extent people actually formulate the ideas themselves in words and sentences. The use of language in this way is only a convenient coding for the ideas. Obviously a speaker of a different language would code the same concepts in different words, and different individuals within one language group might have quite different shades of meaning they normally associate with the same word.

*1.1. The relation between written and spoken language.*

The invention of written forms of language came long after humans had established systems of speech communication, and individuals normally learn to speak long before they learn to read and write. However, the great dependence on written language in modern civilization has produced a tendency for people to consider language primarily in its written form, and to regard speech as merely a spoken form of written text – inferior because it is imprecise and often full of errors. In fact, spoken and written language are different in many ways, and speech has the ability to capture subtle shades of meaning that are quite difficult to express in text, where one's only options are in choice of words and punctuation. Both speech and text have their own characteristics as methods of transferring ideas, and it would be wrong to regard either as an inferior substitute for the other.

The study of how human speech sounds are produced and how they are used in language is an established scientific discipline, with a well-developed theoretical background. The field is split into two branches: the actual generation and

classification of speech sounds falls within the subject of *phonetics*, whereas their functions in languages are the concern of *phonology*. These two subjects need not be studied in detail by students of speech technology, but some phonetic and phonological aspects of the generation and use of speech must be appreciated in general terms.

The normal aim of a talker is to transfer ideas, as expressed in a particular language, but putting that language in the form of speech involves an extremely complicated extra coding process. The actual signal transmitted is predominantly acoustic, i.e. a variation of sound pressure with time. Although particular speech sounds tend to have fairly characteristic properties (better specified in spectral rather than waveform terms), there is great variability in the relationship between the acoustic signal and the linguistic units it represents. In analysing an utterance linguistically the units are generally discrete – e.g. words, phrases, sentences. In speech the acoustic signal is continuous, and it is not possible to determine a precise mapping between time intervals in a speech signal and the words they represent. Words normally join together, and in many cases there is no clear acoustic indication of where one word ends and the next one starts. For example, in "*six seals*" the final sound of the "*six*" is not significantly different from the [s] at the beginning of "*seals*", so the choice of word boundary position will be arbitrary. All else being equal, however, one can be fairly certain that the [s] sound in the middle of "*sick seals*" will be shorter, and this duration difference will probably be the only reliable distinguishing feature in the acoustic signal for resolving any possible confusion between such pairs of words. The acoustic difference between "*sick seals*" and "*six eels*" is likely to be even more subtle.

Although the individual sound components in speech are not unambiguously related to the identities of the words, there is, of course, a high degree of systematic relationship that applies most of the time. Because speech is generated by the human vocal organs the acoustic properties can be related to the positions of the articulators. With sufficient training, phoneticians can, based entirely on listening, describe speech

in terms of a sequence of events related to articulatory gestures. This auditory analysis is largely independent of age or sex of the speaker.

The International Phonetic Alphabet (IPA) is a system of notation whereby phoneticians can describe their analysis as a sequence of discrete units. Although there will be a fair degree of unanimity between phoneticians about the transcription of a particular utterance, it has to be accepted that the parameters of speech articulation are continuously variable. Thus there will obviously be cases where different people will judge a particular stretch of sound to be on the opposite sides of a phonetic category boundary.

Many of the distinctions that can be made in a narrow phonetic transcription, for example between different people pronouncing the same word in slightly different ways, will have no effect on meaning. For dealing with the power of speech sounds to make distinctions of meaning it has been found useful in phonology to define the *phoneme*, which is the smallest unit in speech where substitution of one unit for another might make a distinction of meaning. For example, in English the words "*do*" and "*to*" differ in the initial phoneme, and "*dole*" and "*doll*" differ in the middle (i.e. the vowel sound). There may be many different features of the sound pattern that contribute to the phonemic distinction: in the latter example, although the tongue position during the vowel would normally be slightly different, the most salient feature in choosing between the two words would probably be vowel duration. A similar inventory of symbols is used for phonemic notation as for the more detailed phonetic transcription, although the set of phonemes is specific to the language being described. For any one language only a small subset of the IPA symbols is used to represent the phonemes, and each symbol will normally encompass a fair range of phonetic variation. This variation means that there will be many slightly different sounds which all represent manifestations of the same phoneme, and these are known as *allophones*.

Phonologists can differ in how they analyse speech into phoneme sequences, especially for vowel sounds. Some economize on symbols by representing the long vowels in English as phoneme pairs, whereas they regard short vowels as single

phonemes. Others regard long and short vowels as different single phonemes, and so need more symbols. The latter analysis is useful for acknowledging the difference in phonetic quality between long vowels and their nearest short counterparts.

## 2. Digital Coding of Speech

It is known that some specialized low-data-rate communication channels actually code the speech so that it can be regenerated by synthesis using a functional model of the human speaking system, and some systems even use automatic speech recognition to identify the units for coding. A common method of automatic speech synthesis is to replay a sequence of message parts which have been derived directly from human utterances of the appropriate phrases, words or parts of words. In any modern system of this type the message components will be stored in digitally coded form.

Most of the coding methods were originally developed for real-time speech transmission over digital links, which imposes the need to avoid appreciable delay between the speech entering the coder and emerging from the decoder. This requirement does not apply to the use of digital coding for storing message components, and so for this application there is greater freedom to exploit variable redundancy in the signal structure.

To reproduce an arbitrary audio signal it is possible to calculate the necessary information rate (bits/s) in terms of the bandwidth of the signal and the degree of accuracy to which the signal must be specified within that bandwidth. For typical telephone quality the bandwidth is about 3 kHz and the signal-to-noise ratio might be 40 dB. The information rate in this case is about 40,000 bits/s. For a high-fidelity monophonic sound reproducing system the bandwidth would be about five times greater, and the noise would probably be 60-70 dB below the peak signal level. In this case a rate of about 300,000 bits/s is required to specify any of the possible distinct signals that could be reproduced by such a system.

In contrast to these very high figures, it is known that human cognitive processes cannot take account of an information rate in excess of a few tens of bits per second, thus implying a ratio of information transmitted to information used of between 1,000 and 10,000. This very large ratio indicates that the full information capacity of an audio channel should not be necessary for speech transmission. Unfortunately for the communications engineer, the human listener can be very selective in deciding what aspects of the signal are chosen for attention by the few tens of bits per second available for cognitive processing. Usually the listener concentrates on the message, which, with its normal high degree of linguistic redundancy, falls well within the capacity available. However, the listener may pay attention specifically to the voice quality of the speaker, the background noise, or even to the way certain speech sounds are reproduced.

There are two properties of speech communication that can be heavily exploited in speech coding. The first is the restricted capacity of the human auditory system. Auditory limitations make the listener insensitive to various imperfections in speech reproduction. When designing speech coding systems it can also be advantageous to make use of the fact that the signal is known to be produced by a human talker. The physiology of the speaking mechanism puts strong constraints on the types of signal that can occur, and this fact may be exploited by modelling some aspects of human speech production at the receiving end of a speech link. The potential reduction in digit rate that can ultimately be achieved from this approach is much greater than is possible from exploiting auditory restrictions alone, but such systems are only suited to auditory signals that are speech-like.

Coding methods can be divided into three general classes, thus:

1. Simple waveform coders, which operate at data rates of 16 kbits/'s and above;

2. Analysis/synthesis systems, which are most useful at low rates from 4 kbits/s down to less than 1,000 bits/s and, in the extreme, as low as about 100 bits/s;

3. Intermediate systems, which share some features of both of the first two categories and cover a wide range of rates in the region of 4-32 kbits/s.

Members of each class exploit aspects of production constraints and of perception tolerance, but to varying extents for different types of coders.

Waveform coders, as their name implies, attempt to copy the actual shape of the waveform produced by the microphone and its associated analogue circuits. If the bandwidth is limited, the sampling theorem shows that it is theoretically possible to reconstruct the waveform exactly from a specification in terms of the amplitudes of regularly spaced ordinate samples taken at a frequency of at least twice the signal bandwidth. In its conceptually simplest form a waveform coder consists of a band-limiting filter, a sampler and a device for coding the samples. The sampler operates at a rate higher than twice the cut-off frequency of the filter. The amplitudes of the samples are then represented as a digital code (normally binary) with enough digits to specify the signal ordinates sufficiently accurately. There is obviously no point in making the specification much more accurate than can be made use of for the given input signal-to-noise ratio.

### *2.1. Pulse code modulation*

This principle of coding, known as *pulse code modulation* (PCM), was suggested by J.Reeves (1938), and is now widely used for feeding analogue signals into computers or other digital equipment for subsequent processing (in which case it is known as *analogue-to-digital (A-D) conversion*). The process is not normally used in its simplest form for transmission or for bulk storage of speech, because the required digit rate for acceptable quality is too high. Simple PCM does not exploit any of the special properties of speech production or auditory perception except their limited bandwidth.

The distortion caused by PCM can be considered as the addition of a signal representing the successive sample errors in the coding process. If the number of bits per sample in the code is fairly large (say > 5) this quantizing noise has properties not obviously related to the structure of the speech, and its effect is then perceptually equivalent to adding a small amount of flat-spectrum random noise to the signal. If the number of digits in the binary code is small or if the input signal level exceeds the

permitted coder range, the quantizing noise will have different properties and will be highly correlated with the speech signal. In this case the fidelity of reproduction of the speech waveform will obviously be much worse, but the degradation will no longer sound like the addition of random noise. It will be more similar perceptually to the result of non-linear distortion of the analogue signal. Such distortion produces many intermodulation products from the main spectral components of the speech signal, but even when extremely distorted the signal usually contains sufficient of the spectral features of the original signal for much of the intelligibility to be retained.

The sound pressure waveform of a speech signal has a substantial proportion of its total power (for some speakers more than half) in the frequency range below 300 Hz, even though the information content of the signal is almost entirely carried by the spectrum above 300 Hz. As quantizing noise has a flat spectrum its effect on the signal-to-noise ratio is much more serious for the weaker but more important higher-frequency components. A considerable performance improvement for PCM can be obtained by taking into account this property of speech production, and applying pre-emphasis to the speech signal with a simple linear filter to make the average spectrum more nearly flat. After PCM decoding the received signal can be restored to its original spectral shape by de-emphasis, so reducing the higher-frequency components of the quantizing noise. For normal communication purposes it is not, however, necessary that the de-emphasis should match the pre-emphasis, as speech intelligibility is actually improved by attenuating the low-frequency components, because it reduces the upward spread of auditory masking.

The amplitude of the quantizing noise of simple PCM is determined by the step size associated with a unit increment of the binary code. During low-level speech or silence this noise can be very noticeable, but in loud speech it is masked, partially or in some cases completely, by the wanted signal. For a given perceptual degradation in PCM it is therefore permissible to allow the quantizing noise to vary with signal level, so exploiting a property of perception. The variation can be achieved either by using a non-uniform distribution of quantizing levels or by making the quantizing step size change as the short-term average speech level varies. Both methods have

been adopted, and have enabled excellent quantizing-noise performance to be achieved at 8 bits/sample, and useful communications performance at 4 bits/sample. Civil telephony uses PCM with 8 bits/sample at 8 kHz sampling rate, so needing 64 kbits/s. In this system there is an instantaneous characteristic that gives an approximately exponential distribution of quantizing intervals except at the lowest levels. (The two slightly different variants of this law used by different telephone administrations are known as A-law and U-law.) The sampling rate is generous for the 300-3,400 Hz bandwidth required, but this high sampling rate simplifies the requirements for the band-limiting filters. The time resolution properties of the auditory system ensure that masking of quantizing noise by the higher-level wanted signals is effective for at least a few milliseconds at a time, but instantaneous companding will give finer quantization near zero crossings even for large-amplitude signals. It is obvious that more effective use will be made of the transmitted digits if the step size is not determined by the instantaneous waveform ordinate height, but is changed in sympathy with the short-term average speech level. In this case, however, some means must be devised to transmit the extra information about the quantizing step size. This information can be sent as a small proportion of extra digits interleaved in the digital waveform description, but more usually it is embodied in the waveform code itself. The latter process is achieved by using a feedback loop that modifies the quantal step size slowly up or down according to whether the transmitted codes are near the extremities or near the centre of their permitted range. As the same codes are available at the receiver it is in principle easy to keep the receiver quantizing interval in step with that at the transmitter, but digital errors in the transmission path disturb this process and will thus affect the general signal level besides adding noise to the received audio waveform. Another disadvantage of this method of backward adaptation is that when the signal level increases suddenly it will overload the coder for at least a few samples before the quantizing interval has had time to adapt. Use of a separate channel for forward adaptation of the quantizing control can avoid this problem, but needs a small signal delay to enable the quantizer

to be correctly set before the signal is coded, in addition to the small amount of extra information needed to specify the quantizer step size.

*2.2. Deltamodulation.*

*Deltamodulation* is a very simple alternative type of waveform coding. A deltamodulator uses its transmitted digital codes to generate a local copy of the input waveform, and chooses successive digital codes so that the copy reproduces the input waveform as closely as possible, within the constraints of the coder. In its original and simplest form the quantizer uses only one bit per sample, and merely indicates whether the copy is to be increased or decreased by one quantum. Such a coder offers the possibility of extremely simple hardware implementation, and if run at a high enough sampling rate can approximate waveforms very closely. The process of following the waveform in small steps makes deltamodulation work best on signals in which differences between successive ordinates are small. Thus the low-frequency dominance in speech signals is accommodated directly by deltamodulation without pre-emphasis, and it is acceptable to use a quantal step that is only a very small fraction of the waveform amplitude range. In contrast, a flat-spectrum input would cause frequent slope overloading if used with the same step size and sampling rate. The scheme of deltamodulation is represented in figure 6.1 (borrowed from J.Holmes, W.Holmes, 2002).

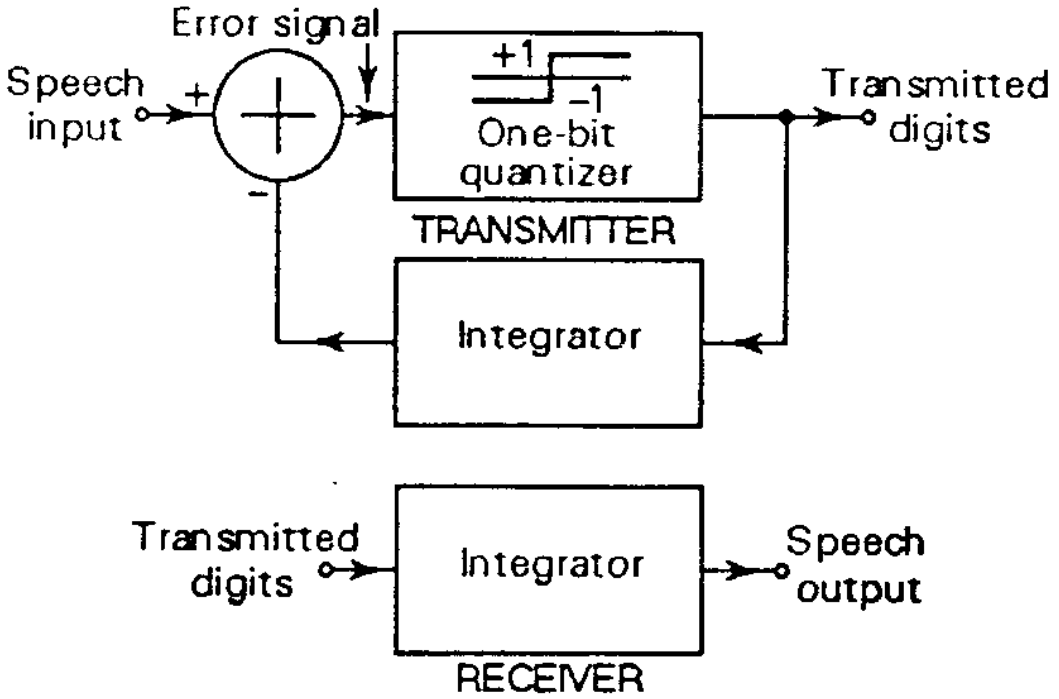



Figure 6.1. Block diagram of a simple deltamodulator

The use of a single bit per sample in deltamodulation is basically inefficient because a sampling rate much in excess of twice the highest frequency in the input signal is needed for close following of the input waveform. However, the intrinsic feedback loop in the coding process gives the coder some 'memory' of coding overload on previous waveform ordinates, for which it continues to compensate on later samples. This advantage of deltamodulation can be combined with those of PCM if a PCM coder is used instead of a one-bit quantizer in the feedback loop. Current terminology describes this arrangement as differential PCM (DPCM).

The advantages of and techniques for level adaptation apply to deltamodulation in the same way as to PCM, and adaptive forms of coder are normally used, so exploiting the noise-masking properties of auditory perception and the slow level changes of speech production. Adaptive DPCM (ADPCM) incorporating an adaptive quantizer seems to be the most efficient of the simpler waveform coding processes. At 16 kbits/s the quantizing noise is noticeable, but slightly less objectionable than the noise given by adaptive deltamodulation or adaptive PCM at the same digit rate.

The wave form in deltamodulation process will look like as presented in figure 6.2. (borrowed from J.Holmes, W.Holmes, 2002):

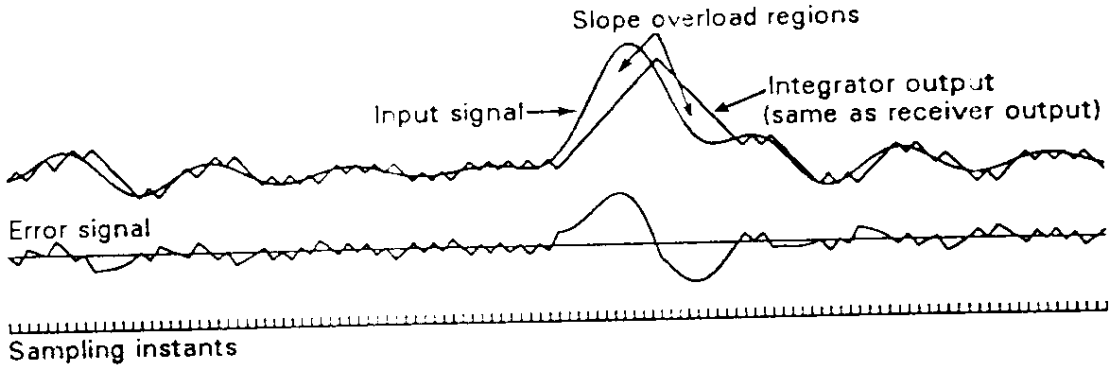



Figure 6.2. Waveforms in a simple deltamodulator

Many authors have also used the term ADPCM to describe waveform-following coders where the adaptation is based on much more complicated models of speech generation, with consequent much greater complexity than the simple coders. Coders of this more complicated type, but referred to as ADPCM, include a group of coders which have been recommended by the International Telecommunications Union (ITU) as standards for network telephony.

There are in fact a variety of waveform-following coders which incorporate adaptation but applied to a speech generation model of some complexity. It seems most useful, therefore, to describe these more elaborate systems in terms of the types of speech generation models they use and, in view of their higher complexity, they will be considered in the intermediate category.

*2.3. Analysis-synthesis systems (Vocoders)*

An alternative to direct waveform coding is to analyse the speech signal in terms of parameters describing its perceptually important characteristics. These parameters are transmitted and used to generate a new waveform at the receiver. The regenerated waveform will not necessarily resemble the original waveform in appearance, but should be perceptually similar. This type of coding system was first described by Homer Dudley of Bell Telephone Laboratories (Dudley, 1939), who called his system

a *vocoder* (a contraction of Voice Coder). The term *vocoder* has since been widely used to refer to analysis/synthesis coders in general.

Most vocoders are based on a model of speech production which exploits the fact that it is possible substantially to separate the operations of sound generation and subsequent spectrum shaping. The sources of sound are modelled by periodic or random excitation, and in several of the more recent vocoders it is also possible to have mixtures of both types of excitation. The excitation is used as input to a dynamically controllable filter system. The filter system models the combined effects of the spectral trend of the original sound source and the frequency response of the vocal tract. The specifications for the sound sources and for the spectral envelope are both derived by analysis of the input speech. By separating the fine structure specification of the sound sources from the overall spectral envelope description, and identifying both in terms of a fairly small number of slowly varying parameters, it is possible to produce a reasonable description of the speech at data rates of 1,000-3,000 bits/s. The structure of a vocoder is representedin fig. 6.3. (borrowed from J.Holmes, W.Holmes, 2002):

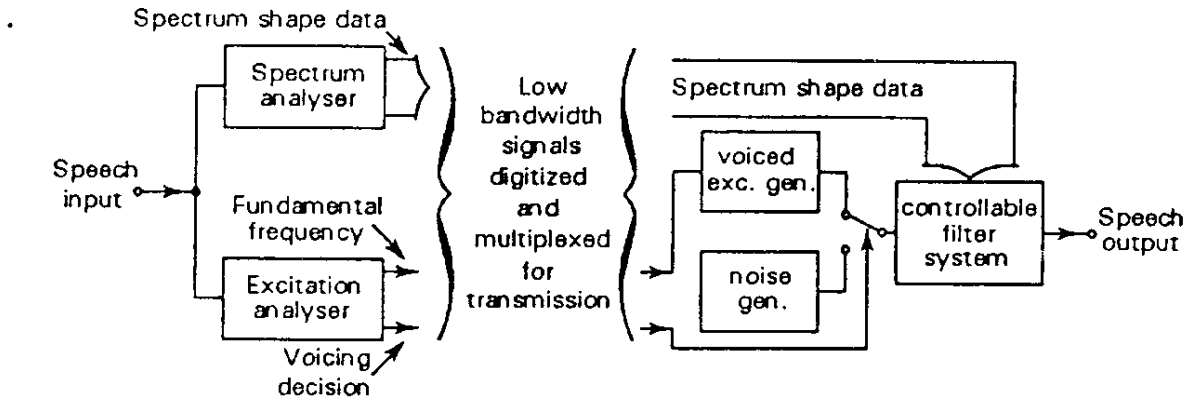



Figure 6.3. Block diagram of the basic elements of a vocoder

There are many different types of coder that use analysis/synthesis. These are channel vocoders, sinusoidal coders, linear predictive coding (LPC) vocoders and formant vocoders. With all these types the data are coded into frames representing speech spectra measured at intervals of 10-30ms. There are also techniques for efficient coding of frames and sequences of frames, and some specialized vocoders which code whole sequences of frames as single units.

