Verbmobil

CSLI
Lecture Notes
No. 33

Verbmobil

A Translation System for Face-to-Face Dialog

Martin Kay, Jean Mark Gawron, and Peter Norvig

CSLI CENTER FOR THE STUDY OF LANGUAGE AND INFORMATION

CSLI was founded early in 1983 by researchers from Stanford University, SRI International, and Xerox PARC to further research and development of integrated theories of language, information, and computation. CSLI headquarters and the publication offices are located at the Stanford site.

CSLI/SRI International	CSLI/Stanford	CSLI/Xerox PARC
333 Ravenswood Avenue	Ventura Hall	3333 Coyote Hill Road
Menlo Park, CA 94025	Stanford, CA 94305	Palo Alto, CA 94304

Printed in the United States
01 00 99 98 97 96 95 94 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Kay, Martin.

Verbmobil: a translation system for face-to-face dialog / Martin Kay, Jean Mark Gawron, and Peter Norvig.

p. cm. – (CSLI lecture notes ; no. 33)

Includes bibliographical references and index.

ISBN 0-937073-96-2 (cloth) — ISBN 0-937073-95-4 (paper)

1. Machine translating. 2. Automatic speech recognition. 3. Natural language processing. I. Gawron, Jean Mark. II. Norvig, Peter. III. Title. IV. Series.

P309.K39 1994
418'.02'0285--dc20 93-38054
CIP

Contents

Preface

The Verbmobil Project

Verbmobil is a portable simultaneous interpretation machine. Carry it to a meeting with speakers of anther language and it will translate what you say for them. Their Verbmobils, if they have them, will enable you to understand what they are saying.

Though it does not exist today, Verbmobil is more than a figment of the imagination. It is, in fact the eventual goal of a program of research recently undertaken by the German Bundesministerium für Forschung und Technologie (BMFT), (the Federal Ministry of Research and Technology).

Verbmobil is obviously an extremely ambitious program. It depends on finding solutions to a great number of problems that have resisted the most determined attacks for decades. For example, Verbmobil must be able to pick words and sentences out of the stream of sounds that impinge on its microphone. But the sounds run together and are confused by background noise. Nothing in the stream corresponds to the spaces that separate written letters and words, or the punctuation marks that set off major phrases in written language. Furthermore, people are often less careful in speech than in writing and, whereas corrections can be erased from the page, they cannot be removed from the follow of sound. Much remains to be explained about how language is used in everyday conversation, about the functions of intonation and rhythmic variation, about how one person yields the floor to another, and so forth. Above all, Verbmobil must be able to translate what it hears into another language and, while this problem has motivated research over the last thirty years, very much more must be learnt before even the most elementary prototype can be built.

This book resulted from a study of Verbmobil conducted at the Center for the Study of Language and Information, Stanford Univer-

sity, at the request of BMFT. It was not a feasibility study because the question never was whether Verbmobil was an appropriate goal to pursue. Taking the goal as given, we were asked to assess the state of the art in relevant fields of science and technology and to chart a course to the first prototype. However, while the overall goal was given, just what should be allowed to count as a prototype was not, and we therefore took this as part of our charter.

The authors wish to thank the following people for their contributions to this report: Jared Bernstein, Jason Christopher, John Etchemendy, Hana Filip, Jerry Hobbs, Michael Inman, Megumi Kameyama, Dikran Karagueuzian, Betsy Macken, Hy Murveit, Ryo Ochitani, Stanley Peters, William Poser, Ivan Sag, Dekai Wu.

1

Introduction

Speech recognition has been a source of fascination for at least forty years (see, for example Denes and Mathews 1960, Davis et al. 1952, and Dudley and Balashek 1958). It is, as it were, the canonical problem for those interested in the relationship between the continuous realm of sound and the discrete realm of language and communication. Even scientifically astute laymen find it hard to understand what can be so difficult about it, especially in view of how little effort, training, or intelligence it seems to require of humans. Yet, despite fairly determined efforts over those forty years, a machine that could automatically transcribe dictation remains little more than a dream.

The history of machine translation has been similar. It has been an object of research, at times quite intense, certainly since the advent of the earliest computers, and it remains a primary motivating problem for computational linguists. If a machine could be made to translate nontrivial texts from one ordinary language to another, it could only be because solutions had been found for a large proportion of the problems of language and linguistic computing. To many laymen, it is incomprehensible that we can build machinery that can convey a man to the moon, but none that can translate even very simple texts into French.

Much of this book will be devoted to the question of why the problems of speech recognition and machine translation are as difficult as they are and whether there is reason to foresee a brighter future for them. These were certainly questions of great interest to the German Bundesministerium für Forschung und Technologie (BMFT), when it initiated a project to build Verbmobil, a portable interpretation machine intended to translate utterances into the language of a foreign interlocutor. On the face of it, the project seems heroic, not to say foolhardy, especially in view of the fact that it will embrace the great majority of the research done in the relevant fields in Germany for

several years. The *prima facie* case for a pessimistic view of Verbmobil is indeed strong and much of what we shall have to say here will serve only to provide further support for this view. A similar project that has been under way for some years at the Advanced Telephony Research Laboratories in Japan has done little to support an opposing view.

However, we must assume that those in the BMFT who framed the idea, though possibly heroic, were not foolhardy. The project that gave rise to this book, and which was carried out at the behest and with the generous sponsorship of the BMFT, was intended to uncover what was already known about the relevant fields of endeavor and to estimate the severity of the difficulties that lay ahead. Recommendations were solicited on the most promising approaches to take, but not on whether the enterprise as a whole should go forward—we were given to understand that that decision had already been taken. On the other hand, recommendations were solicited on the question of what short-term goals it would be reasonable to adopt, the milestones that the project should be expected to pass, and how its results should be assessed. To this extent, then, we were given a fairly liberal charter to say what we thought Verbmobil should be. At the end of this chapter, we will summarize the recommendations that we made in response to that part of our charter. However, since this book is intended for a wider audience, a few words are in order in support of the project as a whole.

In the early seventies, the Advanced Projects Research Agency of the U. S. Department of Defense (ARPA) initiated a major program of research on speech recognition and, like the BMFT, they began by empanelling a committee of advisors to chart the course of the project (Newell 1973). Prominent among the findings of that committee was that the signal that reaches a person's ears probably contains less information than the person perceives in it, the rest being supplied by expectations based on context. Therefore, an approach to speech recognition that concentrated too narrowly on acoustic and other low-level processes would be unlikely to succeed. The project, as it developed, placed a great deal of emphasis on the integration of acoustic with other linguistic and pragmatic processes. The systems were built to understand what they heard only in the light of what they expected to hear. The report recommended that, while always deriving maximum benefit from information directly available from the speech signal, including rhythm and prosody, speech recognition should also profit to the greatest possible