Of the four types of vocoder mentioned above, up until around the late 1980s LPC and channel vocoders predominated. The advantages and disadvantages of the two types were nearly equally balanced; both gave usable but rather poor speech. Speech coding always involves a three-way compromise between data rate, speech quality and algorithm complexity.

Simple waveform coders, using *pulse code modulation* or *deltamodulation*, can achieve fairly good quality with very simple equipment, but require a high data rate. Adaptation of the quantizer in these coders improves the performance at any data rate with only a small increase in complexity. Analysis/synthesis systems ('vocoders') provide much lower data rates by using some functional model of the human speaking mechanism at the receiver. The excitation properties and spectral envelope are usually specified separately. Different types of vocoder describe the slowly varying spectral envelope in different ways. Channel vocoders specify the power in a set of contiguous fixed band-pass filters, and sinusoidal coders specify frequencies, amplitudes and phases of sinusoids. LPC vocoders use an all-pole sampled-data filter to model the short-term speech spectrum. Formant vocoders specify the frequencies and intensities of the lowest-frequency formants. Currently the most successful coders for real-time speech communication at 2,400 bits/s use sinusoidal coding or mixed-excitation linear prediction. Intermediate systems have some of the advantages both of vocoders and of simple waveform coders, and often use digit rates in the 4-16 kbits/s range.

Many intermediate systems use linear prediction analysis to exploit the resonant properties of speech production, but with different ways of coding the prediction residual for use as excitation in the receiver. Adaptive predictive coding, multipulse linear prediction and code-excited linear prediction can all give excellent speech quality at data rates well below 16 kbits/s - very low data rates of a few hundred bits/s can be achieved by coding whole sequences of frames as single units using segment or phonetic vocoders, but at the expense of complex processing and often quite poor speech quality. Ideally speech coders need to be evaluated by

subjective tests of both quality and naturalness, but objective comparison measures can also be useful.

## 3. Message Synthesis from Stored Human Speech Components

Several years ago the term "speech synthesis" was used almost exclusively for the process of generating speech sounds completely artificially in a machine which to some extent modelled the human speaking system. The applications were mainly for research in speech production and perception. These days, particularly in an engineering environment, speech synthesis has come to mean provision of information in the form of speech from a machine, in which the messages are structured dynamically to suit the particular circumstances required. The applications include information services, reading machines for the blind and communication aids for people with speech disorders. Speech synthesis can also be an important part of complicated man-machine systems, in which various types of structured dialogue can be made using voice output, with either automatic speech recognition or key pressing for the human-to-machine direction of communication. A conceptually simple approach to message synthesis is to concatenate fragments of human speech for the message components.. These synthesis techniques can be used for preparing limited sets of known messages, but they are also frequently used as the speech-generation component of more general systems for speech synthesis from unrestricted text.

### *3.1. Concatenation of whole words, sub-word units and waveform segments*

An obvious way of producing speech messages by machine is to have recordings of a human being speaking all the various words, and to replay the recordings at the required times to compose the messages. The first significant application of this technique was a speaking clock, introduced into the UK telephone system in 1936, and now provided by telephone administrations all over the world. The original UK

Speaking Clock used optical recording on glass discs for the various phrases, words and part-words that were required to make up the full range of time announcements. Some words can be split into parts for this application, because, for example, the same recording can be used for the second syllables of "twenty", "thirty", etc. The next generation of equipment used analogue storage on magnetic drums. For general applications of voice output there is a serious disadvantage with analogue storage on tapes, discs or drums: the words can only start when the recording medium is in the right position, so messages need to be structured to use words at regular intervals in order to avoid delays approaching the duration of one word or more. The use of recorded natural speech means that the technical quality of the reproduction can be extremely high.

There are applications where this method has worked extremely well. In the late 1960s it was used for some announcing machine applications in association with general-purpose computers, such as to provide share prices from the New York Stock Exchange to telephone enquirers.

The development of large cheap computer memories has made it practicable to store speech signals in digitally coded form for use with computer-controlled replay. As long as sufficiently fast memory access is available, this arrangement overcomes the timing problems of analogue waveform storage. Digitally coded speech waveforms of adequate quality for announcing machines generally use digit rates of 16-32 kbits per second of message stored, so quite a large memory is needed if many different elements are required to make up the messages.

For several years now there have been many computer voice-response systems commercially available that work on the principle of stored digitally coded message elements derived from human speech. The simplest of these systems involve merely recording the required components of the messages, which are then concatenated together without any modification to the individual elements. This simple concatenation can work well when the messages are in the form of a list, such as a simple digit sequence, or if each message unit always occurs in the same place in a sentence, so that it is comparatively easy to ensure that it is spoken with a suitable

timing and pitch pattern. Where a particular sentence structure is required, but with alternative words at particular places in the sentence, it is important that the alternative words should all be recorded as part of the right sort of sentence, because they would otherwise not fit in with the required sentence intonation. For list structures it is desirable to record two versions of every element that can occur either in the final or non-final position. The appropriate falling pitch can then be used for the final element in each list. Even for messages that are suitable for simple stored waveform concatenation, great care has to be taken in recording and editing the separate message components, so that they sound reasonably fluent when presented in sequence. For any large body of messages it is worthwhile to provide a special interactive editing system, in which any section of waveform can be marked and replayed, either in isolation or joined to other sections. By this means it is possible to select the best available recording and choose the precise cutting points for greatest fluency. Even with these special tools the editing is labour-intensive, and it can be very time-consuming to achieve good results with a message set of moderate size.

There are a number of difficulties associated with using stored speech waveforms for voice output when a variety of different messages are required. In normal human speech the words join together, and the inherently slow movements of the articulators mean that the ends of words interact to modify the sound pattern in a way that depends on the neighbouring sounds. The pitch of the voice normally changes smoothly, and intonation is very important in achieving fluency and naturalness of speech. It therefore follows that if single versions of each word are stored they cannot produce fluent speech if simply joined together in all the different orders that might be needed for a wide variety of messages.

Over 30 years ago laboratory experiments with arbitrary messages generated in this way demonstrated that the completely wrong rhythm and intonation made such messages extremely difficult to listen to, even though the quality of the individual words was very high.

*Synthesis by concatenating vocoded sub-word units*

Vocoder parameters for the sequence of synthesis units can be simply joined together, applying any necessary duration modifications. When shortening is required frames can be removed, and lengthening can be achieved by interpolating the synthesis parameters for the region to be lengthened. Interpolation across the boundary between two units has the advantage of reducing any discontinuities in the parameters. Thus, by only storing short transition regions , interpolation will usually be required to lengthen the units and at the same time minimize discontinuities. Any remaining discontinuities can be reduced after concatenation by applying a smoothing function to the parameters, in the same way as for concatenating vocoded words. Pitch modifications are easily achieved by varying the separate fundamental frequency parameter.

The quality of speech synthesized by vocoder-based concatenation cannot be better than the vocoder method employed. Although formant synthesizers can produce very natural-sounding speech if the controls are set appropriately, the quality of speech from formant vocoders suffers due to the difficulties involved in deriving these controls automatically. If careful hand-editing is used to correct analysis errors, a formant vocoder could be applied to generate the synthesis units. However, mainly because of the ease of analysis and availability of very-low-cost synthesis chips, LPC methods are much more widely used. The underlying quality is then limited to that possible from an LPC vocoder.

*Synthesis by concatenating waveform segments*

Consider the problem of joining together two segments of vowel waveform. Discontinuities in the combined waveform will be minimized if the join occurs at the same position during a glottal cycle for both the segments. This position should correspond to the lowest-amplitude region when the vocal-tract response to the current glottal pulse has largely decayed and just before the following pulse. Thus the two segments are joined together in a pitch-synchronous manner. To obtain a smooth join, a tapered window is applied to the end of the first segment and to the start of the second segment, and the two windowed signals are overlapped before being added

together. Because the method involves a combination of pitch-synchronous processing with an overlap-add (OLA) procedure to join the waveform segments, it is known as pitch-synchronous overlap-add (PSOLA).

The PSOLA technique can be used to modify pitch and timing directly in the waveform domain, without needing any explicit parametric analysis of the speech. The position of every instance of glottal closure (i.e. pitch pulse) is first marked on the speech waveform. These pitch markers can be used to generate a windowed segment of waveform for every pitch period. For each period, the window should be centred on the region of maximum amplitude, and the shape of the window function should be such that it is smoothly tapered to either side of the centre. A variety of different window functions have been used, but the *Hanning window*) is a popular choice. The window length is set to be longer than a single period's duration, so that there will always be some overlap between adjacent windowed signals. The OLA procedure can then be used to join together decomposing speech waveforms into a sequence of pitch-synchronous overlapping windows. For two voiced speech segments, pitch markers and window placement are shown in the top plots, and the outputs of the analysis windows are shown in the middle plots. The bottom plot shows the waveform that is obtained if the PSOLA technique is used to join the last analysis window of the first segment to the first analysis window of the second sequence of windowed signals, where each one is centred on a pitch marker and is regarded as characterizing a single pitch period. By adding the sequence of windowed waveform segments in the relative positions given by the analysed pitch markers, the original signal can be reconstructed exactly. However, by adjusting the relative positions and timber of the pitch markers before resynthesizing, it is possible to alter the pitch and timing, as described below.

### *3.2. Pitch level and time modification*

The pitch of the signal can be raised by reducing the spacing between the pitch markers, and lowered by increasing this spacing. As the degree of overlap between successive windows is altered, the energy in the resynthesized signal will tend to

vary, but a normalization factor can be applied to compensate for this artefact of the technique.

To be successful, the pitch-modification technique needs to change the pitch of the signal (given by the repetition rate of the pitch pulses) while not altering the spectral envelope (i.e. the formant frequencies and bandwidths). Thus the analysis window length needs to be short enough to be dominated by only a single pitch pulse, but long enough to capture the formant structure with sufficient accuracy.

The effect of this windowing of the signal tends to cause some widening of the formant bandwidths when the pitch is modified, but a moderate degree of widening does not seem to be perceptually significant. Widening of formant bandwidths becomes more severe as the pitch of the analysed signal increases, so the analysis window becomes shorter and hence there is a decrease in the accuracy with which the formant structure is preserved.

*Time modification*

It is straightforward to use PSOLA to modify the timing of an utterance by careful selection of the sequence of pitch markers to use for synthesis. Pitch markers can be replicated where lengthening is required, and removed when a region is to be shortened. The sequence of pitch markers gives the order of the analysis windows to use when constructing the synthesized signal. Synthesis is achieved by applying the OLA procedure to join these windowed segments together at a spacing corresponding to the required synthesis pitch period. When choosing the sequence of pitch markers to use in order to achieve the required timing, it is necessary to take into account the changes in duration that will occur as a by-product of any pitch modifications. If the pitch is altered, some adjustment to the sequence of pitch markers will be needed even to keep the timing the same as for the original signal.

Timing can be modified with little acoustic distortion using the above method to achieve the effect of increasing speaking rate by a factor of up to about four, but to reduce speaking rate by rather less. When slowing down unvoiced regions by more than a factor of about two, the regular repetition of identical segments of signal tends

to introduce a buzzy quality to the synthesized speech. This buzziness can be avoided by reversing the time-axis for every alternate segment, after which reasonable quality is obtained for slowing down by a factor of up to about four.

*Performance of waveform concatenation*

For PSOLA to work well, the positions of instances of glottal closure must be marked accurately on all the waveform segments. There are methods for determining these pitch markers automatically from the speech waveform, but these methods generally make some errors which need to be corrected by hand based on expert visual inspection of the waveform. More reliable automatic extraction of pitch markers is possible by using a *laryngograph* to record glottal activity simultaneously with the speech recordings. Whatever method is used to derive the pitch markers, part of this process will involve identifying unvoiced regions of the speech. For these regions, the positions of the analysis windows are not critical, and it is generally sufficient to place the pitch markers in arbitrary positions at a constant rate (although some care is needed for stop consonants).

Once speech segments and associated pitch markers are available, the PSOLA method described above is extremely simple to implement and requires very little computation, but it does need a lot of memory for storing the units. Some memory saving is possible by using a simple waveform coding technique such as DPCM (which typically more than halves the amount of memory required). However, the more complex coding methods that would be needed to obtain greater compression are not generally used with time-domain waveform synthesis, mainly because they would add considerable complexity to an otherwise simple synthesis procedure.

Because the individual message parts are obtained directly from human utterances, speech synthesized by waveform concatenation can be very natural-sounding. However, this naturalness is only achieved if any two segments to be concatenated have similar pitch periods and spectral envelopes that match at the join. Concatenation of waveforms provides no straightforward mechanism for using

phonetic synthesis by rule produce speech which is very intelligible for much of the time, but which does not sound natural and has a 'machine-like' quality.

More recently, with the advent of PSOLA and low-cost computer memory phonetic synthesis by rule has been largely abandoned for text-to-speech systems in favour of waveform-based concatenative techniques, which currently give more natural-sounding synthetic speech. However, formant synthesis by rule has important advantages in its inherently smooth model of co-articulation, and also in the flexibility to easily incorporate effects due to changes in speaking rate, voice quality, vocal effort and so on, by applying appropriate transformations to just the relevant controls. Although this flexibility is shared to some degree by parametric concatenative methods, it can be achieved in a more disciplined way with rule-driven synthesis. Techniques for automatic optimization using natural speech data may offer the opportunity for much higher-quality formant synthesis by rule to be achieved in the future.

To sum it up, we state that phonetic synthesis by rule involves applying acoustic-phonetic rules to generate synthesizer control parameters from a description of an utterance in terms of a sequence of phonetic segments together with prosodic information.

A convenient implementation is to store the rules as tables of numbers for use by a single computational procedure. Typically, a table for each phone holds some target synthesizer control values, together with transition durations and information used to calculate the controls at the nominal boundary between any pair of phones. Such a system can capture much of the co-articulation effects between phones.

Separate tables can be included for any allophonic variation which is not captured by the co-articulation rules. The total number of different units will still be far fewer than the number required in a concatenative system.

Acoustic-phonetic rule systems have tended to be set up 'by hand', but automatic procedures can be used to derive the parameters of these systems, based on optimizing the match of the synthesized speech to phonetically transcribed natural speech data.

## 4. Speech Synthesis from Textual or Conceptual Input

The previous chapters have described two different methods for generating an acoustic waveform from an input phoneme sequence together with prosodic information. Either of the methods can form one component of a more general speech synthesis system in which the input is at some higher level, which may be orthographic text or even concepts that are somehow represented in the machine.

When human beings speak, many factors control how the acoustic output is related to the linguistic content of their utterances. At one level, there are constraints determined by the physiology of their vocal apparatus. Although the physiology is generally similar between people, there are also clear differences of detail, partly related to age and sex, but also caused by genetic differences between individuals.

For a given vocal system, the speech depends on the sequence of muscular actions that control the articulatory gestures. These gestures are learnt from early childhood, and their details are determined partly by the properties of the inherited central nervous system, but also very much by the speech environment in which the child grows up. The latter feature is entirely responsible for determining the inventory of available phonetic productions of any individual, which is closely tied to his/her native language. At a (higher level, the relationship between the ideas to be expressed by the choice? of words, with their pitch, intensity and timing, is entirely determined by the language.

In acquiring competence in speech the human has two forms of feedback. On the one hand, auditory self-monitoring is paramount for comparing the acoustic patterns produced with those heard as model utterances. The second main form of feedback is the response by other human beings to imperfect utterances produced during language acquisition. Once the right types of utterances can be produced and the necessary gestures have been learnt, kinaesthetic feedback can be used for detailed control of articulatory positions, and can ensure continuation of competent speech even if auditory feedback is not available for any reason.

All the above aspects of speech acquisition imply that the human develops a set of rules at many different levels, to convert concepts to speech. Although some parts of these rules are determined by inherited physiology and some by learning from the environment, it is not easy to separate these two aspects. However, it is clear that there must be a set of rules to guide humans generating speech, although in many cases the utterances will be modified by chance or by creative variation within the limits of what is acceptable to retain the desired effect on the listeners.

To embody the complete process of human speaking, these rules must be fantastically complicated – particularly in the linguistic process of expressing subtle shades of meaning by choice of words and prosody.

The aim for computer speech synthesis from either textual or conceptual input is to imitate the characteristics of the typical human speaking process well enough to produce synthetic speech that is acceptable to human listeners. Synthesis from text should be able to apply the rules used by a good reader in interpreting written text and producing speech. In its most advanced form such a system should be able to apply semantic interpretation, so that the manner of speaking appropriate for the text can be conveyed where this is not immediately obvious from the short-span word sequences alone. Synthesis from concept poses rather different challenges, as the computer will already have some representation of the meaning to be conveyed, but an appropriate sequence of words must be generated for the required concepts before the words can be further converted into their acoustic realization. Most work on speech synthesis has concentrated on text-to-speech (TTS) conversion.

### *4.1. Converting from text to speech*

The generation of synthetic speech from text is often characterized as a two-stage analysis-synthesis process. The first part of this process involves analysis of the text to determine underlying linguistic structure. This abstract linguistic description will include a phoneme sequence and any other information, such as stress pattern and syntactic structure, which may influence the way in which the text should be spoken. The second part of the TTS conversion process generates synthetic speech

from the linguistic description. This synthesis stage can be further subdivided into prosody generation followed by generation of a synthetic speech waveform from the phonemic and prosodic specifications.

Both the analysis and synthesis processes of TTS conversion involve a number of processing operations represented in fig. 6.4.:

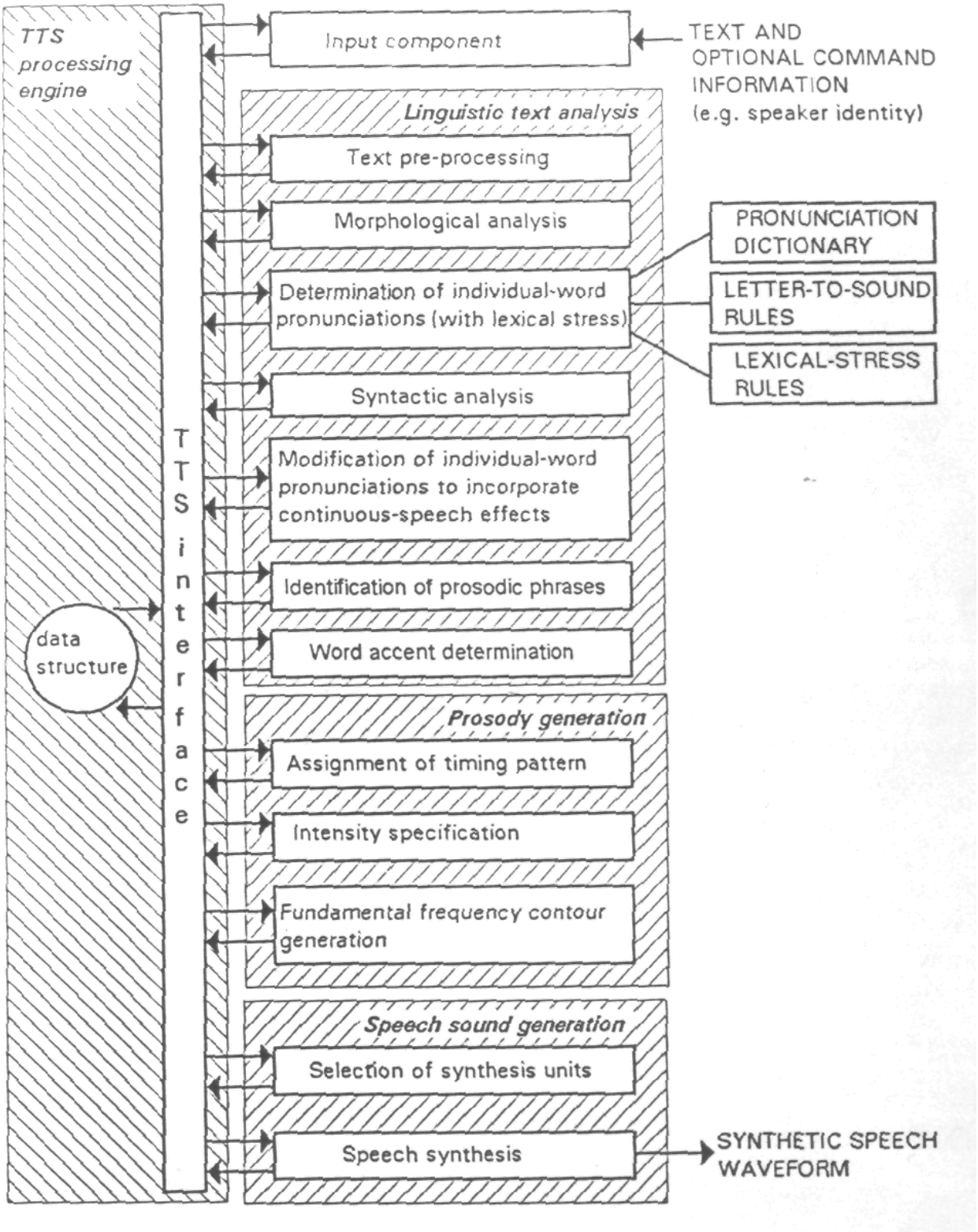


Figure 6.4. TTS conversion

Most modern TTS systems incorporate these different operations within a modular architecture. When text is input to the system, each of the modules takes some input related to the text, which may need to be generated by other modules in the system, and generates some output which can then be used by further modules, until the final synthetic speech waveform is generated.

For a language such as English the separation into words is fairly easy as words are usually delimited by white space. The detection of sentence boundaries is less straightforward. For example, a full stop can usually be interpreted as marking the end of a sentence, but is also used for other functions, such as to mark abbreviations and as a decimal point in numbers.

Any unrestricted input text is likely to include numerals, abbreviations, special symbols such as %, *, etc., capitalization and a variety of punctuation and formatting information (white space, tab characters, etc.). It is therefore usual for the text pre-processing to also include a process of text normalization, in which the input text is converted to a sequence of pronounceable words. The normalized text will typically consist of a sequence of explicitly separated words, consisting only of lower-case letters, and with punctuation associated with some of the words. For example, the text "*Dr. Smith lives at 16 Castle St*." could be converted to:

([doctor][smith][lives][at][sixteen][castle][street]),

where square brackets have been used to delimit each individual word and curly brackets – to delimit the sentence. Each word can be marked with tags to indicate detection of an expanded abbreviation, expanded numerals, capital letters and so on. In this way, all of the information can be passed on but at the same time the text is put into a format which is more suitable for further processing. Most TTS systems include a large number of rules to deal with the variety of text formats that may be encountered, and a few examples are given in the following paragraphs.

In the case of numerals, the correct pronunciation will depend on the context. In many contexts a four-digit number beginning in *1* represents a year and should therefore be pronounced according to the conventions for dates, but in other cases it

will be "*one thousand*" followed by the hundreds, tens and units (e.g. "1999" could be the year "*nineteen ninety nine*" or "*one thousand nine hundred and ninety nine*"). Telephone numbers in English are usually pronounced as a sequence of separate digits. A number with two decimal places will be pronounced as a sum of money if preceded by a currency symbol (e.g. "S24.75" becomes "*twenty-four dollars and seventy-five cents*"), but will otherwise include the word "*point*" (e.g. "24.75" becomes "*twenty-four point seven five*".

Conversions for abbreviations and special symbols can be provided in a lookup table. Special symbols are replaced by the relevant words (e.g. "%" is changed to "*per cent*", and "&" to "*and*"), and certain abbreviations need to be expanded as appropriate (e.g. "Mr." to "*mister*", and "etc." to "*et cetera*"). Some abbreviations are ambiguous and context needs to be taken into account to determine the correct expansion. Commonly cited examples are "Dr.", which can expand to "*doctor*" or to "*drive*", and "St.", which can expand to "*saint*" or to "*street*". While some abbreviations need to be expanded, others (e.g. "USA", "GMT") must be spelled out and these will be replaced by the appropriate sequence of letter names.

In general, dealing with abbreviations is quite straightforward as long as they are known in advance and have been included in a conversion table. It will, however, be impossible to predict all abbreviations that might occur in any arbitrary text, and so it is usual to include rules for detecting abbreviations. The presence of full stops between the letters can be taken as a good indication that the letter names should be pronounced separately. A word in capitals is also likely to be an abbreviation, at least if the surrounding words are in lower case. If the sequence of letters forms a pronounceable word, it is probably an acronym (e.g. "*NATO*") and should therefore be treated as a word, but otherwise the abbreviation can be pronounced as a sequence of letter names. However, some pronounceable sequences should also be spelled out as individual letters (e.g. "*MIT*"). The best strategy is probably to treat abbreviations of four or more letters as words if they are pronounceable. For shorter abbreviations, or if the letter combination is unpronounceable, it is more appropriate to spell out the individual letters.

Text pre-processing rules of the types described above can cope adequately with many text formatting phenomena, but unrestricted text is always likely to contain some formatting features which will be difficult to decode without sophisticated analysis of syntax and even meaning. It may be possible to overcome any ambiguity by delaying decisions that cannot be resolved at a pre-processing stage until the later stages of text analysis. Currently, however, the best results are still obtained if the designer prepares the TTS system for a known restricted range of applications, so that the pre-processing can be tailored appropriately.

*4.2. Morphological analysis*

Morphemes are the minimum meaningful units of language. For example, the word "*played*" contains two morphemes: "*play*" and a morpheme to account for the past tense. Morphemes are abstract units which may appear in several forms in the words they affect, so that for example the word "*though*t" comprises the morpheme "*think*" together with the same past-tense morpheme as was one used in the previous example. When there is a direct mapping between the abstract morphemes and segments in the textual form of the word, these text segments are referred to as morphs. In many words, such as "*carrot*", the whole word consists of a single morph. Others, such as "*lighthouse*", have two or more. Morphs can be categorized into roots and affixes, and the addition of common affixes can vastly increase the number of morphs in a word. For example, "*antidisestablishmentarianism*" has six morphs if "*establish*" is regarded as a single root morph. A high proportion of words in languages such as English can be combined with prefixes and/or suffixes to form other words, but the pronunciations of the derived forms are closely related to the pronunciations of the root words.

Rules can be devised to correctly decompose the majority of words (generally at least 95% of words in typical texts) into their constituent morphs. This morphological analysis is a useful early step in TTS conversion for several reasons:

It is then not necessary for all derived forms of regularly inflected words to be stored in the pronunciation dictionary. Instead, the pronunciation of any derived word

can be determined from-the pronunciation of the root morphs together with the normal pronunciations of the affixes. For example, inclusion of the word "*prove*" would enable the correct pronunciation of "*improvement*", "*proving*", etc. to be determined. (Note that it is necessary to take account of the fact that in many words a final "*e*" needs to be removed before the addition of certain suffixes.)

If the pronunciation of individual morphs is known, it is possible to deal with the many compound words of English and cover a high proportion of the total vocabulary while keeping the dictionary at a manageable size. Complete words need then only be included in the dictionary if they do not follow the regular morpheme composition rules of the language. A morph lexicon is also useful in predicting the pronunciation of unknown words. While the words in a language are continually changing, it is rare for a new morpheme to enter a language.

Even when it is necessary to apply letter-to-sound rules, some attempt to locate morph boundaries is beneficial as many of the rules for the pronunciation of consonant clusters do not apply across morph boundaries. For example, the usual pronunciation of the letter sequence "*th*" does not apply in the word "*hothouse*", due to the position of the morph boundary.

Morphological analysis gives information about attributes such as syntactic category, number, case, gender (in the case of some languages) and so on. This information is useful for later syntactic analysis.

Extensive use of morphological analysis and morph dictionaries was pioneered in the MITalk system (B.Allen et al., 1987), which covered over 100,000 English words with a morph lexicon of about 12,000 entries and hence moderate cost in terms of storage. While storage cost is no longer such an issue, the other advantages of morph decomposition are such that the better TTS systems all include at least some morphological analysis.

### *4.3. Phonetic transcription*

It is usual for the task of determining the pronunciation of a text to begin by assigning an idealized phonemic transcription to each of the words individually.

Nowadays, most TTS systems use a large dictionary. This dictionary will generally only contain the root forms of words and not their morphological derivatives, except for those derivatives that cannot be correctly predicted by rule. For words with alternative pronunciations, both possibilities can be offered by the dictionary and syntactic analysis may be used to choose between them.

A typical strategy for determining word pronunciation is to start by searching the dictionary to check whether the complete word is included. If it is not, the component morphs can be searched for. Provided that the individual morphs are in the dictionary, the pronunciation of the derived word can then be determined by rule from the pronunciation of its component morphs. When deriving pronunciations of derived words from their root form, it is necessary to take into account any pronunciation-modification rules associated with the affixes. For example, the suffix "*ion*" changes the phonemic interpretation of the final /t/ sound in words like "*create*".

For any words (or component morphs) whose pronunciation cannot be determined using the dictionary, letter-to-sound rules are needed. The complexity of the relationship between the spellings of words and their phonemic transcription is different for different languages. However, even in a language such as English which has a particularly complicated mapping between letters and phonemes, it is obvious that human readers must have some rules for relating spelling to phoneme sequences because they can usually make a reasonable guess at the pronunciation of an unfamiliar word. While it cannot be guaranteed that letter-to-sound rules will always give the pronunciation that most people would regard as correct, a human reader will also often make errors with unfamiliar words. However, these words are usually quite rare and the nature of any errors tends to be such that the incorrect phoneme sequence is often sufficient to indicate the intended word. Predicting the pronunciation of proper names is especially challenging, as names often follow quite different pronunciation rules from ordinary words and may be from many different languages. Many TTS systems include special rules for names, sometimes using a scheme based on analogy with known names (e.g. the pronunciation of "*Plotsky*" can be predicted by analogy with the pronunciation of "*Trotsky*").

In naturally spoken continuous speech, word pronunciations are influenced by the identities of the surrounding words. TTS systems incorporate these effects by applying post-lexical rules to make phonetic adjustments to the individual-word phonemic transcriptions. For example, the correct pronunciation of the vowel in the word "the" depends on whether the following word begins with a vowel (e.g. "*the apple*") or a consonant (e.g. "*the dog*"). Other effects on pronunciation are related to the consequences of co-articulation and the preferred option may depend on the speaking style. For example, the consonant sequence in the middle of "*handbag*" may be pronounced [ndb] in highly articulated speech, but would more usually be reduced to [nb], and may even become [mb] in casual speech.

### *4.4. Syntactic analysis and prosodic phrasing*

Some syntactic analysis is needed both to resolve pronunciation ambiguities and to determine how the utterance should be structured into phrases. Possible syntactic classes can be included with each entry in the dictionary, and the morphological analysis will also provide useful information about likely parts of speech. However, very many English words may be used as both nouns and verbs, and several can also be adjectives, so very little definite information about syntax can be resolved without taking into account the relationships between the words.

Assignment of syntactic classes, or part-of-speech tags, is often achieved using a statistical model of language, based both on probabilities for particular tags appearing in a certain context and on probabilities of the tags being associated with the given words. The model probabilities can be derived from large amounts of correctly marked text, and the modelling technique itself is one that is widely used for language modelling in automatic speech recognition.

Once the part of speech has been decided for each word in a sentence, the phrase structure of the sentence can be determined. In order for suitable prosody to be generated, it is necessary to decide on sentence type (declarative, imperative or question), and to identify phrases and clauses. Some systems have included full syntactic parsing, while others perform a more superficial syntactic analysis, for

example to locate noun phrases and verb phrases and possibly group these phrases into clauses. There are also methods for using statistical models trained to predict prosodic phrases directly from information about parts of speech, stress, position in the sentence and other relevant factors. The general aim is to produce a reasonable analysis for any text, even if the text contains syntactic errors. There will always be instances for which correct assignment of appropriate phrasing cannot be achieved without incorporating semantic and pragmatic constraints, but current TTS systems do not have more than very limited capability to apply such constraints.

*4.5. Assignment of lexical stress and pattern of word accents*

In the case of polysyllabic words, there is normally one syllable that is given primary stress, and other syllables are either unstressed, or carry a less prominent secondary stress. These lexical stress markings can be included for each entry in the dictionary. When the pronunciation of a word is obtained by combining morphs, the stress pattern for the individual morphs may be changed, so it is necessary to apply rules to determine the stress pattern for the complete word. For example, the addition of the suffix "*ity*" to "*electric*" moves the primary stress from the second syllable to the third. For some words, such as "*permit*", the stress assignment depends on syntactic category, so the choice between alternative stress patterns must be made following the syntactic analysis.

With any word whose pronunciation has to be obtained by letter-to-sound rules, additional rules are also needed to assign lexical stress. For many polysyllabic words of English the placement of primary and secondary stresses on the syllables can be determined reasonably accurately using very complicated rules that depend on how many vowels there are in the word, how many consonants follow each vowel, the vowel lengths, etc. There are, however, many words for which the normal rules do not apply, as exemplified by the fact that some pairs of words are of similar structure yet are stressed differently. Examples are "*Canada*" and "*camera*", contrasting with "*Granada*" and "*banana*". Words such as these will need to be included in the dictionary to ensure correct lexical stress assignment.

One of the last tasks in text analysis is to assign sentence-level stress to the utterance, whereby different words in a sentence are accented to different extents. Assignment of accents depends on a number of factors. Function words (such as articles, conjunctions, prepositions and auxiliary verbs) serve to indicate the relationships between the content words that carry the main information content of an utterance. Function words are not normally accented, whereas content words tend to be accented to varying degrees dependent on factors such as parts of speech and the phrase structure. In addition to the syntax-driven placement of stress, emphasis may be placed on important words in the sentence. For example, when a speaker wishes to emphasize his or her attitude towards the truth of something, words such as "*surely*", "*might*" and "*not*" may be used with stress. Stress may also be used to make a distinction between new and old information, or to emphasize a contrast. Some TTS systems include rules to model a number of these types of effects. The pattern of accents on the different words will usually be realized as movements in fundamental frequency, often referred to as pitch accents.

*4.6. Prosody generation*

The acoustic correlates of prosody are intensity, timing pattern and fundamental frequency. Intensity is mainly determined by phone identity, although it also varies with stress for example. From the perspective of prosody, intensity variations are in general less influential than variations in timing pattern and in fundamental frequency contour, which are discussed in the following sections.

Both in concatenative synthesis and in most synthesis-by-rule methods, utterances are generated as sequences of speech segments. For any utterance, a duration needs to be chosen for each segment such that the synthesized speech mimics the temporal structure of typical human utterances. The temporal structure of human speech is influenced by a wide variety of factors which cause the durations of speech segments to vary. Observations about this variability include the following:

1. The inherent durations of different speech sounds differ considerably. Some vowels are intrinsically short and others long. The vowels in the words "*bit*" and

"*beet*" in English differ in this way. Diphthongs are usually longer than monophthongs, and consonant sounds also show systematic differences.

2. Durations differ according to speed of speaking, but sounds that are mainly steady in character, such as fricatives and vowels, tend to vary in duration more than inherently transient sounds, such as the bursts of stop consonants.

3. If a particular word in a sentence is emphasized, its most prominent syllable is normally lengthened.

4. Durations of phones vary according to their position in a word, particularly if there are several syllables.

5. When at the end of a phrase, a syllable tends to be longer than when the same syllable occurs in other locations in a phrase.

6. Vowels before voiced consonants are normally longer than occurrences of the same vowels before unvoiced consonants. For example, in the English words "*feed*" and "*feet*" the vowel is substantially longer in "*feed*". There are also other systematic duration modifications that depend on the identities of neighbouring phones.

7. Some evidence suggests that, in a 'stress-timed' language such as English, unstressed syllables tend to be shorter if there are several of them between two stressed syllables. However, the empirical evidence is less conclusive for this effect than for the other effects listed above.

A number of systems have been developed for deriving segment durations by applying a succession of rules. These rules operate on phonetic transcriptions with the stressed syllables marked, and assume that some decision has been made about speed of speaking. It is then possible to estimate a suitable duration for each phone by having some intrinsic duration for the phone, and to modify it by various amounts according to each of the circumstances mentioned above. The amount of the modification could in general depend on the circumstances causing it and on the identity of the phone whose duration is being calculated. Sets of rules have been devised and refined based on phonetic knowledge in combination with statistics of speech segment durations and the results of small-scale experiments investigating the

effect of varying different factors on synthesis quality. While reasonable success has been achieved in producing acceptable timing patterns, this approach is not able to guarantee that the rules are optimized simultaneously to the very wide range of utterances that general TTS systems must be able to deal with.

In recent years, as large speech corpora and increased computational resources have become available, there has been a growth in alternative approaches using automatic optimization to derive the parameters of a general model based on large databases of segmented and labelled speech. It is quite straightforward to apply these data-driven methods to derive a reasonable duration model for a new language, provided that sufficient labelled speech data are available.

Automatic methods have achieved some improvement over the older rule-based systems. However, current TTS systems are still not able to produce the rhythm that humans can adopt naturally in sentences containing rhyming clauses, or to generate other systematic variations related to meaning. Speech synthesized from text also lacks the pattern of pauses and decelerations that are found in speech from a good human reader, and which serve to enhance a listener's comprehension. More elaborate linguistic analysis would be necessary to produce all these effects.

### *4.7. Fundamental frequency contour*

The fundamental frequency of voiced speech, which determines the perceived pitch, is widely used by all languages to convey information that supplements the sequence of phonemes. In some languages, such as Chinese, pitch changes are used to distinguish different meanings for syllables that are phonetically similar. In most Western languages pitch does not help directly in identifying words, but provides additional information, such as which words in a sentence are most prominent, whether a sentence is a question, statement or command, the mood of the speaker, etc. Even for these Western languages, the type of intonation pattern that is used to achieve particular effects varies considerably from one language to another, and even between accents of the same language. Obviously the model for generating a suitable

intonation pattern must be developed to suit the required language. For the purposes of this book, examples will be given for typical southern British English.

Most sentences in English show a general tendency for pitch to fall gradually from beginning to end of each sentence, but with many local variations around this trend. Two major factors determining these variations are the way in which the sentence is subdivided into phrases and the sentence stress pattern. The most significant pitch variations occur at major phrase boundaries and on words that the user wishes to be more prominent. In the case of polysyllabic words, the syllable with primary stress carries the main pitch movement.

The normal structure of English is such that the last syllable carrying primary stress in any breath group is given the biggest pitch change, and is known as the nuclear syllable. Usually the nuclear tone (i.e. the pitch pattern on the nuclear syllable) on a simple statement is a pitch fall, but a number of other patterns are possible to indicate other types of utterance. (The number of possible nuclear tones is at least three, but some workers have claimed that there are up to six significantly different patterns.) The nuclear tone for a question expecting a yes/no answer shows a substantial pitch rise. On the non-final stressed syllables the pitch usually shows a local small rise and then continues its steady fall. The amount of this rise and the subsequent rate of fall can depend on the syntactic function of the word in the sentence: verbs, for example, generally have less pitch variation than nouns and adjectives. At the beginning of an utterance the pitch often starts fairly low, and then rises to a high value on the first stressed syllable.

In addition to these pitch changes caused by the pattern of stressed syllables, there are smaller pitch variations that are influenced by the phonetic detail of an Taking the example of speaker recognition, speaker verification involves deciding whether or not a given voice sample was spoken by one known individual, based on how well this sample matches the reference for the voice of the one speaker. If the match is good enough, the utterance is accepted Speaker identification (within a closed set) entails deciding on the speaker identity from a set of known' speakers, by finding the one from the reference set that gives the closest match to an 'unknown'

voice sample. Open-set speaker identification requires both types of decision to be made, so as to make a choice between speakers or to reject the sample if it does not match any of the known speakers well enough. Similar distinctions apply in language recognition, although most research has concentrated on (closed-set) language identification.

In either identification or verification tasks, it is possible to make an estimate of confidence m the recognition decision. For certain applications (such as speaker verification for secure access), further input can be requested when a confident decision cannot be made based only on the original input utterance Performance tends to improve as the amount of speech material increases, and hence more information becomes available, until some maximum performance level is reached.

## 5. Applications and Performance of Current Speech Technology

In the past few years there has been a large and continuing increase in the number and range of products and services that incorporate speech technology. More and more people have experience of an application that uses speech technology in some way. This increase in applications is due partly to the advances in the methods that are used in speech synthesis and recognition, but also to the more general progress that has been made in computer technology. The increases in the computer power and memory that have become available at decreasing cost have contributed to the growth of speech technology in two ways. Firstly, the advances in computers have been a crucial factor in much of the recent progress in the speech synthesis and recognition techniques themselves. Secondly, the widespread use of computers has opened up new opportunities for exploiting speech technology. The fantastic growth of the Internet has created a demand for easy ways of accessing and retrieving all the information and sendees that are becoming available. Also highly relevant to the application of speech technology are the more general developments that have taken place in telecommunications, including the growth in mobile telephony. There are

now a vast number of automated telephone-based services, for which voice is the most natural means of communication.

Speech is not always the most appropriate or the easiest means of communication between humans and machines. Speech technology must offer some tangible advantage over alternative options if it is to be successful in any given application. *Potential advantages* include:

1. Cost savings may be obtained by automating services or by making human operators more efficient.

2. Effectiveness may be improved, for example in terms of speed and quality of output or in terms of ease with which a goal can be achieved.

3. Safety may be increased by using an additional modality for communication.

Situations in which there are obvious advantages to be gained from applying speech technology can be categorized as follows:

1. *Hands busy, eyes busy*: The usual mode of communication with a computer or other machine is to input commands using the hands, and to receive output visually. However, in situations when it is not possible to use the hands and/or the eyes, speech can provide a valuable alternative means of communication. Such situations arise when a person's hands and eyes are occupied, for example operating some piece of equipment, but also when hands or eyes cannot be used for some other reason such as disability, or the need to operate in darkness or to wear special equipment that makes manual operation difficult.

2. *Remoteness*: The telephone makes it possible to communicate with computers remotely. Although touch-tone phones are now widespread and can be used to input information, speech is more natural and is much easier for many types of information. For the machine-human direction of communication, speech is the only really viable option.

3. *Small devices*: Computers are becoming miniaturized. For communicating with palm-top computers and other small devices, there are many circumstances in which speech is easier than using a pointing device or limited keyboard. Similarly,

when there is only a small display available, speech can provide a better means for output of many types of information.

Speech technology performance does not yet approach human performance, but there are many tasks for which the technology is useful. In situations where speech technology is providing users with a facility they would not otherwise have, those users will generally be more tolerant of limitations in the technology than will users of applications for which there are other alternatives. However, for any application, achieving success is critically dependent upon designing the system and its user interface to take into account the strengths and weaknesses of the particular technology that is to be used given the requirements of the application.

### *5.1. Speech Synthesis Technology*

When spoken messages are required, currently the main choice is between a text-to-speech (TTS) system and digitally recorded speech, possibly compressed using some speech coding algorithm. Recorded speech offers the best quality, or alternatively, with some loss in quality, can be used in coded form very cheaply on simple DSP chips. The principal disadvantage is lack of flexibility: if a new word or a different type of message is required, it is necessary to make a new recording. In addition, there may be practical difficulties in using recorded speech if a large number of different messages are required. For applications where the messages are unpredictable or are likely to be changed frequently, TTS is the only practical option. As it has already been discussed, the best TTS systems produce speech that is highly intelligible and sounds fairly natural on short, straightforward utterances. However, for longer passages, especially those requiring the expression of emotion, the perceived quality can drop dramatically. Thus at present the most successful applications for TTS synthesis are those needing only short utterances with simple intonation patterns, or those applications that need more complex utterances but for which lower quality will be tolerated because the system provides the users with a facility which they would not otherwise have.

For some speech synthesis applications (such as telephone-based information sendees), the requirement is really for a message preparation system. Such a system needs the flexibility of message content that TTS offers, but not the full range of capabilities to deal with any arbitrary text because the service provider has control over the input to the synthesis system. Given suitable tools together with the TTS system, a service provider can easily correct errors that may occur in word pronunciations or in the initial text analysis, and also 'mark up' the text to indicate, for example, which words should be stressed or where to put pauses. These types of facilities can be used to get around many of the limitations of current TTS while still providing much greater flexibility than is possible with pre-recorded messages.

Another issue that is relevant to the application of speech synthesis concerns memory requirements. At the moment the TTS systems which give the best quality use a lot of memory. A system of this type may be practical when it can be held centrally and used to service many telephone lines for example, but will generally not be an option for incorporating in a small, low-cost product.

*Examples of speech synthesis application.*

*Aids for the disabled*

One of the longest-established applications of TTS synthesis is in reading machines for the blind. The first such machine, combining an optical character reader with a TTS synthesizer, was produced by *Kurzweil Computer Products* in the 1970s. Even now, this speech synthesis task is very difficult as the machine must cope with any arbitrary text, and the quality of the speech that is generated would be regarded as insufficient by many people. However, these systems provide the visually impaired with the facility to read text that would not otherwise be available to them. Because these users are very motivated, they tend to be much more tolerant of errors and will learn to understand even low quality TTS output very well. Indeed, someone who is familiar with the speech may choose to increase the speed of speaking to several times faster than normal speed and still understand the speech well enough to successfully search for some particular part of a document which is of interest.

The requirements for aids for people with speech impairments are rather different. Here the speech synthesizer acts as a means to communicate with other people, so the speech must be intelligible, preferably natural-sounding and ideally with a voice that is appropriate to the person who is using it. However, because the user has control over what the machine is required to speak, text pre-processing is not an issue and mark-up facilities can be used to improve quality and expressiveness. Commonly required utterances can even be prepared in advance.

*Spoken warning signals, instructions and user feedback.*

Speech synthesis can be used to provide spoken warnings in emergencies. Spoken warnings are especially useful in eyes-busy, stressful environments where visual warnings may go unnoticed. A good example is the cockpit of a fighter aircraft. Speech synthesis may also be used more generally in hands-busy, eyes-busy situations such as when operating or repairing complicated equipment, to provide spoken instructions, feedback and so on. For all these types of applications, the messages may be recorded specially or a TTS-based message-preparation system can be used, depending on whether there is expected to be a requirement to change the messages. Care is needed to choose a voice that is the most effective for attracting attention in the environment in which the system is to be used.

*Education, toys and games*

Beginning with Texas Instruments' "Speak & Spell" in the 1970s, dedicated speech synthesis chips have been used in educational toys and in other toys and games. These synthesis chips are typically used to provide a fixed set of coded messages at low cost. There are also many opportunities for applying speech synthesis in the field of education. Possibilities include teaching foreign languages, teaching vocabulary and pronunciation to children learning their native language, and tools to assist in correcting speech defects. These applications can also incorporate a speech recognition component to provide feedback to learners about the accuracy of their pronunciations. For educational applications, high-quality output is normally

very important, and so recorded speech is generally used at present. However, TTS synthesis allows much greater flexibility and the recent advances in speech quality are making it more viable as an alternative.

*Telecommunications*

- Information services and interactive voice response systems

There is a vast quantity and variety of information that is stored on computers and for which there is a demand to be able to access remotely over the telephone. If the message structure is controlled and the words are not likely to change, it is practical for these systems to use recorded speech. Speaking clocks and directory enquiries services are examples for which a large number of different messages are required but the structure and vocabulary is sufficiently constrained for recorded speech to be applicable. For other applications, requiring a large vocabulary or messages of an unpredictable nature, it is more appropriate to use TTS synthesis. Examples include services providing access to public information, such as current stock market prices, news and weather reports, sports results, and so on. Other examples involve accessing more personal information, such as recent bank-account transactions, or the status of an order made through a mail-order catalogue. In many situations, rather than just passively accessing information over the telephone, a person may wish to interact with and influence the remote system. Automated systems of this type, based on spoken output, are generally referred to as interactive voice response (IVR) systems. Examples include making banking transactions, booking travel tickets and placing orders from a mail-order catalogue. In many IVR systems the person is required to communicate with the machine using a touch-tone keypad, but alternatively ASR can be used.

*Remote e-mail readers.*

A specialized but very useful application of TTS synthesis is to provide remote access to e-mail from any fixed or mobile telephone. For an e-mail reader, a full TTS conversion facility is required because the messages may contain any text characters.

E-mail messages are often especially challenging, due to the tendency to errors of spelling and grammar as well as the special nature of the language, abbreviations and so on that are often used. There are also many formatting features that are specific to e-mail. For example the message header needs to be processed appropriately to extract relevant information, such as who the message is from and when it was sent. Other important facilities include an ability to navigate through messages, with options such as repeating, going back to a previous message or on to the next one. Commands from the user to the system may be implemented using speech recognition technology, or using the telephone keypad.

There are a number of commercial products available for remote reading of e-mail. Although the quality that can be achieved using TTS synthesis is still rather limited for this application, these products can be very useful because they make it possible to keep in touch with e-mail without needing to carry a computer around.

### *5.2. Speech Recognition Technology*

The task of an ASR system is to respond appropriately to spoken human input, and the difficulty of this task is affected by a whole range of factors related to characteristics of the users' speech and the environment in which they are speaking. The main parameters which influence the difficulty of ASR tasks are:

1. *Vocabulary choice*: It is easier to distinguish between a small number of words that are acoustically very different than to choose between a much larger number of words or between words that are acoustically very similar.

2. *Speaking mode*: In isolated-word recognition tasks, the speaker leaves a gap between each word and so co-articulation effects between words are avoided. Continuous speech recognition is more difficult due to between-word co-articulation and the difficulty of locating word boundaries.

3. *Speaker enrolment*: If the task is to be performed by known individuals and each person can provide sufficient suitable speech to train the recognizer, a speaker-dependent system can be set up for each person. Speaker-independent recognition is much more difficult, as here it may be necessary to recognize speech

from any arbitrary person without knowledge of relevant factors such as gender, age group, dialect or even language. Substantial improvement over raw speaker-independent performance is possible by employing speaker adaptation techniques. Speaker adaptation is most effective if a person uses the system over a long period of time and corrects any recognition errors.

4. *Speaking style*: Read speech is generally easier to recognize than spontaneous speech, which usually contains more hesitations, errors and corrections, mispronunciations and so on. The recognition task is also easier when the talker can be trained to follow a strict syntax specifying allowed utterance constructs, than when unconstrained natural language must be accommodated. For situations in which it is applicable, a carefully designed syntax can ensure that the effective vocabulary at any one point is small and distinct, even if the total vocabulary size is large. Another aspect of speaking style is speech level: recognition performance tends to be best for speech spoken at a consistent moderate sound level, and worse when speech is shouted or whispered.

5. *Environment characteristics*: Recognition accuracy tends to be higher in a quiet environment than a noisy one, but the most important factor is to match the training environment as closely as possible to the environment in which the recognizer will be used. Conditions of time-varying noise are especially problematic. Another difficulty, which is associated with environmental characteristics, is that users often change the way they speak when the environment changes, for example shouting in an effort to be heard above the level of any noise.

6. *Channel characteristics*: In general, if the bandwidth of the speech signal is limited (as in speech transmitted over the telephone), the recognition task becomes more difficult because less information is available. Other problems that can occur with telephone-based systems include distortions due to handsets and telephone networks; cellular networks are especially problematic. More generally, the type of microphone affects the speech quality and, for example, speech recognizers tend to work better when a close-talking microphone is used than if it is necessary to use a

far-field microphone. As with other factors, performance tends to degrade with any variation in the microphone that is used.

7. *Physiological/psychological condition of speaker*: Mild illness such as a common cold can change an individual's voice characteristics. Other relevant factors include fatigue and both emotional and physical stress (such as the high g- force experienced in jet aircraft). Any change to the speaker's voice is yet more variation that can present problems to a speech recognizer.

For an ASR application to be successful, the recognizer capabilities must match the requirements for the task in terms of the parameters listed above and also in terms of recognition accuracy and any other relevant considerations, such as cost, memory and processing requirements, real-time operation, etc. It may be difficult to find a system that both satisfies the necessary economic criteria and meets all the task requirements, while giving a sufficiently high level of recognition performance. However, by making some compromises in what is required of the task, it may be possible to achieve high-enough recognition accuracy. In general the different parameters can be traded against each other so that, for example, in a hostile, stressful, noisy environment a small vocabulary of command words may be practical. On the other hand, in quiet conditions with a known user and a close-talking microphone it may be possible to achieve useful performance recognizing natural language with a large vocabulary.

*Typical recognition performance for different tasks*

When assessing the accuracy of a recognizer in operational use, it is difficult to control all the factors that may affect the performance level. However, a useful indication of performance can be obtained from laboratory tests on databases of speech that have been previously collected under known conditions. A variety of databases are available, including the ones used for the competitive evaluations. It is evident that the technology performs well enough to be applicable to digit-recognition tasks and to tasks requiring recognition of considerably larger vocabularies in a fairly constrained domain, such as airline travel information.

For large-vocabulary tasks, involving recognition of 10,000 words or more, recognition performance is greatly affected by the type of speech material. Read newspaper texts are easier to recognize than television and radio broadcasts because newspaper texts have a quite specific, consistent style. Live broadcasts, on the other hand, may contain a great variety of different material as well as particularly difficult background noise (including speech, music and so on). Conversational speech is even more challenging, especially when the conversations are between individuals who know each other very well. In these situations the familiarity between talker and listener is such that speech tends to be produced very casually, and the talker often relies on the listener using shared knowledge and experience to understand a message with minimal acoustic cues. Current ASR systems do not possess the personal knowledge that people rely on in these situations and so for this type of speech the percentage of recognition errors is several times that for read speech, even when the vocabulary size is much smaller. Thus, while large-vocabulary recognition is good enough to be deployed when the situation is constrained and the environment is controlled, performance is not yet sufficiently high for transcription of less restricted material.

*Achieving success with ASR in an application*

To be successful in an application, ASR technology must give adequate recognition performance for the required task. Word accuracy is a useful measure, especially for a task requiring accurate transcription of what a person says (dictation for example). However, when speech is used to retrieve data or to give commands, success means achieving the required result with each spoken input, not necessarily recognizing every word accurately. Some recognition errors (of function words for example) will not matter, whereas others will be critical.

Any recognition system will make errors sometimes. Users' perception of ASR technology depends very much on how errors are handled and on other aspects of the user interface. It is important to provide appropriate feedback to the user so that he or she is made aware of any recognition errors, and to provide a means for the user to

correct these errors. Sometimes the user may not speak clearly or may give a response that does not match any of the allowed options, and it is therefore often helpful to estimate 'confidence' in the recognition accuracy. If confidence is low, the system can respond by, for example, requesting clarification or repeating the allowed options.

The system must always respond to the user quickly, and allow any input to be 'undone' in the case of errors either by the system or by the user. In addition it is important to always make clear to the users what is expected of them at any point in an interaction, especially in systems designed for naive users. At the same time provision must be made for the expert user, for example to barge-in over spoken prompts. These human factors considerations are crucial to the successful application of ASR, and the detailed design of the system and its user interface will depend on the application.

*Examples of Automatic Speech Recognition (ASR)*

*Command and control*

The term "command and control" is used to refer to applications in which a person uses simple voice commands to control functions of a machine. These applications tend to be associated with situations in which hands-free, eyes-free operation is required. Voice control is best suited to functions requiring selection between a discrete set of choices, rather than to selection of continuous quantities or to positional control, and ASR is of course not suitable for safety-critical functions.

Command-and-control systems often have to work in difficult, noisy environments, possibly with the users under stress. However, many of these applications are successful with current technology because the vocabulary size tends to be small, the users are generally known to the system and in some cases may even be highly trained. Well-established applications of this type can be found in the military environment; for example, ASR has been operated successfully in fighter aircraft for functions such as setting radio frequencies and controlling flight displays, and has been included in the *Eurofighter 2000* aircraft from its earliest design stages.

Another traditional area for command-and-control applications is in factories and other industrial environments, to enable machinery to be operated without requiring hands and eyes to be distracted from the primary task.

There are also commercial command-and-control applications. For example, software packages exist which enable users to customize their PCs for voice control of functions such as menu selection, Web browsing, etc. One application area where voice control is of obvious benefit is in cars, for controlling equipment such as the car radio and, in particular, for voice-controlled dialling of mobile telephones. A number can be entered by speaking the required digit sequence or by speaking some previously programmed repertory entry, such as a name or a descriptor such as "*home*". Although repertory dialling requires the user first to train the system by speaking the required words, subsequent recognition performance will generally be better than can be obtained for long digit strings and the usability of the voice-dialling facility is greatly enhanced. Voice dialling is an attractive facility that is now included with many mobile telephones.

*Education, toys and games*

Speech recognition can be used in the field of education for a variety of applications, closely linked to the speech synthesis applications. Current uses of ASR generally involve assessing the accuracy of pronunciation of specified words. PC-based software products are available, both for foreign-language teaching and for assisting children in learning to read.

There is potentially a very large consumer market for ASR technology in games and interactive toys. Low-cost special-purpose speech recognition chips are available and have been used in toys incorporating some simple speech recognition capability. Although in the past attempts to incorporate ASR in toys have not achieved widespread success, the situation is rapidly changing with the capabilities of current technology and the growing demand for toys that are interactive.

An alternative to typing large amounts of text is to speak the words and have them transcribed automatically. Dictation applications of ASR are now established as

a commercial reality. Early products required the user to speak words in an 'isolated' style, leaving short pauses between each word, but in 1997 both *Dragon Systems* and IBM introduced PC software products that accept continuous speech. Several companies now offer ranges of products (with different capabilities and in various languages), many of which can be purchased from a computer store for less than £100. Quoted word accuracies are around 95-98%.

*Dictation products*

Current dictation products typically have active vocabularies of tens of thousands of words, but are intended for use with a close-talking microphone, in a quiet environment and in a speaker-dependent mode. Before first using the system, it is necessary to train it by speaking some specified text, and it will then continue to adapt to the individual's voice (both acoustics and choice of words) as that person uses it over a period of time. In the initial period it is likely that the system will make many errors, and care and patience are required on the part of the user to correct these errors so that the adaptation can work properly. Speaking style is also very important: speech must be clear and spoken at a steady rate, without extraneous noises such as coughs, "urs" and "ers". Over time, not only does the system adapt to the user, but committed and successful users of these products adapt their speaking style to optimize the performance of the technology. At the moment, it seems that such dedication and prolonged training are necessary to get good recognition rates. Another crucial component of these products is the user interface, and in particular the ease of error correction. If it is easy to correct errors, users' perceptions are greatly enhanced, even if the product makes mistakes.

If voice dictation is used for preparing many documents of a particular type, productivity is much improved by using 'macros' to call up standard formats (such as letters, including commonly used addresses), as well as standard paragraphs and phrases. With facilities of this type, voice dictation applications have proved to be very successful in specialized areas such as medical reporting. For example, radiologists are responsible for interpreting X-rays and reporting their findings, and

this would traditionally have been achieved by speaking into a tape recorder for later typing by a transcriptionist. There are now specially tailored ASR systems that allow radiologists to dictate directly into a computer, with resulting savings in cost and efficiency. This application, and other dictation applications involving professionals (such as doctors and lawyers) who are used to dictating documents in a very standard format, have proved very successful with current ASR technology.

*Data entry and retrieval*

We use the term "data entry" to refer to the input of information to a computer's data file (rather than dictation, which involves direct transcription of the words spoken). Data retrieval is the reverse process of accessing information that is stored in a computer system. Aside from telephone applications, which we will consider separately in the next section, typical application areas for data entry and retrieval via speech recognition involve hands-busy, eyes-busy scenarios. For example, speech recognition can be used in manufacturing to enter quality control information while inspecting product parts, and in dentistry to allow a dentist to carry out an oral examination of a patient and record the results at the same time without needing an assistant.

Data-retrieval applications of speech recognition include requesting instructions or detailed information such as specific measurements while conducting assembly or repairs. The information from the computer system can be provided using pre-recorded speech or speech synthesis.

When ASR is used to communicate with computers in military, industrial or medical applications, restrictions can be placed on the vocabulary and it is reasonable for the users to be trained to follow a defined syntax. A very different type of data-retrieval application for ASR involves cataloguing and extracting information from broadcasts or other recorded speech material. This task is very challenging because information must be extracted from material that is often completely uncontrolled. At a simple level, some classification of speech material into 'topics' (e.g. weather

forecasts) is possible by extending keyword-spotting techniques to look for groups of words that typify particular topics.

*Telecommunications*

ASR enables people to interact with computers over the telephone in a much more natural and flexible manner than is possible using only a touch-tone keypad. Some applications are aimed at cost saving by removing or reducing the need for human attendants, while others provide new services that were not previously available. The users can be expected to be more tolerant of technology limitations for the latter type of application, but any telephone system for use by the general public has to cope with a very wide variety of voices (including people in different age groups, from different dialect regions and even non-native speakers). Thus very robust speaker-independent recognition is required. In addition, users of the system will often range from experts to first-time users, and the users cannot be relied upon to respond in the way that the system expects, even when given precise instructions. For systems that are intended to recognize only a limited vocabulary for keyword spotting and detection of out-of-vocabulary words can enable some sensible response to be given to most input. More elaborate systems include some spoken language understanding capability.

*Automation of functions in telephone networks*

Voice dialling has already been mentioned above, and there are also voice-directory products that are used by several companies for internal callers and by some hotels both for employees and for guests. These systems remove the need for paper directories and make internal communication much easier. A related application is in 'automated voice attendants' which some companies and department stores are now using to answer calls made to the main switchboard and then route these calls to a named department or person. AT&T Laboratories in the U.S. have developed a sophisticated voice attendant system which has enabled a great reduction in the need for human operators in the AT&T network. Rather than restricting the user to name a

person or department, this system answers a call simply with "*How may I help you*?" and, based on the reply, enters into a dialogue with the caller to obtain additional information or clarification in order to process the call. The aim is to classify the call and pass it on to another automated system or to a human operator if necessary. Some understanding capability is needed, but not necessarily accurate recognition of every word in the utterances.

*Information services and IVR systems*

Many IVR (Information Voice Recognition) systems rely on touch-tone selection, but this method can be very restrictive and is often unpopular with users who may respond by defaulting to the human operator because it is not obvious to them how to achieve their goal with the automatic system. Speech recognition provides a more intuitive interface and an easy way to select between large numbers of different alternatives. For example, there are ASR-based systems for providing stock quotes, and *United Parcel Service* in the U.S. uses ASR for a service that allows customers to arrange collections and to track packages. A number of companies offer services whereby people can call a single number and speak keywords to access a variety of different types of information (such as restaurant listings, traffic reports, sports scores and so on).

Some companies use speech recognition to handle travel information and reservations. As demonstrated in the ATIS research, this type of application requires some understanding and dialogue capabilities if it is to deal with the wide range of likely enquiries. Speech recognition can also be used in automated telephone banking facilities, allowing customers to check on account balances, credits and debits and to conduct simple transactions. One successful example is the "*Answer*" system from NTT in Japan, which was first introduced in 1981. Although this early system could only cope with isolated words from a very limited vocabulary, it was highly successful because the user interface was well designed and the system offered obvious advantages to the users in providing them with easy and immediate access to information about their banking transactions.

*Remote access to e-mail, voice mail and messaging systems*

Speech recognition provides a natural way to access the remote e-mail reading application of TTS synthesis. ASR can also be used for remote access to conventional voice mail messages, and to the growing number of 'unified messaging' systems. The key concept in unified messaging is to provide the user with access at any time to a single system for handling e-mail, voice mail, fax and pager messages. TTS synthesis can be used to regenerate communications (such as e-mail and fax) that were not originally in spoken form.

Some companies now offer the service of a personalized 'telecommunications assistant' that integrates several functions under a single voice interface. Typically, these systems handle messaging functions, screen and forward calls, allow voice dialling by name from a contact list, and may also provide other facilities such as news and stock quotes. Automated personal assistants are a fairly new, but expanding, commercial application for speech technology.

*5.3. Applications of Speaker and Language Recognition*

While there are fewer applications for speaker recognition technology than there are for speech synthesis and speech recognition, the deployment of speaker recognition systems has increased in recent years. The applications can be divided into two general categories:

1. Authentication for access restriction and fraud prevention: A number of companies now offer speaker verification products for access control and fraud prevention. These products are often combined with speech recognition, and some are available for several different languages. Systems have been deployed for controlling access to telephone-based services, such as telephone banking and home shopping. Other applications include controlling building access, and validating users of the Internet or users of mobile phones.

2. Monitoring and forensics: Automatic speaker recognition can be used for general monitoring of voice recordings, or more specifically for checking on the

whereabouts of particular individuals. For example, speaker verification is used for the automatic monitoring of offenders who have been released under restrictions such as home detention. Forensic evidence based on identification of individuals from voice recordings has a long but controversial history. Traditionally this task is performed by a forensic phonetician, but automatic systems are sometimes used to assist in the process.

Automatic language identification has applications for surveillance and monitoring of communications, which are of interest to the military, for example. To the authors' knowledge there are not yet any automatic language recognition systems in commercial use. Potential applications include automatic routing of multilingual telephone calls. For example, calls to the emergency services could be directed to an operator who can converse in the relevant language. Language identification can also form a component of systems for multilingual speech recognition or spoken language translation, which have so far been demonstrated as research systems but which should achieve commercial realization in the future.

In conclusion, it should be said that to be successful in an application, speech technology must offer an advantage (e.g. in terms of cost, effectiveness or safety) over alternative options. Speech technology offers most benefit when manual/visual communication between human and machine is difficult: when the hands and eyes are busy, to communicate remotely via the telephone, or if the machine is very small.

For voice output, there is a choice between using recorded natural speech or an ITS system. TTS is more flexible, but the perceived quality is limited, especially for long pieces of text. A compromise that is suitable for some applications is to use a TTS-based message preparation system.

Speech synthesis applications include: aids for the disabled; systems for giving spoken warning signals, instructions and feedback to users of complex machines; voice output for toys and educational systems for teaching native or foreign languages; telephone services such as remote e-mail readers, information systems and interactive voice response (IVR) systems.

Many factors influence the difficulty of a speech recognition task: choice and size of vocabulary; speaking mode (isolated words versus continuous speech); whether or not speakers are enrolled to use the system (speaker-dependent versus speaker-independent recognition); speaking style (read versus spontaneous versus conversational speech); environment and channel characteristics; physiological and psychological condition of the speaker.

Recognition of spontaneous speech in a constrained domain (e.g. airline travel information), and of read speech for large vocabularies but under controlled conditions, is adequate for deployment now. Error rates for transcription of conversational speech are still too high for widespread application.

A good user interface, including provision for correcting the inevitable recognition errors, is crucial to the successful application of ASR technology. ASR applications include: 'command and control' of machine functions for military (e.g. fighter aircraft) and commercial applications (e.g. voice dialling for mobile telephones); educational systems for pronunciation learning; voice-controlled toys and games; voice dictation systems; data entry and retrieval; telephone services, such as automated operator services and access to e-mail readers, information services and IVR systems.

The continuing growth of the Internet, mobile telephony and ever-smaller computers offers much potential for future applications of speech technology.

## 6. Future Research Directions in Speech Synthesis and Recognition

The commercial exploitation of speech technology looks set to continue to expand in the coming years, closely linked to more general developments in information technology. Telephone and Internet applications will continue to grow. Computers are progressing in the direction of small-scale and embedded computing devices, and intelligent software 'agents' are being developed to manage interactions. With these developments, the ability to interact by voice will become more important.

Although many applications already incorporate both speech synthesis and recognition, much more integration and incorporation into multimedia interfaces is expected in the future. Future applications of speech technology will require higher capability in spoken language understanding and natural language generation, including a greater ability to deal with multiple languages and translate between languages.

The last decade of the twentieth century saw a substantial growth in the capabilities of speech technology. Current performance of speech synthesizers and recognizers makes them already extremely useful for a variety of practical tasks, and they are now deployed in many applications. However, the performance of current technology still falls far short of what is normally achieved with ease by human beings. In speech synthesis, very good quality is possible for a restricted set of messages, but if complete flexibility of message content is required, even the best systems are significantly deficient in both intelligibility and naturalness when compared with typical human speech. In recognition, even the most advanced systems cannot provide the same level of accuracy that is achievable by a competent human speaker of the target language, except when the task is so constrained that the machine has very few output choices at any one time. In both synthesis and recognition, the gap between human and machine performance widens as the conditions become more difficult, for example involving spontaneous speech, emotional speech or noisy environmental conditions.

Although the task of improving performance of speech input/output devices is not trivial, there are a number of lines of work that show considerable promise for leading to significant improvements in technology capabilities.

The first point to emphasize is that, although immense complexity will be required in more powerful systems, the availability of computational resources is increasing all the time and not likely to be the limiting factor. In addition, as a result of the large number of data collection exercises that have been conducted in recent years, there are now plenty of speech databases available for training and testing recognition and synthesis systems. What seems to be required is to develop more

powerful and robust techniques for recognition and synthesis, making better use of the resources that are already available.

### *6.1. Speech Synthesis*

The best TTS systems are now able to produce synthetic speech, in a neutral reading style, that sounds both intelligible and natural for many short passages of text. However, the systems that give the best speech quality are offered with very few different voices, and usually with very little flexibility to change the characteristics of a voice or the speaking style. Furthermore, on passages of more than just a few sentences even the best systems quickly become very boring to listen to, and some words may give problems even for short passages. If an expressive speaking style is required, such as would be appropriate when reading a story for example, the quality of speech produced by TTS systems is still not really good enough to be useful except to highly motivated users. Achieving variety, flexibility and appropriate expressiveness in speech synthesis will require research into improving all levels of the automatic speech generation process.

At the moment the best synthetic speech quality is provided by systems that use PSOLA-type concatenate techniques with a large inventory of variable-length segments. Typically the systems either use time-domain waveform concatenation or they include some waveform coding method that preserves most of the detail of the original waveform (e.g. LP-PSOLA, MBR-PSOLA). Good quality is only achieved by both a careful choice of the inventory of segments and careful extraction of suitable examples. Recent research has led to the development of automatic techniques for optimizing this selection and extraction operation, so that the process of setting up concatenative systems for new talkers is becoming easier. However, with a waveform-based coding method, the only changes that can be made are in the selection of the segments and there is no obvious way to enforce appropriate formant transitions across segment boundaries, or to model systematic changes in formant frequencies or bandwidths for example. This deficiency is widely recognized and is

motivating current research into representations and methods that allow some co-articulatory and other spectral changes to be made to the concatenated units, while still retaining the high-quality coding of the speech. For example, if units are coded in a way that can be related to formant frequencies fairly reliably, it is possible to apply some small formant modifications such as are required for smoothing transitions across segment boundaries.

The search for representations and methods that facilitate greater manipulation of speech characteristics within a waveform-based concatenative framework will probably continue to be a focus for speech synthesis research for several years to come. However, even with these methods it seems unlikely that it will be possible to model co-articulation phenomena or to vary the voice characteristics to the same extent that can be achieved with rule-driven approaches.

Current output of phonetic synthesis by rule is considerably worse than that of the best concatenative systems and is unlikely to be mistaken for a recording of human speech, even if correct phonemic and prosodic specifications are provided. The potential sources of the deficiencies are in the speech production model and in the rules for controlling it. However, demonstrations made nearly 30 years ago showed for a few sentences that a parallel formant synthesizer could produce synthetic utterances that were almost indistinguishable perceptually from recordings of the natural utterances that they were copying. There is thus good evidence that the limitations of synthesis by rule are almost entirely in the rules for converting a phonemic/prosodic description into control signals for the synthesizer.

The difficulty in choosing appropriate phonetic rules to mimic real speakers has been the major obstacle to achieving truly natural-sounding synthetic speech by rule. While automatic training methods are well-established in speech recognition, and have more recently also been used for concatenative synthesis systems, they have not been widely adopted in the case of phonetic synthesis by rule. However, automatic training methods are applicable to a rule system. More generally, if an appropriate synthesis model were available, it should be possible to train its parameters automatically. There is some recent research interest in developing this type of

statistical speech model for application both to recognition and to synthesis. In the long term, this line of research may lead to the best solution to achieving natural-sounding speech synthesis, by modelling speech dynamics and capturing the effects of co-articulation. A model for variability may be required to prevent the synthetic speech from sounding too monotonous.

An advantage of automatic techniques is that they can be applied to derive synthesis parameters for any talker of any language or dialect, given enough labelled speech data to train the system. It should also be possible to apply speaker adaptation techniques to transform an existing set of models based on just a small quantity of speech data from a new talker. The long-term aim should be to develop a synthesis model that characterizes all the attributes that distinguish different individuals' voices. This way it may ultimately be possible to achieve truly flexible and natural synthesis of any specified voice quality on demand, rather than relying on a speech database for a particular talker as is the case at the moment.

Prosody is often cited as the major limitation to the quality of the speech generated by current TTS systems. Recent research has led to the development of automatic techniques for deriving the parameters of prosodic models, but further work is still needed to improve these models and to find the best automatic methods for training them. If the sound generation component is developed to provide an accurate model of co-articulation effects, the realization of different sounds should then vary appropriately with changes in the timing of articulatory movements, which should in turn facilitate development of improved models of timing. Intonation prediction may be more complex because it depends on the choice of an abstract representation that captures every attribute that contributes to the characteristics of intonation. Prosodic transcription schemes such as *TOBI* (J.Silverman et al, 1992) have proved useful, but may not capture all the relevant information, especially if different intonational correlates of emotion need to be included.

Whatever prosodic labelling scheme is used, in a practical TTS system it will be necessary to derive the prosodic labels, as well as the phonemic labels, from analysis of text. It is these higher levels of TTS conversion that seem to present the most

difficult challenge. Conversion of arbitrary text into a really accurate, detailed phonemic and prosodic description must require at least some understanding of the meaning of the text. Even if a representation of the underlying concepts is available, the problem of converting from concept to the phonemic and prosodic specification seems just as daunting. However, applications that require speech output within a limited domain (such as train timetable enquiry systems) are more manageable. Because such systems normally only need to offer a restricted range of messages in a known domain, relevant information such as syntactic phrasing and semantic focus is relatively easy to obtain. There is, however, still a need for better models to capture the relationship between this semantic and syntactic information and the required prosodic structure.

TTS systems of the future will need to be able to speak in different styles depending on the type of text (e.g. news report, e-mail message, children's story). Research is needed to find the best way to model these stylistic differences with controllable synthesis parameters, and to find methods for automatically training the models from suitable text and speech data. Sophisticated text analysis will also be required to automatically determine the style and structure of documents.

### *6.2. Automatic Speech Recognition*

Current ASR performance can be very impressive, even for tasks involving very large vocabularies. However, some marked deficiencies remain. In particular, ASR systems tend to be very sensitive to variation: changes in the acoustic environment, transmission channel, talker identity, speaking style and so on all cause much more problem for recognition by machines than for recognition by humans. The most successful ASR systems to date have been almost universally based on HMMs. Over the years there have been many refinements to the way in which the HMMs are used, and it seems likely that these incremental improvements will continue in the immediate future. However, there is also research interest in more substantial changes to and advances beyond the current HMM methods. The hope is that this research could eventually lead to a step improvement in recognition performance, especially in

terms of robustness to all the inevitable, but often systematic, variability that is found in speech.

While current HMM pattern-matching methods have serious limitations, they also have some very desirable properties, as already explained in previous chapters. In this section we will revisit some of the important advantages, before discussing the limitations in the next section.

The HMM methodology provides a tractable mathematical framework with straightforward algorithms for recognition and for training to match some given speech data. The spectral characteristics and the temporal characteristics (the Markov chain with its transition probabilities) are treated separately but within the one consistent framework. As a consequence, segmentation of an utterance arises automatically as part of the training and recognition processes. In addition the models can be made to generalize quite naturally to unseen data, either by smoothing estimated discrete distributions (typically used for language models), or by using a parameterized continuous distribution such as a multi-variate Gaussian (now widely used for acoustic models).

During recognition, a result is only output when the partial trace-back through possible word sequences coalesces into a single path. This coalescence can cause the identities of a whole sequence of words to be determined simultaneously, and in fact the implied decision about the phonemic content of an early word in the sequence can be changed as a result of either acoustic or linguistic evidence for a later word. For example, assume that an early word is acoustically ambiguous between two possibilities. If linguistic knowledge (as expressed in a language model) indicates that a later word for which there is strong acoustic evidence could not follow one of the early candidates, the overall decision on the utterance will be biased strongly against that earlier word.

It can thus be seen that for any given utterance one-pass continuous recognition algorithms make a single decision about the word sequence as soon as they can reliably do so, after weighing up all the available evidence, both acoustic and linguistic. Provided no significant information has been lost in the acoustic analysis,

the decision is made without prematurely discarding any relevant information. Experiments with human speech perception (e.g. Marslen-Wilson, 1980) strongly suggest that human speech recognition behaves in a similar way.

Early systems for large-vocabulary recognition used a knowledge-based approach, whereby recognition was attempted by trying to detect and recognize phonetic features. The comparatively poor performance shown by these systems is due to the difficulties involved both in identifying all the required knowledge for performing speech recognition and in finding an appropriate way for specifying that knowledge. In contrast the stochastic methods require only a general structure whose parameters are trained automatically using a large amount of training data. The best HMM systems incorporate knowledge about speech, but this knowledge takes the form of constraints on the more general model structure. Examples include the choice of unit inventory (e.g. context-dependent sub-word models) and the selection of a model topology that only allows a subset of plausible transitions. Typical methods of acoustic feature analysis are also chosen taking account of knowledge about the phonetically important characteristics of a speech signal that need to be preserved.

HMMs provide a structure that is broadly appropriate to represent the spectral and temporal variation in speech. However, some assumptions are made in the HMM formalism that are clearly inappropriate for modelling speech patterns. Firstly, it is assumed that a speech pattern is produced by a piece-wise stationary process, with instantaneous transitions between stationary states. This assumption is in direct contradiction with the fact that speech signals are produced by a continuously moving physical system – the vocal tract. Secondly, in a first-order Markov model, the only modelling of dependency between observations occurs through constraints on possible state sequences. Successive observations generated by a single HMM state are treated as independent and identically distributed. By making this independence assumption, the model takes no account of the dynamic constraints of the physical system that has generated a particular sequence of acoustic data, except inasmuch as these constraints can be incorporated in the feature vector associated with a state. In a typical speaker-independent HMM recognizer in which each modelling unit is

represented by a multi-modal Gaussian distribution to include all speakers, the model in effect treats each frame of data as if it could have been spoken by a different speaker.

HMM recognition systems are usually designed to reduce the impact of the inappropriate modelling assumptions. For example, a generous allocation of states allows a sequence of piece-wise stationary segments to make a fair approximation to speech dynamics, and time-derivative features help to mitigate the effects of the independence assumption as well as capturing some information about local dynamics explicitly.

HMMs are well suited to modelling acoustic features obtained by short-time spectral analysis using a fixed window size (typically 20-25 ms) at fixed time intervals (typically 10ms). It is well known that this type of analysis has to compromise between capturing temporal properties and representing spectral detail, and is not a very good model for many properties of human auditory perception. For example, studies of human speech perception have shown the importance of dynamic information over many different timescales, ranging from as short as 2-3 ms to around 20-50 ms. In the case of prosodic information, much longer time intervals are also relevant. In order to incorporate such information operating at this wide range of timescales, it seems necessary to modify not only the methods of feature analysis, but also the nature of the models. For example, prosody provides information that helps in speech understanding, but simply adding prosodic information to an acoustic feature vector at 10ms intervals for modelling with HMM states seems unlikely to capture the necessary prosodic cues.

A first-order Markov process cannot capture more than very immediate linguistic influences, and long-range syntactic and semantic constraints are difficult to incorporate. As is the case for speech synthesis, recognition systems of the future will need more powerful models of language understanding, especially when dealing with spontaneous or noisy speech. We will return to this issue in Section 16.5 after first discussing how the acoustic modelling might be improved.

At present data-driven statistical methods have proved to give better recognition performance than knowledge-based methods, even though as presently formulated the data-driven systems ignore much of the information in the acoustic signal that we know is important for human speech recognition. In principle the ability to learn by example that is characteristic of the data-driven approach could be extended to incorporate a richer structure and to learn more complicated phonetic features.

One way of addressing the HMM assumptions of independence and piece-wise stationarity is to associate models with variable-length sequences, or 'segments', of acoustic feature vectors. It is then possible to characterize both the duration of the segments and the relationship between the vectors in the sequence associated with a state, usually incorporating the concept of a trajectory to describe how the features change over time in the segment. A variety of segment models have been investigated, using different trajectories and different ways of describing have been investigated, using different trajectories and different ways of describing the probability distributions associated with those trajectories. In comparison with conventional HMMs, some improvements in recognition performance have been demonstrated by modelling trajectories of typical acoustic feature vectors. However, success is often dependent upon a careful choice of trajectory model and distribution modelling assumptions. It seems that, when introducing more structural assumptions into the model, the accuracy of the assumptions is critical to success. If the assumptions are not sufficiently accurate, performance may actually be worse than the performance of a conventional HMM, which, although it is a crude model, makes only a few very general assumptions.

The general concept of modelling the relationship between successive acoustic feature vectors seems desirable. However, the motivation for modelling dynamics and trajectories originates in the nature of speech production. It may therefore not be most appropriate to apply these models directly to transformed acoustic features, which have a very complex relationship with the speech production system. To obtain the full benefit of trajectory modelling, it may be necessary to apply the models to features that are more directly related to speech production. These features could be

some form of articulatory features, or alternatively they could be acoustic features that are closely linked with articulation, such as formant frequencies. Some success has been achieved in extracting and using articulatory and formant information for speech recognition, especially as a supplement to general acoustic features. However, formant or articulatory analysis is very difficult to perform reliably without any prior knowledge about the speech sounds, due to the complex many-to-one mapping that exists between articulation and its acoustic realization. Furthermore, articulatory or formant features do not provide all the information that is needed to make certain distinctions, such as those relying on excitation source differences.

The desirability of modelling articulatory or formant dynamics, together with the difficulties involved in extracting the relevant information, have led a number of workers to suggest that this information is best incorporated in a multiple-level modelling framework. The idea is to introduce an intermediate level between the abstract phonological units and the observed acoustic features. This intermediate level represents trajectories of articulators, formants or other parameters closely related to speech production. Some complex mapping is required between this underlying level and the observed acoustic features. The aim is for the trajectory model to enforce production-related constraints on possible acoustic realizations without requiring explicit extraction of articulatory or formant information. To gain the full benefit of such a model, it is important to incorporate a model of co-articulation in the underlying trajectories, so that different trajectories arise naturally for different sequences of speech sounds rather than requiring very many context-dependent models. For example, some approaches model a sequence of hidden targets which are then filtered to obtain a continuously evolving trajectory.

While a lot of further research will be needed, multiple-level models that capture some salient aspects of speech production would seem to be a promising line of investigation, which may lead to more powerful, constraining models than are provided by current HMM systems. In particular, these models should provide a meaningful way of capturing differences, both between talkers and within any one individual, due to effects such as stress or simply differences in speaking rate.

Variations of this type can be expected to be much more systematic, and hence predictable, at the level of the speech production system. At this level it should therefore be much easier to adapt to changes than when simply linking acoustic variation directly to the phonological units, as is the case for current HMM systems. It should also be possible to move away from the rather artificial notion in current recognition (and synthesis) systems that speech can be represented as a sequence of contiguous but distinct phonetic segments. Many modem phonological theories view speech as being generated in terms of several different articulatory features operating at different and overlapping timescales, and it is easier to see how these might be accommodated in a model with an intermediate layer that is related to articulation.

A related issue concerns the choice of acoustic features upon which any recognition process must operate. Ideally the feature analysis should preserve all the perceptually relevant information in the speech signal. The typical long analysis window blurs highly transient events such as the bursts of stop consonants and rapid formant transitions, yet these portions of the signal convey some of the most important perceptual information. It ought to be advantageous to make more use of auditory models in ASR systems. The human auditory system shows sensitivity to transitional information at a range of timescales, and automatic systems should be improved by the development of better methods for both extracting and modelling the relevant information.

The need to model information at different timescales has recently been addressed by research into extending HMMs to use multiple feature sets in parallel, with the associated probabilities being combined at some stage as part of the recognition process. These multi-stream methods potentially provide a way of incorporating many diverse information sources, including those obtained at different timescales.

The modelling approaches that have been mentioned above are just a few of the ideas that are currently being pursued as part of research aimed at improved speech modelling for ASR. More generally, there is also a growing interest in applying statistical formalisms that are used in other areas of pattern recognition, including

methods that can be viewed as more powerful generalizations of the first-order HMMs that are still the most widely used model for ASR at present.

### *6.3. Relationship between Synthesis and Recognition*

Most speech research groups have for many years tended to specialize in one particular facet of speech processing, and it has been fairly unusual for the same research workers to be involved in both speech synthesis and speech recognition. In the past the actual techniques that have been used in the two areas have seemed to be almost completely unrelated. Yet it is evident that, for real advancement in both subjects, the predominant need is for knowledge about the structure of speech and its relationship to the underlying linguistic content of utterances. There has been much debate about the relationship between speech production and speech perception in humans, and similar issues apply to the development of automatic systems. However at the very least it ought to be beneficial to take more account of
the constraints of production in speech recognition and to take more advantage of perceptual influences in speech synthesis.

In recent years there has been some convergence of the two fields, as automatic data-driven techniques have become widely adopted in synthesis as well as in recognition. At present, the most successful systems for both technologies use large inventories of acoustic segments and make minimal assumptions about the underlying structure of those segments. HMMs have been used to identify segments for use in concatenative synthesis. Of course an HMM is a generative model and can therefore be viewed as a synthesizer, but a rather crude one that generates an acoustic signal as a sequence of piecewise-stationary chunks. In the previous section we argued that ASR could be improved by using a more appropriate model of speech production, to capture the dynamic properties of speech in a better way by somehow modelling the human speech production mechanisms more closely. For example, in synthesis it is necessary to generate a speech waveform, but for recognition it will probably be better to work directly from some analysis of this waveform.

*6.4. Automatic Speech Understanding*

There is a growing demand for interactive spoken-language systems that involve a two-way dialogue between a person and a computer. In the future greater naturalness will be required both in the language that the human can use and in the responses generated by the system. Such naturalness is only likely to be achievable if the machine has a good model of the interaction and some 'understanding' of the information being communicated. The difficulties here are generally regarded as problems of artificial intelligence, but most language processing work in artificial intelligence has so far only considered textual forms of language. Although the achievements in this field are impressive, spoken language poses additional challenges. In particular, spontaneous speech can be vastly different from read speech: not only does the speech tend to be more casual and include hesitations, corrections and so on, but the use of language is very different when it is part of an interactive communication. Future conversational systems will need to model these effects, both for high-performance speech recognition and for natural-sounding speech generation. Another aspect of growing importance is a demand for spoken-language systems to be multilingual. A truly multilingual system needs to include an underlying representation of concepts, together with methods for relating those concepts to utterances in different languages in ways that exploit the commonality between different languages while also modelling the differences between them. Such a capability is probably necessary if major advances are to be achieved in the most challenging problem of spoken language translation (requiring recognition in one language and synthesis in another, with a translation stage in between the two). Artificial intelligence methods will need to use information about the current speech communication task whether performing synthesis or recognition. In particular, knowledge about the subject matter is extremely important for producing and interpreting utterances in man-machine dialogue, and must include the effect of previous utterances on the expectations of what will follow. So far, the best spoken

dialogue systems have been the result of a lot of hand-tuning specific to their application domain (such as air travel), so setting up a system for a new domain is time-consuming and labour-intensive. Good automatic methods are needed for training models for the domain semantics from appropriate existing material for the relevant domain. The processes that will be needed to interpret the phonetic and prosodic properties of speech signals as text or as concepts will have their counterparts in going from text or concept into speech, and both directions of processing need to be taken into account in dialogue design. Improvements in both synthesis and recognition technologies should come about with the development of better models of spoken language, capturing all levels in the relationship between abstract linguistic concepts and the generated acoustic signal.

## Conclusion

Although speech synthesis and recognition technology are now good enough to be useful in many applications, performance is still poor in comparison with that of humans. The problem is not in the computational power achievable with electronic technology. Large quantities of data are now available for training, but better models are needed to make the most effective use of these data.

Data-driven methods have been applied to concatenative synthesis but have not yet been widely applied to synthesis by rule. A model-driven approach to synthesis offers the scope to capture co-articulation, and to include variability and flexibility in a way that is not possible with concatenative methods.

The most difficult synthesis problems are in making the style of speech appropriate for the intended meaning. Development of artificial intelligence techniques will be necessary.

Future developments in ASR will need to retain a data-driven approach to training, and a recognition framework that delays decisions until there is sufficient evidence and does not discard information too early. Improvements over current methods should be possible by incorporating a richer structure in the model. Promising developments include the use of multiple-layer frameworks to incorporate constraints of speech production and models that use parallel streams to capture information at differing timescales.

Advanced systems for both synthesis and recognition need the same knowledge about speech and language, including an understanding capability. It should therefore be advantageous for the two applications to be studied together.

# CHAPTER VII.
# LINGUISTICS AND ROBOTICS

## 1. Computational Linguistics and Robotics

The goal of Computational Linguistics in the domain of Robotics is to reproduce the natural transmission of information by modeling the speaker's production and the hearer's interpretation on a suitable type of a computer. This amounts to the construction of autonomous cognitive machines (robots) which can communicate freely in natural language.

*Turing test.*

The task of modeling the mechanism of natural communication on the computer was described in 1950 by Alan Turing (1912-1954) in the form of an 'imitation game' known today as the Turing test. In this game, a human interrogator is asked to question a male and a female partner in another room via a teleprinter in order to determine which answer was given by the man and which by the woman. The people running the test count how often the interrogator classifies his communication partners correctly and how often (s)he is fooled by them.

Subsequently one of the two humans is replaced by a computer. The computer passes the Turing test if it simulates the man or the woman which it replaced so well that the guesses of the interrogator are just as often right and wrong as with the previous set of partners. In this way Turing wanted to replace the question "*Can machines think?*" by the question "*Are there imaginable digital computers which would do well in the imitation game?*"

*Eliza program.*

In its original intention, the Turing test requires the construction of an artificial cognitive agent with a verbal behavior so natural that it cannot be distinguished from

that of a human native speaker. This presupposes complete coverage of the language data and of the communicative functions in real time. At the same time, the test tries to avoid all aspects not directly involved in verbal behavior.

However, the Turing test does not specify what cognitive structure the artificial agent should have in order to succeed in the imitation game. For this reason, it is possible to misinterpret the aim of the Turing test as fooling the interrogator rather than providing a functional model of communication on the computer. This was shown by the Eliza program of J.Weizenbaum (1965).

The Eliza program simulates a psychiatrist encouraging the human interrogator to talk more and more about him- or herself. The structure of Eliza is based on sentence templates into which certain words used by the interrogator, now in the role of a patient, are inserted. For example, if the interrogator mentions the word *mother*, Eliza uses the template *"Tell me more about your…"* to generate the sentence "*Tell me more about your mother"*.

Because of the way in which Eliza works, we know that Eliza has no understanding of the dialog with the interrogator/patient. Thus, the construction of Eliza is not a model of communication. If we regard the dialog between Eliza and the interrogator/patient as a modified Turing test, however, the Eliza program is successful insofar as the interrogator/patient feels him- or herself understood and therefore does not distinguish between a human and an artificial communication partner in the role of the psychiatrist.

*The purpose of computational linguistics* is the real modeling of natural language communication, and not a mimicry based on exploiting particular restrictions of a specific dialog situation, as in the Eliza program. Thus, computational linguistics must:

1) explain the mechanism of natural communication theoretically;
2) verify this explanation in practice. The latter is done in terms of a complete and general implementation which must prove its functioning in everyday communication rather than in the Turing test.

## 2. Modeling Natural Communication

Designing a talking robot provides an excellent occasion for systematically developing the basic notions as well as *the philosophical, mathematical, methodological, and programming aspects of computational linguistics*. This is because modeling the mechanism of natural communication requires:

- a theory of language which explains the natural transfer of information in a way that is functionally coherent, mathematically explicit, and computationally efficient;

- a description of language data which is empirically complete for all components of this theory of language, i.e., the lexicon, the morphology, the syntax, and the semantics, as well as the pragmatics and the representation of the internal context;

- a degree of precision in the description of these components which is sufficient for computation.

Fulfilling these requirements will take hard, systematic, goal-oriented work, but it will be worth the effort.

*For theory development*, the construction of talking robots is of interest because an electronically implemented model of communication may be tested both externally in terms of the verbal behavior observed, and internally via direct access to its cognitive states. The work towards realizing unrestricted human-computer communication in natural language is facilitated by the fact that the functional model may be developed incrementally, beginning with a simplified, but fully general system to which additional functions as well as additional natural languages are added step by step.

*For practical purposes*, unrestricted communication with computers and robots in natural languages will make the interaction with these machines maximally user friendly and permit new, powerful ways of information processing. Artificial

programming languages may then be limited to specialists developing and servicing the machines.

*Using parsers*

Computational linguistics analyzes natural languages automatically in terms of software programs called *parsers*. The use of parsers influences the theoretical viewpoint of linguistic research, distribution of funds, and everyday research practice as follows:

- *Competition*

Competing theories of grammar are measured with respect to the new standard of how well they are suited for efficient parsing and how well they fit into a theory of language designed to model the mechanism of natural communication.

- *Funding*

Computationally efficient and empirically adequate parsers for different languages are needed for an unlimited range of practical applications, which has a major impact on the inflow of funds for research, development, and teaching in this particular area of the humanities.

- *Verification*

Programming grammars as parsers allow testing their empirical adequacy automatically on arbitrarily large amounts of real data in the areas of word form recognition/synthesis, syntactic analysis/generation and semantic-pragmatic interpretation in both the speaker and the hearer mode.

The verification of theories of language and grammar by means of testing electronic models in real applications is a new approach which clearly differs from the methods of traditional linguistics, psychology, philosophy, and mathematical logic.

*Theoretical levels of abstraction*

So far there are no electronic systems which model the functioning of natural communication so successfully that one can talk with them more or less freely.

Furthermore, researchers do not agree on how the mechanism of natural communication really works. One may therefore question whether achieving a functional model of natural communication is possible in principle.

Today's situation in computational linguistics resembles the development of mechanical flight before 1903. For hundreds of years humans had observed sparrows and other birds in order to understand how they fly. Their goal was to become air-borne in a similar manner. It turned out, however, that flapping wings did not work for humans. This was taken by some as a basis for declaring human flight impossible in principle, in accordance with the pious cliché "If God had intended humans to fly, He would have given them wings."

Today human air travel is commonplace. Furthermore, we now know that a sparrow remains air-borne in accordance with the same aero-dynamic principles as a jumbo jet. Thus, there is a certain level of abstraction at which the flights of sparrows and jumbo jets function in the same way.

Similarly, the modeling of natural communication requires an abstract theory which applies to human and artificial cognitive machines alike. Thereby, one naturally runs the risk of setting the level of abstraction either too low or too high. As in the case of flying, the crucial problem is finding the correct level of abstraction.

A level of abstraction which is too low is exemplified by closed signal systems such as vending machines. Such machines are inappropriate as a theoretical model because they fail to capture the diversity of natural language use, i.e., the characteristic property that one and the same expression can be used meaningfully in different contexts.

A level of abstraction which is too high, on the other hand, is exemplified by naive anthropomorphic expectations. For example, a notion of 'proper understanding' which requires that the computational system be subtly amused when scanning Finnegan's Wake is as far off the mark as a notion of 'proper flying' which requires mating and breeding behavior from a jumbo jet.

## 3. Human cognition analysis

The history of mechanical flight shows how a natural process (bird flight) proposes a conceptually simple and obvious problem to science. Despite great efforts it was un-solvable for a long time. In the end, the solution tamed out to be a highly abstract mathematical theory. In addition to being a successful foundation of mechanical flight, this theory is able to explain the functioning of natural flight as well.

This is why the abstract theory of aero-dynamics has led to a new appreciation of nature. Once the development of biplanes, turboprops, and jets resulted in a better theoretical and practical understanding of the principles of flight, interest was refocused again on the natural flight of animals in order to grasp their wonderful efficiency and power. This in turn led to major improvements in artificial flight, resulting in less noisy and more fuel-efficient air planes.

Applied to computational linguistics, this analogy illustrates that our highly abstract and technological approach does not imply a lack of interest in the human language capacity. On the contrary, investigating the specific properties of human language communication is theoretically meaningful only after the mechanism of natural language communication has been modeled computationally and proven successful in concrete applications on massive amounts of data.

### *3.1. Linguistic verification*

In science we may distinguish between internal and external truths. *Internal truths* are conceptual models, developed and used by scientists to explain certain phenomena, and held true by relevant parts of society for limited periods of time. Examples are the Ptolemaic (geocentric) view of planetary motion or Bohr's model of the atom.

*External truths* are the bare facts of external reality which exist irrespective of whether or not there are cognitive agents to appreciate them. These facts may be measured more or less accurately, and explained using conceptual models.

Because conceptual models of science have been known to change radically in the course of history, internal truths must be viewed as *hypotheses*. They are justified mainly by the degree to which they are useful for arriving at a systematic description of external truths, represented by sufficiently large amounts of real data.

Especially in the natural sciences, internal truths have improved dramatically over the last five centuries. This is shown by an increasingly close fit between theoretical predictions and data, as well as a theoretical consolidation exhibited in the form of greater mathematical precision and greater functional coherence of the conceptual (sub)models.

In contrast, contemporary linguistics is characterized by a lack of theoretical consolidation, as shown by the many disparate theories of language and the overwhelming variety of competing theories of grammar. As in the natural sciences, however, there is external truth also in linguistics. It may be approximated by completeness of empirical data coverage and functional modeling.

The relation between internal and external truth is established by means of *a verification method*. The verification method of the natural sciences consists in the principle that experiments must be repeatable. This means that, given the same initial conditions, the same measurements must result again and again.

On the one hand, this method is not without problems because experimental data may be interpreted in different ways and may thus support different, even conflicting, hypotheses. On the other hand, the requirements of this method are so minimal that by now no self-respecting theory of natural science can afford to reject it. Therefore the repeatability of experiments has managed to channel the competing forces in the natural sciences in a constructive manner.

Another aspect of achieving scientific truth has developed in the tradition of *mathematical logic*. This is the principle of formal consistency, as realized in the method of axiomatization and the rule-based derivation of theorems.

Taken by itself the quasi-mechanical reconstruction of mathematical intuition in the form of axiom systems is separate from the facts of scientific measurements. As the logical foundation of natural science theories, however, the method of axiomatization has proven to be a helpful complement to the principle of repeatable experiments.

In linguistics, corresponding methods of verification have been sorely missed. To make up for this shortcoming there have been repeated attempts to remodel linguistics into either a natural science or a branch of mathematical logic. Such attempts are bound to fail, however, for the following reasons:

- *The principle of repeatable experiments* can only be applied under precisely defined conditions suitable for measuring. The method of experiments is not suitable for the objects of linguistic description because they are conventions that have developed over the course of centuries and exist as the intuitions ('Sprachgefiihl') of the native speaker-hearer.

- *The method of axiomatization* can only be applied to theories which have consolidated on a high level of abstraction, such as Newtonian mechanics, thermodynamics, or the theory of relativity. In today's linguistics, there is neither the required consolidation of theory nor completeness of data coverage. Therefore, any attempt at axiomatization in current linguistics is bound to be empirically vacuous.

Happily, there is no necessity to borrow from the neighboring sciences in order to arrive at a methodological foundation of linguistics. Instead, theories of language and grammar are to be implemented as electronic models which are tested automatically on arbitrarily large amounts of real data as well as in real applications of spontaneous human-computer communication. This method of verifying or falsifying linguistic theories objectively is specific to computational linguistics and may be viewed as the counterpart of the repeatability of experiments in the natural sciences.

### *3.2. Empirical data and their theoretical framework*

The methodology of computational linguistics presupposes a theory of language which defines the goals of empirical analysis and provides the framework into which components are to be embedded without conflict or redundancy. The development of such a framework can be extraordinarily difficult, as witnessed again and again in the history of science.

For example, in the beginning of astronomy scientists wrestled for centuries with the problem of providing a functional framework to explain the measurements that had been made of planetary motion and to make correct predictions based on such a framework. It was comparatively recently that J.Kepler (1571-1630) and I.Newton (1642-1727) first succeeded with a description which was both empirically precise and functionally simple. This, however, required a radical revolution in the theory of astronomy.

The revolution affected the structural hypothesis (transition from geo- to heliocentrism), the functional explanation (transition from crystal spheres to gravitation in space), and the mathematical model (transition from a complicated system of epicycles to the form of ellipses). Furthermore, the new system of astronomy was constructed at a level of abstraction where the dropping of an apple and the trajectory of the moon are explained as instantiations of one and the same set of general principles.

In linguistics, a corresponding scientific revolution has long been overdue. Even though the empirical data and the goals of their theoretical description are no less clear in linguistics than in astronomy, linguistics has not achieved a comparable consolidation in the form of a comprehensive, verifiable, functional theory of language.

## 4. The SLIM theory of language

The analysis of natural communication should be structured in terms of methodological, empirical, ontological, and functional principles of the most general kind. *The SLIM theory of language* is based on surface compositional, linear, internal matching. These principles were defined by R.Hausser (2002) as follows:

1. *S*urface compositional (methodological principle)

Syntactic-semantic composition assembles only concrete word forms, excluding the use of zero-elements, identity mappings, or transformations.

2. *L*inear (empirical principle)

Interpretation and production of utterances are based on a strictly time-linear derivation order.

3. *I*nternal (ontological principle)

Interpretation and production of utterances are analyzed as cognitive procedures located inside the speaker-hearer.

4. *M*atching (functional principle).

Referring with language to past, current, or future objects and events is modeled in terms of pattern matching between language meaning and context.

These principles originate in widely different areas (methodology, ontology, etc.), but within the SLIM theory of language they interact very closely. For example, the functional principle of (4) matching can only be implemented on a computer if the overall system is handled ontologically as (3) an internal procedure of the cognitive agent. Furthermore, the methodological principle of (1) surface compositionality and the empirical principle of (2) time-linearity can be realized within a functional mechanism of communication only if the overall theory is based on internal matching (3,4). In addition to what its letters stand for, the acronym *slim* is motivated as a word with a meaning like *slender*. This is so because detailed mathematical and computational investigations have proven SLIM to be efficient in the areas of syntax, semantics, and pragmatics - both relatively in comparison to existing alternatives, and absolutely in accordance with the formal principles of mathematical complexity theory.

The SLIM theory of language is defined on a level of abstraction where the mechanism of natural language communication in humans and in suitably constructed cognitive machines is explained in terms of the same principles of surface compositional, linear, internal matching.

Moreover, the structural hypothesis of the SLIM theory of language is a regular, strictly time-linear derivation order - in contrast to grammar systems based on constituent structure. The functional explanation of SLIM is designed to model the mechanism of natural communication as a speaking robot - and not some tacit language knowledge innate in the speaker-hearer which excludes language use (performance). The mathematical model of SLIM is the continuation-based algorithm of LA-grammar (Left-associative derivation order), and not the substitution-based algorithms of the last 50 years.

This is an important precondition for unrestricted human- computer communication in natural language. Its realization requires general and efficient solutions in the following areas.

First, the hearer's understanding of natural language must be modeled. This process is realized as the automatic reading-in of propositions into a database and - most importantly - determining the correct place for their storage and retrieval. The foundation of the semantic primitives is handled in terms of natural or artificial recognition and action.

Second, how the speaker determines the contents to be expressed in language must be modeled. This process, traditionally called conceptualization, is realized as an autonomous navigation through the propositions of the internal database. Thereby speech production is handled as a direction reflection (internal matching) of the navigation path in line with the motto: Speech is verbalized thought.

Third, the speaker and the hearer must be able to draw inferences on the basis of the contents of their respective databases. Inferences are realized as a special form of the autonomous time-linear navigation resulting in the derivation of new propositions. Inferences play an important role in the pragmatic interpretation of natural language, both in the hearer and the speaker.

The formal basis of time-linear navigation consists in concatenated propositions stored in a network database as a set of word tokens. A word token is a feature structure with the special property that it explicitly specifies the possible continuations to other word tokens, both within its proposition and from its proposition to others. This

novel structure is called a word bank and provides the 'railroad tracks' for the navigation of a mental focus point. The navigation is powered and controlled by suitable LA-grammars (motor algorithms) which compute the possible continuations from one word token to the next.

The word bank and its motor algorithms constitute the central processing unit of an artificial cognitive agent called slim machine. The word bank is connected to external reality via the SLIM machine's recognition and action. The interpretation of perception, both verbal and nonverbal, results in concatenated propositions which are read into the word bank. The production of action, both verbal and nonverbal, is based on realizing some of the propositions traversed during the autonomous navigation.

The vision of unrestricted natural language communication between humans and machines is like the vision of motorized flight a hundred years ago: solved theoretically, but not yet realized in practical systems. At this point, all it will take to really succeed in computational linguistics is a well-directed, concentrated sustained effort in cooperation with robotics, artificial intelligence, and psychology.

## 5. Human-Machine Communication

Computers are comfortable for entering, editing, and retrieving natural language, at least in the medium of writing, for which reason they have replaced electric typewriters. For utilizing the computers' abilities beyond word processing, however, commands using artificial languages must be applied. These are called programming languages, and are especially designed for controlling the computer's electronic operations.

In contrast to natural languages, which are flexible and rely on the seemingly obvious circumstances of the utterance situation, common background knowledge, the content of earlier conversations, etc., programming languages are inflexible and refer directly, explicitly, and exclusively to operations of the machine. For most potential users, a programming language is difficult to handle because (a) they are not familiar

with the operations of the computer, (b) the expressions of the programming language differ from those of everyday language, and (c) the use of the programming language requires great precision.

Such an expanded notion of human-machine communication should be avoided, however, because it fosters misunderstandings. Machines without general input/output facilities for language constitute the special case of nonverbal human-machine communication, which may be neglected for the purposes of computational linguistics.

Consider, for example, a standard database which stores information about the employees of a company in the form of records:

| | **Last name** | **First name** | **place** | **…** |
|---|---|---|---|---|
| A1 | Schmidt | Peter | Bamberg | … |
| A2 | Meyer | Susanne | Nurnberg | … |
| A3 | Sanders | Reinhard | Schwabach | … |

The rows, named by different attributes like first name, last name, etc., are called the fields of the record type. The lines Al, A2, etc., each constitute a record. Based on this fixed record structure, the standard operations for the retrieval and update of information in the database are defined.

To retrieve the name of the representative in, for example, *Schwabach*, the user must type in the following commands of the programming language (here, a query language for databases) without mistake.

*Database query:*

Query:

*select A#*

*where city ='Schwabach'*

Result:

*result: A3 Sanders Reinhard*

The correct use of commands such as 'select' initiates quasi-mechanical procedures which correspond to filing and retrieving cards in a filing cabinet with many compartments. Compared to the nonelectronic method, the computational system has many practical advantages. The electronic version is faster, the adding and removing of information is simpler, and the possibilities of search are much more powerful because various different keywords may be logically combined into a complex query. Is it possible to gradually extend such an interaction with a computer to natural language?

Standard computers have been regarded as general purpose machines for information processing because any kind of standard program can be developed and installed on them. From this point of view, their capabilities are restricted only by hardware factors like available speed and memory. In another sense, the information processing of standard computers is not general purpose, however, because their input and output facilities are restricted to the language channel.

A second type of computer not subject to this limitation is autonomous robots. In contradistinction to standard computers, robots are not restricted to the language channel, but designed to recognize their environment and to act in it.

Corresponding to the different technologies of standard computers and robots, there have evolved two different branches of artificial intelligence. One branch, dubbed classic Al by its opponents, is based on standard computers. The other branch, which calls itself nouvelle Al, requires the technology of robots.

Classic Al analyzes intelligent behavior in terms of manipulating abstract symbols. A typical example is a chess-playing program. It operates in isolation from the rest of the world, using a fixed set of predefined pieces and a predefined board. The search space for a dynamic strategy of winning in chess is astronomical. Yet the technology of a standard computer is sufficient because the world of chess is closed.

Nouvelle Al aims at the development of autonomous agents. In contrast to systems which respond solely to a predefined set of user commands and behave otherwise in isolation, autonomous agents are designed to interact with their real

world environment. Because the environment is constantly changing in unpredictable ways they must continually keep track of it by means of sensors.

For this, nouvelle AI uses the strategy of task level decomposition. Rather than building and updating one giant global representation to serve as the basis of automatic reasoning, nouvelle AI systems aim at handling their tasks in terms of many interacting local procedures controlled by perception. Thereby low-level inferencing operates directly on the local perception data.

A third type of machine processing information - besides standard computers and robots - is systems of virtual reality (VR). While a robot analyzes its environment in order to influence it in certain ways (such as moving in it), a VR system aims at creating an artificial environment for the user. Thereby the VR system reacts to the movements of the user's hand, the direction of his/her gaze, etc., and utilizes them in order to create as realistic an environment as possible.

The different types of human-computer communication exemplified by standard computers, robots, and VR systems may be compared schematicallyin fig.7.1:

**standard computer** **autonomous robot** **virtual reality**

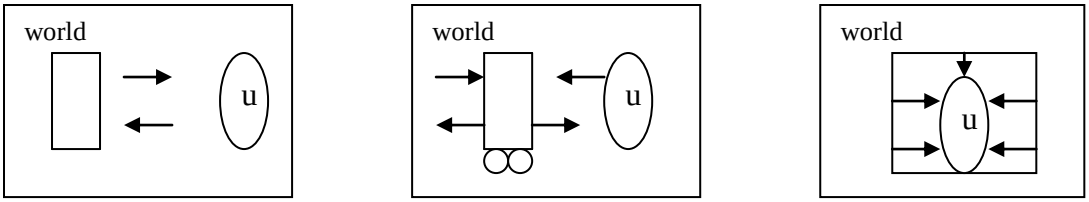



Figure 7.1. Three types of human-computer interaction:

The ovals represent the users who face the respective systems in the 'world.' The arrows represent the interaction of the systems with their environment and the user.

A standard computer communicates with users who initiate the interaction. A robot interacts independently with its environment and its users. A VR system does not interact with its environment, but rather creates an artificial environment for the user. In robots and VR systems, communication with the user in terms of language is optional and may be found only in advanced systems. These systems must always have a language-based 'service channel', however, for the installation and upgrading of the system software.

A speaker of English knows the meaning of a word like *red*. When asked to pick the red object among a set of non-red objects, for example, a competent speaker-hearer will be able to do it. A standard computer, on the other hand, does not 'understand' what red means, just as a piece of paper does not understand what is written on it.

In the interaction with a standard computer, the understanding of natural language is restricted largely to the user. For example, if a user searches in a database for a red object, (s)he understands the word red before it is put into - and after it is given out by - the standard computer. But inside the standard computer, the word red is manipulated as a sign which is uninterpreted with respect to the color denoted.

What is true for standard computers does not apply to human-computer communication in general, however. Consider for example a modern robot which is asked by its master to get an object it has not previously encountered, for example, the new blue and yellow book on the desk in the other room. If such a robot is able to spontaneously perform an open range of different jobs like this, it has an understanding of language which at some level may be regarded as functionally equivalent to the corresponding cognitive procedures in humans.

### *5.1. Linguistic aspects of Human-Machine Interaction*

The communication with a robot may be based on either artificial or natural language. The use of natural language is much more challenging, however, and much preferable in many situations. As a first step towards achieving unrestricted human-computer communication in natural language, let us consider the current state of linguistics.

In this field of research, three basic approaches to grammatical analysis may be distinguished, namely (1) traditional grammar, (2) theoretical linguistics, and (3) computational linguistics. They differ in their methods, goals, and applications.

*Traditional Grammar* uses the method of informal classification and description based on tradition and experience; it has the goal to collect and classify the

regularities and irregularities of the natural language in question as completely as possible, and is applied mostly in teaching languages (originally Latin).

While traditional grammar has long been shunted aside by theoretical linguistics, it has been of great interest to computational linguistics because of its wealth of concrete data.

*Theoretical Linguistics* uses the method of mathematical logic to describe natural languages by means of formal rule systems intended to derive all and only the well-formed expressions of a language - which has the advantage of stating grammatical hypotheses explicitly, has pursued the goal of describing the 'innate human language ability' (competence), whereby aspects of language use in communication (performance) have been excluded, and has had rather limited applications because of its computational inefficiency and because of its fragmentation into different schools.

Theoretical linguistics is relevant to computational linguistics in the area of formal language analysis and mathematical complexity theory.

*Computational Linguistics* combines the methods of traditional grammar and theoretical linguistics with the method of effectively verifying explicit hypotheses by implementing formal grammars as efficient computer programs and testing them automatically on realistic -i.e., very large - amounts of real data. It has as its goal modeling the mechanism of natural language communication, which requires a complete morphological, lexical, syntactic, semantic, and pragmatic analysis of a given natural language within a functional framework, and has applications in all instances of human-computer communication far beyond letter-based 'language processing.'

Computational linguistics analyses natural language at the level of abstraction which is independent of any particular medium of manifestation, e.g. sound.

Despite their different methods, goals, and applications, the three variants of language science divide the field into the same components of grammar, namely phonology, morphology, lexicon, syntax, semantics, and the additional field of

pragmatics. The role played by these components and the ways in which they are handled scientifically differs, however, within the three different approaches.

*Phonology: Science of language sounds.*

Phonology is the study of historical changes as well as synchronic alternations, such as trisyllabic laxing in English or final devoicing in German, in terms of generative grammars.

For theoretical linguistics, phonology is important: it is used as a kind of sand table on which different schools try to demonstrate the innateness of their current universals and grammar variants.

In computational linguistics, the role of phonology is marginal at best. One might conceive of using it in automatic speech recognition and synthesis, but the appropriate science is in fact phonetics. Phonetics investigates the (1) articulatory, (2) acoustic, and (3) auditive processes of speech. In contrast to phonology, phonetics is traditionally not considered part of the grammar.

*Morphology*: *Science of word form structure*

In the field of morphology the words of a language are classified according to their art of speech (category), and the structure of word forms is described in terms of inflection, derivation, and composition. To traditional grammar, morphology has long been central, as shown by the many paradigm tables in, for example, grammars of Latin.

In theoretical linguistics, morphology has played a minor role. Squeezed between phonology and syntax, morphology has been used mostly to exemplify principles of either or both of its neighboring components.

In computational linguistics, morphology appears in the context of automatic word form recognition. It is based on an on-line lexicon and a morphological parser which (1) relates each word form to its base form (lemmatization) and (2) characterizes its morpho-syntactic properties (categorization). Automatic word form

recognition is presupposed by all other rule-based techniques of automatic language analysis, such as syntactic and semantic parsing.

*Lexicon*: *Listing analyzed words.*

The words of a language are collected and classified in lexicography and lexicology. Lexicography deals with the principles of coding and structuring lexical entries, and is a practically oriented border area of natural language science. Lexicology investigates semantic relations in the vocabulary of a language and is part of traditional philology.

In computational linguistics, electronic lexica combine with morphological parsers in the task of automatic word form recognition. The goal is maximal completeness with fast access and low space requirements. In addition to building new lexica for the purpose of automatic word form recognition, there is a great interest in utilizing the knowledge of traditional lexica for automatic language processing ('mining of dictionaries').

*Syntax: Science of assembling word forms into sentences.*

In communication, the task of syntax is the composition of meanings via the composition of word forms (surfaces). One aspect of this is characterizing well-formed compositions in terms of grammatical rules. The other is to provide the basis for a simultaneous semantic interpretation.

In theoretical linguistics, syntactic analysis has concentrated on a description c grammatical well-formedness. The problem with analyzing well-formedness in isolation is that any finite set of sentences may be described by a vast multitude of different grammars. In order to select the one type of description which turns out to be correct in the long run, theoretical linguistics has vainly searched for 'universals' supposed to characterize the 'innate human language faculty.'

A more realistic and effective standard is to make the grammar suitable to serve as a component in an artificial cognitive agent communicating in natural language. Thereby, the descriptive and functional adequacy of the grammar may be tested

automatically on the full range of natural language data. This presupposes a grammatical algorithm with low mathematical complexity. Furthermore, the algorithm must be input-output equivalent with the mechanism of natural language communication.

*Semantics: The Science of literal meanings.*

The semantics of natural language may be divided into lexical semantics, describing the meaning of words, and compositional semantics, describing the composition of meanings in accordance with the syntax. The task of semantics is a systematic conversion of the syntactically analyzed expression into a semantic representation based on the function-argument structure underlying the categories of basic and complex expressions.

The beginning of traditional grammar contributed considerably to the theory of semantics, for example Aristotle's distinction between subject and predicate. However, these contributions were passed on and developed mostly within philosophy of language. In traditional grammar instruction, the treatment of semantics did not reach beyond the anecdotal.

In theoretical linguistics, semantics was initially limited to characterizing syntactic ambiguity and paraphrase. Subsequently, logical semantics was applied to natural language: based on a metalanguage, natural language meanings were defined in terms of truth conditions.

Computational linguistics uses procedural semantics instead of metalanguage-based logical semantics. The semantic primitives of procedural semantics are based on operations of perception and action by the cognitive agent. The semantics is designed to be used by the pragmatics in an explicit modeling of the information transfer between speaker and hearer.

*Pragmatics: The Science of using language expressions.*

Pragmatics is the study of how grammatically analyzed expressions are used relative to the context of interpretation. Therefore, pragmatics is not part of the

grammar proper, but concerned with the interaction between the expressions and the context, presupposing the grammatical analysis of the expressions and a suitable description of the context.

In traditional grammar, phenomena of pragmatics have been handled in the separate discipline of rhetoric. This has been an obstacle to integrating the analysis of language structure and language use.

In theoretical linguistics, the distinction between semantics and pragmatics has evolved only haltingly. Because theoretical linguistics has not been based on a functional model of communication, pragmatics has served mostly as the proverbial 'wastebasket' (Y. Bar-Hillel 1964).

The components phonology, morphology, lexicon, syntax, and semantics are part of the grammar proper because they deal with the structure of word forms, complex expressions, and sentences.

In computational linguistics, the need for a systematic theory of pragmatics became most obvious in natural language generation - as in dialogue systems or machine translation, where the system has to decide what to say and how to say it in a rhetorically acceptable way.

That the different approaches of traditional grammar, theoretical linguistics, and computational linguistics use the same set of components to describe the phenomena of natural language - despite their different methods and goals - is due to the fact that the division of phenomena underlying these components is based on different structural aspects, namely sounds (phonology), word forms (morphology), sentences (syntax), literal meanings (semantics), and their use in communication (pragmatics).

## 6. Cognitive Mechanisms of Human-Machine Communication

The application of linguistic knowledge to the creation of natural language understanding robots is the reason why the SLIM theory of language aims from the outset at modeling the mechanism of natural language communication in general. Thereby verbal and nonverbal contents are represented alike as concatenated

propositions, defined as sets of bidirectional proplets in a classic network database. This new format is not only suitable for modeling production and interpretation in natural human-computer communication, but also as a universal Interlingua system.

The mechanism of natural communication is described in terms of constructing a robot named CURIOUS.

### *6.1. Prototype of communication*

The question of how natural language functions in communication may seem complicated because there are so many different ways of using language. Consider the following samples of communication:

- two speakers are located face to face and talk about concrete objects in their immediate environment;
- two speakers talk on the phone about events experienced together in the past;
- a merchant writes to a company to order merchandise in a certain number, size, color, etc., and the company responds by filling the order;
- a newspaper article reports a planned extension of public transportation;
- a translator reconstructs an English short story in German;
- a teacher of physics explains the law of gravitation;
- a registrar issues a marriage license;
- a judge announces a sentence;
- a conductor says: *End of the line, everybody please get off!*;
- a sign reads: *Don't walk on the grass!*;
- a professor of literature interprets an expressionistic poem;
- an author writes a science fiction story;
- an actor speaks a role.

These different variants are not an insurmountable obstacle to designing a general model of communication. They only require finding a basic mechanism which works for all of them while accommodating their respective differences.

The slim theory of language proceeds on the hypothesis that there is a basic prototype which includes all essential aspects of natural communication. This

prototype is defined as follows: *the basic prototype of natural communication* is the direct face-to-face discourse of two partners talking about concrete objects in their immediate environment.

Possible alternatives to the basic prototype of natural communication would be approaches which take as their basic model, for example, complete texts or the signs of nature, such as smoke indicating fire.

The prototype hypothesis is proven in two steps. First, a robot is described which allows unrestricted natural human-computer communication within the basic prototype. Second, it is shown that all the other variants mentioned above are special cases or extensions which can easily be integrated into the cognitive structure of the robot.

Realizing the prototype of communication as a functioning robot requires an exact definition of the following *components of basic communication*:

1. Specification of the task environment;
2. Structure of the cognitive agent;
3. Specification of the language.

The notion *task environment* was introduced by A. Newell & H. Simon in 1972. It refers to the robot's external situation. The robot-internal representation of the task environment is called *the problem space.*

The task environment of the robot CURIOUS is a large room with a flat floor. Distributed randomly on the floor are objects of the following kinds:

- *triangles (scalene, isosceles, etc.);*
- *quadrangles (square, rectilinear, etc.);*
- *circles and ellipses.*

These objects of varying sizes and different colors are elements of the real world.

### *6.2. CURIOUS project for man-machine communication*

The robot is called CURIOUS because it is programmed to constantly observe the state of its task environment. The task environment keeps changing in

unforeseeable ways because the human 'wardens' remove objects on the floor, add others, or change their position in order to test CURIOUS ' attention.

CURIOUS knows about the state of its task environment by exploring it regularly. To avoid disturbing the objects on the floor, CURIOUS is mounted on the ceiling. The floor is divided into even-sized fields which CURIOUS can visit from above.

The basic cognition of CURIOUS includes an internal map divided into fields corresponding to those on the floor and a procedure indicating its current external position on the internal map. Furthermore, CURIOUS can specify a certain goal on its internal map and then adjust its external position accordingly.

When CURIOUS finds an object while visiting a certain field, the object is analyzed and the information is stored as, for example, *Isosceles red triangle in field D2*. By systematically collecting data of this kind for all fields, CURIOUS is as well-informed about its task environment as its human wardens.

### *6.3. Perception and recognition*

The first crucial aspect of this setup is that the task environment of CURIOUS is an open world: the objects in the task environment are not restricted to a fixed, predefined set, but can be processed by the system even if some disappear and new ones are added in unpredictable ways.

The second crucial aspect is that the task environment is part of the real world. Thus, for the proper functioning of CURIOUS a nontrivial form of reference must be implemented, allowing the system to keep track of external objects.

The cognitive functioning of CURIOUS presupposes the real external world as a given. This is in accordance with the approach of nouvelle AI, which is based on the motto *The world is its own best model*. CURIOUS' internal representations do not attempt to model the external world completely, but are limited to properties necessary for the intended interaction with the external world, here the perception and recognition of two-dimensional geometric objects of varying colors.

The performance of the system is evaluated according to the following two criteria:

1. Behaviour test, measuring active and reactive behaviour
2. Cognition test: measuring cognitive processing directly.

The behaviour test is the conventional method of observing actions and reactions of cognitive agents in controlled environments. If only behavior tests are available -as is normally the case with natural cognitive agents - the examination of cognitive functions is limited.

Behaviour tests with humans may include the use of language by interviewing the subjects about their experience. This however, (1) introduces a subjective element and (2) is not possible with all types of cognitive agents.

The cognition test consists in evaluating the cognitive performance of a system directly. This kind of test presupposes that the internal states can be accessed and accurately interpreted from the outside.

While we can never be sure whether our human partners see the world as we do and understand us the way we mean it, this can be determined precisely in the case of CURIOUS because its cognition may be accessed directly. Thus, the problem of solipsism may be overcome in CURIOUS.

A cognitive agent interacts with the world in terms of recognition and action. Recognition is the process of transporting structures of the external world into the cognitive agent. Action is the process of transporting structures originating inside the cognitive agent into the world.

The processes of recognition and action may be described at different levels of abstraction. Modeling vision, for example, is complicated by such problems as separating objects from the background, completing occluded portions, perception of depth, handling reflection, changes in lighting, perception of motion, etc.

For purposes of grounding a semantics procedurally, however, a relatively high level of abstraction is appropriate. As an illustration consider a robot without language observing its environment by means of a video camera. The robot's

recognition begins with an unanalyzed internal image of the object in question, e.g., a blue square, as presented in fig.7.2. (borrowed from R. Hausser 2002).

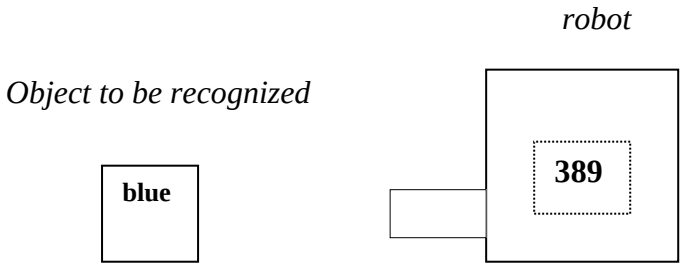
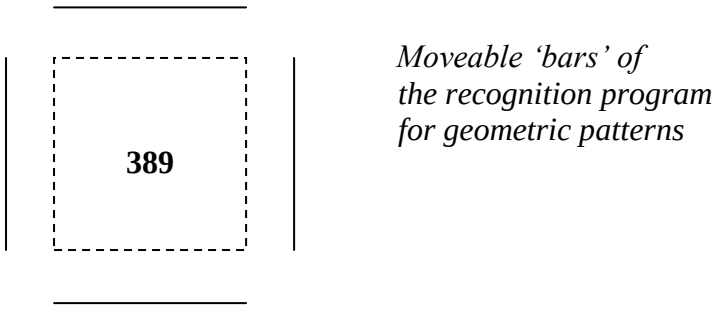



Figure 7.2. Schema of internal bitmap representation of external object

Inside the robot, the blue square is represented as a bitmap outline and its color as the electromagnetic frequency measured, i.e., 389 nm. Just as an OCR system analyzes bitmap structures to recognize letters, the robot recognizes the form of objects in its task environment by matching their bitmap structures with corresponding patterns.

The recognition of geometric forms may be viewed as a three step process. First, a suitable program approximates the bitmap outline with movable bars resulting in a reconstructed pattern:




Second, the reconstructed pattern is logically analyzed in terms of the number of corners, their angles, the length of the edges, etc. In our example, the logical analysis results in an area enclosed by four lines of equal length forming four right angles.

Third, the logical analysis is classified in terms of abstract concepts, to be discussed in the following section. The classification results in the recognition of the object in question. For the sake of conceptual simplicity, the reconstructed pattern, the logical analysis, and the classification are described here as separate phases. In practice, these three aspects may be closely interrelated in an incremental procedure. For example, the analysis system may measure an angle as soon as two edges intersect, the counter for corners may be incremented each time a new comer is found, a hypothesis regarding a possible matching concept may be formed early so that the remainder of the logical analysis is used to verify this hypothesis, etc. The basic recognition procedure illustrated above with the example square may be extended to other types of geometric objects, to properties like colors, and to relations like *A* being contained in *B*.

A system like CURIOUS is anchored in its task environment by means of its perception and recognition. This means that the relevant aspects of its task environment are represented internally and are updated constantly. The current internal representation is called the (nonverbal) context of a cognitive agent.

The context of a cognitive agent (CA) at a given point of time *t* includes:

1. the total of all current cognitive parameter values ($CA_t$);
2. the logical analyses of the parameter values and their combinations (reconstructed patterns);
3. the conceptual structures used to classify the reconstructed patterns and their combinations.

The cognitive processing of CURIOUS described so far illustrates the difference between perception and recognition. The raw data provided by the video camera are what CURIOUS perceives. Their classification with respect to geometric shape and colour constitute what CURIOUS recognizes.

## Conclusion

Transmitting information by means of a natural language like Chinese, English, or German is a real and well-structured procedure. This becomes evident when we attempt to communicate with people who speak a foreign language. Even if the information we want to convey is completely clear to us, we will not be understood by our hearers if we fail to use their language adequately.

Modeling the mechanism of natural communication in terms of a computationally efficient, general theory has a threefold motivation in computational linguistics. First, theoretically it requires discovering how natural language actually works - surely an important problem of general interest. Second, methodologically it provides a unified, functional viewpoint for developing the components of grammar on the computer and allows objective verification of the theoretical model in terms of its implementation. Third, practically it serves as the basis for solid solutions in advanced applications.

The development of speaking robots is not a matter of fiction, but a real scientific task. Remarkably, however, theories of language have so far avoided a functional modeling of the natural communication mechanism, concentrating instead on peripheral aspects such as methodology (behaviorism), innate ideas (nativism), and scientific truth (model theory).

## BIBLIOGRAPHY


1. Андрійчук Н.І. Прикладна лінгвістика: концепція підготовки фахівців та перспективи розвитку спеціальності у державному університеті „Львівська політехніка” // Матеріали 3-ї Міжнародної наукової конференції „Комп’ютерна лінгвістика та викладання чужоземних мов у вищих навчальних закладах”. – Львів. – 1998. – С. 5-6
2. Баранов А.Н. Введение в прикладную лингвистику: М.: Эдиториал УРСС, 2001. – 358 с.
3. Волошин В.Г. Комп’ютерна лінгвістика. – Суми: Університетська книга, 2004. – 468 с.
4. Звегінцев В.А. Теоретическая и прикладная лингвистика. – М.: Просвещение, 1968. – 324 с.
5. Златоустова Л.В., Потапова Р.К. Общая и прикладная фонетика. – М.: изд-во Моск. ун-та, 1997. – 415 с.
6. Использование ЕВМ в лингвистических исследованиях. – Киев: Наукова думка, 1990. – 226 с.
7. Кейтер Дж. Комп’ютеры – синтезаторы речи. – М.: Мир, 1985. – 236 с.
8. Морфологический анализ научного текста на ЭВМ. – Киев: Наука думка, 1989. – 264 с.
9. Пещак М.М. Нариси з комп’ютерної лінгвістики. – Ужгород: Закарпаття, 1999. – 200 с.
10. Потапова Р.К. Введение в лингвокибернетику. – М.: Изд-во Московского Ордена Дружбы Народов государственный лингвистический университет, 1990. – 140 с.

11. Чейф У.Л. Память й вербализация прошлого опыта // Новое в зарубежной лингвистике. - в. 12. - С. 35-40.
12. Agger B. The Decline of Discourse: Reading, writing, and resistance in postmodern capitalism. – New York: Falmer Press, 1990
13. Armstrong S. Using Large Corpora. – Cambridge, MA: MIT Press, 1994
14. Asher N., Lascarides A. Intentions and information in Discourse // Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. – San Francisco: Morgan Kaufmann Publishers, 1994. – P.34-41
15. Atkins B.T.S.: Tools for Computer-Aided Lexicography: The Hector Project // Papers in Computational Lexicography: Complex 1994. – Budapest; Research Institute for Linguistics, Hungarian Academy of Sciences. – 1994. – P. 1-59
16. Atkins B.T.C., Zampolli A. Computational Approaches to the Lexicon. – Oxford: Oxford U. P., 1994
17. Atkins B.T.S., Levin B. Building on a Corpus // International Journal of Lexicography. - № 8. – 1995. – P. 85-114
18. Bar-Hillel Y. Language and Information. Selected Essays on Their Theory and Application. – Mass., USA: Addison-Wesley, 1964
19. Bartoli M.G. Saggi di linquistica spaziale. – Torino: Bona, 1945
20. Bates M, Boisen S., Makhoul J. Developing an Evaluation Methodology for Spoken Language Systems // DARPA Speech and Natural Language Workshop. – Hidden Valley, PA: Morgan Kaufmann Publishers, 1991. – P.102-108
21. Bates M., Weischedel R.M. Challenges in Natural Language Processing. – Cambridge: Cambridge U. P., 1993
22. Belevitch V. Langage des machines et langage humain. – Brussels: Labegue & Nationlale, 1956
23. Bernhardt S.A. The Shape of Text to Come: The texture of print on screens // College Composition and Communication. – № 44. – 1993. – P.141-175
24. Biber D., Conrad C., Reppen R. Corpus-based Approaches to Issues in Applied Linguistics // Applied Linguistics. – № 15. – 1994. – P. 169-189

25. Bloch B., Trager G.L., Outline of Linguistic Analysis. – Baltimore: Linguistic Society of America, 1942
26. Bloomfield L. Language. – New York: Holt, 1933
27. Bolter J. Writing Space: The computer, hypertext, and the history of writing. – Hillsdale, NJ: Lawrence Erlbaum, 1991
28. Bolter J. Turing's Man: Western culture in the computer age. – Chapel Hill, NC: University of North Carolina Press, 1984
29. Brandt D. Literacy as Involvment: The acts of writers, readers, and texts. – Carbondale, IL: Southern Illinois U. P., 1990
30. Brent M.R. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax // Computational Linguistics. – № 19. – 1993. –P. 243-262
31. Brill E. A simple Rule-Based Part-of-Speech Tagger // Proceedings of the third Conference on Applied Natural Language Processing. – Trento, Italy: Morgan Kaufmenn Publishers, 1992. – P. 152-155
32. Britton K. Communication. – New York: Harcourt, 1939
33. Brøndal V. Essais de linguistique générale. – Copenhagen: Munksgaard, 1943
34. Brunot F. La pensée et la langue. – Paris: Masson, 1926
35. Buchanan R.A. Texbase Technology: Writing with reusable text // Computers and Writing: State of the Art. Ed. by P.O.Holt and N.Williams. – Oxford: Intellect, 1992. – P. 254-265
36. Burke S. The Death and Return of the Author. – Edinburgh: Edinburgh U. P., 1992
37. Bush V. As We May Think // Atlantic Monthly. – № 176 (July). – 1945. – P. 101-108
38. Carnap R. Logishe Syntax der Sprache. – Vienna: Springer, 1942
39. Cassirer E. Philosophie der symbolishen Formen. – vol.1. – Teil: Die Sprache. – Berlin: Cassirer, 1923
40. Chierchia G., McConnell-Ginet S. Meaning and Grammar: An Introduction to Semantics. – Cambridge, MA: MIT Press, 1990
41. Chomsky N. Syntactic Structures. – The Hague: Mouton, 1957

42. Clear J. The British National Corpus // The Digital Word: Text-based computing in the humanities. – London: MIT Press, 1993. – P. 163-188
43. Crother W. A Common facts Data Base // Speech and Natural language. – San Mateo, CA: Margan Kaufmann Publishers Inc., 1980. – P.89-93
44. Cruse D.A. Lexical Semantics. – Cambridge: CUP, 1986
45. Damourette J., Pichon E. Des mots á la pensée. Essai de grammaire de la langue française. – 7 vols. – Paris: d'Artrey, 1952
46. Dermatas D.E., Kokkinakis G. Automatic Stochastic Tagging of Natural language Texts // Computational linguistics. – № 21. – 1995. – P. 137-163
47. Diaper D., Sanger C. CSCW in Practice. – New York: Springer-Verlad, 1993
48. Eastman C.M., McLean D.C. On the Need for Parsing Ill-Formed Input // American Journal of Computational linguistics. – № 7. – 1981. – P. 257-271
49. Eisenstein E. The Printing Press as an Agent of Change: Communication and Cultural Transformations in Early Modern Europe. – Cambridge: Cambridge U. P., 1979
50. Ferreiro E., Teberosky A. Literacy before schooling (English translation). – Mexico D.F.: Siglo Veintiuno Editors, 1982
51. Firth J.R. Papers in Linguistics. – London: Oxford U.P., 1957
52. Flower L.S., Hayes J.R. The Dynamics of Composing: Making plans and juggling constraints // Cognitive Processing in Writing. Ed. by L.W. Gregg, E.Steinberg. – Hillsdale, NY: Lawrence Erlbaum, 1980. – P.31-50
53. Flowerdew L., Tong A. Entering Text. – Hong Kong: The Language Centre, 1994
54. Fouché P. Traité de pronunciation française. – Paris: Klincksieck, 1956
55. Frei H. La grammaire des fautes. Introduction á la linguistique functionnelle. – Paris: Geuthner, 1929
56. Frei H. Le livre des deux mille phrases. – Geneva: Droz, 1953
57. Fries C.C. The Structure of English. – London: Longmans, 1957
58. Fries U., Tottie G., Schneider P. Creating and Using English Corpora. – Amsterdam: Rodopi, 1994

59. Fromkin V.A. Janua Linguarum, Series Maior 77: Speech Errors as Linguistic Evidence. – The Hague: Mouton, 1973
60. Gardiner A.H. The Theory of Speech and Language. – London: Oxford U.P., 1932
61. Gilliéron J., Edmont E. Atlas linguistique de la France. – Paris: Champion, 1912
62. Goldstein K. Language and language Disturbances. – New York: Grune, 1948
63. Gouldner A.W. The Dialectic of Ideology and Technology. – New York: Oxford U. P., 1976
64. Greenberg J.H. A Quantitative approach to the Morphological Typology of Language // Method and Perspective in Anthropology // Minneapolis: Minnesota U. P., 1954
65. Grosz B.J., Joshi A.K., Weinstein S. Centering: A Framework for Modeling the Local Coherence of Discourse // Computational linguistics. – № 21. – 1995. – P.203-225
66. Guillaume G. Langage et science du langage. – Paris: Nizet, 1963
67. Guiraud P. Langage, connaissance et information // J. de Psychologie, - 1958
68. de Haan P. Postmodifying Clauses in the English Noun Phrase: A corpus-based study. – Amsterdam: Rodopi, 1989
69. Hajicova E., Skoumalova H. and Segall P. An Automatic Procedure for Topic Focus identification // Computational linguistics – № 21. – 1995. – P. 81-94
70. Halliday M.A.K. Categories of the Theory of Grammar // Word. – Vol. 17. – 1959
71. Harris R. The Origin of Writing. – London: Duckworth, 1986
72. Harris Z.S. Methods in Structural Linguistics. – Chicago: Chicago University Press, 1951
73. Hausser R. Foundations of Computational Linguistics. – Berlin, NY: Springer, 2002
74. Heim M. Electric Language: A philosophical study of word processing. – New Haven, CT: Yale U. P., 1987

75. Herdan G. The Calcilis of Linguistic Observations. – The Hague: Mouton, 1962
76. Herdan G. Type-Token Mathematics: a textbook of mathematical linguistics. – The Hague: Mouton, 1960
77. Hinrichs E.W., Ayuso D.M., Scha R. The Syntax and Semantics Of the JANUS Semantic Interpretation Language // Research and Development in Natural Language Understanding as Part of the Strategic Computing Program. Annual Technical Report Dec. 1985 – Dec. 1986. – BBN Laboratories. – № 6522. – P. 27-31
78. Hjelmslev L., Uldall H.J. Outline of Glossematics. A study in the methodology of the humanities with special reference to linguistics. – Copenhagen: Nordisk Sprog og Kulturforlag, 1957
79. Hirsch E.D. Cultural Literacy: What every American needs to know. – Boston: Houghton Miffin, 1987
80. Holmes J., Holmes W. Speech Synthesis and Recognition. – London, New York:Taylor & Francis, 2002
81. Jakobson R., Halle M. Fundamentals of Language. – The Hague: Mouton, 1956
82. Jespersen O. The Philosophy of Grammar. – London: Allen & Unwin, 1924
83. Joos M. (ed.) Readings in linguistics. The development of descriptive linguistics in America since 1925. – Washington: American Council of Learned Societies, 1957
84. Kaiser L. Manual of phonetics. – Amsterdam: North Holland, 1957
85. Kuznecov P.S. Die Morphologische Klassifikation der Sprachen. – Halle: Niemeyer, 1956
86. Ladefoged P. Elements of Acoustic Phonetics. – Edinburgh: Oliver & Boyd, 1962
87. Landow G.P. Hypertext: The convergence of contemporary critical theory and technology. – Baltimore: Johns Hopkins University Press, 1992

88. Landow G.P., Delany P. The Digital Word. – Cambridge, MA: MIT Press, 1993
89. Langacker R.W. An Overview of Cognitive Grammar // Topics in Cognitive Linguistics. – Philadelphia: John Benjamins, 1988. – P.3-48
90. Lankshear C., McLaren P.L. Critical Literacy: Politics, praxis and the postmodern. – Albany, NJ: SUNY Press, 1993
91. Larsen M.T. Literacy and Society. – Copenhagen: Centre for Research in the humanities, 1989
92. Locke W.N., Boot A.D. Machine Translation of Languages. – New York: Wiley, 1955
93. Mackey W.F. Language Teaching Analysis. – London: Longmans, 1983
94. Mair C. Infinitival Complement Clauses in English. – Cambridge: Cambridge U. P., 1990
95. Manning C.D. Automatic Acquisition of a Large Subcategorization Dictionary From Corpora // Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. – San Francisco: Morgan Kaufmann Publishers, 1993. – P. 235-242
96. McMillan B. Current Trends in Information Theory. – Pittsburgh: Pittsburgh U.P., 1953
97. McRoy S.W. Using Multiple Knowledge Sources for Word Sense Disambiguation // Computational Linguistics. – № 18. – 1992. – P. 1-30
98. Meredith G.P. The communication of Scientific Concepts and models of semantic Mechanisms // The advancement of Science. – №18. – P.110-117
99. Merialdo B. Tagging English Text with a Probabilistic Model // Computational Linguistics. – № 20. – 1994. – P. 155-171
100. Miller G.A. Language and communication. – New York: McGraw-Hill, 1951
101. Montague R. Pragmatics and Intensional Logic // Syntheses. – № 22. – 1970. – P. 68-94

102. Montague R. Formal Philosophy: Selected Papers of Richard Mantague. – New Haven,CT: Yale University Press, 1974

103. Morais J., Alegria J., Cary L., Bertelson P. Does awareness of speech as a sequence of phones arise spontaneously? // Cognition. – № 7. – 1979. – P. 323-331

104. Morris C. Signs, Language and Behaviour. – New York: Prentice-Hall, 1946

105. Morris C.W. Foundations of the Theory of Signs. – Chicago: Chicago U.P., 1938

106. Neff M., Boguraev B. Dictionaries, Dictionary Grammars and Dictionary Entry Parsing // Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. – Van Couver, British Columbia, 1989. – P. 91-101

107. Nelson T. Literacy Online: The promise (and peril) of reading and writing with computers. – Pittsburgh PA: University of Pittsburgh Press, 1992

108. Neuwirth C.M., Kaufer D.S., Chimera R., Gillespie T. The Notes Program: A hypertext application for writing from source texts. – Pittsburgh, PA: Carnegie-Mellon University, 1987

109. von Newman J. The Computer and the Brain. – London: Oxford U. P., 1958

110. Nystrand M. A Social-Interactive Model of Writing // Written Communication. – № 6. – 1989. – P. 66-85

111. Oettinger A.G., Automatic Language Translation: lexical and technical aspects. – Cambridge (Mass.): Harvard U. P., 1960

112. Ogden C.K., Richards I.A. The Meaning of Meaning. A study of the influence of language upon thought and of the science of symbolism. – London: Routledge & Kegan Paul, 1949

113. Ong W. Ramus: Method and the Decay of Dialogue: From the art of Discourse to the art of Reason. – Cambridge: Cambridge U. P., 1958

114. Ooi V.B.Y. Computer Corpus Lexicography. – Edinburgh: Edinburgh University Press, 1998
115. Oostdijk N., de Haan P. Corpus-based Research into Language. – Amsterdam: Rodopi, 1994
116. Penfield W.C., Roberts L. Speech and Brain Mechanisms. – Princeton: Princeton U. P., 1959
117. Peterson G. An Oral Communication // Language. – № 31. – P. 414-427
118. Pike K.L. Language in Relation to a Unified Theory of the Structure of Human Behaviour. – 3 vols. – Santa Ana (California, USA): Summer Institute of Linguistics, 1960
119. Quirk R., Greenbaum S., Leech G., Svartvik J. A comprehensive Grammar of the English Language. – London: Longman, 1985
120. Reppen R. Variation in Elementary Student Language: A multi-dimensional perspective. – Flagstaff, AZ: Northern Arizona University. – Ph.D. diss. - 1994
121. Ries J. Was ist ein Satz? Beiträge zur Grundlegung der Syntax. – Prague: Taussig & Taussig, 1931
122. Rich C., Waters R.C., Automatic Programming: Myths and Prospects. – NY: MIT Press, 1988
123. Ross D., Hunter D. m-Eyeball: An interactive system for producing stylistic descriptions and comparisons // Computers in the Humanities. – № 28. – 1994. – P. 1-11
124. Russel B. An Inquiry into Meaning and Truth. – New York: Norton, 1940
125. Saussure F. de Course in General Linguistics. – New York: Philosophical Library, 1959
126. Shannon C., Weaver W. The Mathematical Theory of Communication. – Urbana: Illinois U. P., 1949
127. Sharples M. Computer Supported Collaborative Writing. – New York: Springer-Verlad, 1993

128. Shea V. Netiquette. – San Francisko, CA: Albion Books, 1994
129. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford U. P., 1991
130. Skinner B.F. Verbal Behaviour. – New York: Appleton, 1957
131. Spang-Hanssen H. Probability and Structural classification in Language Description. – Copenhagen: Rosenkilde & Bagger, 1959
132. Spring M.b. Electronic Printing and Publishing: The document processing revolution. – New York: Marcel Dekker Inc., 1991
133. Stallard D. Answering Questions Posed in an Intensional Logic: A Multilevel Simantics Approach // Weischedel R., Ayuso D. etc. Research and Development in Natural Language Understanding as Part of Strategic computer Program. – Cambridge: BBN Laboratories, 1987. – P. 35-47
134. Summers D. Longman / Lancaster English language corpus – criteria and design. // International Journal of Lexicography. – № 6. – 1993. – P. 181-208
135. Svartvik J. The London-Lund Corpus of Spoken English: Description and Research. – Lund, Sweden: Lund U. P., 1990
136. Thompson B.H. Linguistic Analysis of Natural language Communication with Computers // Proceedings of the eighth International Conference on Computational linguistics. – 1980. – P. 190-201
137. Togeby K. Structure immanente de la langue française. – Copenhagen: Nordisk Sprog- og Kulturforlag, 1951
138. Trager G.L. The Field of Linguistics: SIL Occasional Paper I. – Norman (Okla): Battenburg, 1950
139. Trubetzkoy J.M. Principes de Phonologie. – Paris: Klincksieck, 1949
140. Turner J. A Theory of Social Interaction. – Palo Alto, CA: Stanford U. P., 1988
141. Urban W.M. Language and Reality. – London: Allen & Unwin, 1939
142. Virbel J. Reading and Managing Texts on the Bibliothéqie de France station // The Digital Word. Ed. by G.Landow and P.Delany – Cambridge, MA: MIT Press, 1993

143. Watts W. The Brave New World of Desktop Publishing // Computers in the Humanities. – № 26. – 1992. – P. 457-461
144. Wayner P. Boxweb: A structured outline program for writers // Computers and Writing: State of the art. Ed. by P.O.Holt, N.Williams. – Oxford: Intellect, 1992 . – P. 78-89
145. Weischedel R.M. A Hybrid Approach to Representation in the Janus Natural language Processor // Proceedings of the 27the Annual Meeting of the Association for Computational Linguistics. – Cambridge: ACL, 1989. – P. 193-202
146. Weizenbaum J. Elisa – a computer program for the study of natural language communication between man and machine // Communications of the ACM. – № 9.1. – 1966. – P. 36-45
147. Wells H.G. World Brain. – Garden City, NY: Doubleday, 1938
148. Whatmough J. Language: a modern synthesis. – New York: St. Martin's, 1956
149. Whitehead A.N., Russel B. Principia Mathematica. – 3 vol. – Cambridge: CUP, 1925-27
150. Whorf B.L. Language, Thought and Reality. – New York: Wiley, 1956
151. Wiebe J.M. Tracking Point of View In Narrative // Computational Linguistics. – № 20. – 1994. – P. 233-287
152. Wiener N. Cybernetics. New York: Wiley. 1948
153. Williams N. Writer's Problems and computer solutions // Computer Assisted Language Learning Journal. – № 2. – 1990. – P. 5-25
154. Wittgenstein L. Philosophical Investigations. – Oxford: Blackwell, 1960
155. Woods W.A. Transition Network Grammars for Natural Language Analysis // Communication of the association for Computing Machinery. – № 13. – 1970. – P. 591-606
156. Zipf G.K. Human Behaviour and the principle of Least Effort: An Introduction to Human Ecology. – Cambridge: Addison-Wesley, 1949

157. Zuboff S. In the Age of Smart Machine: The future of work and power. – New York: Basic, 1988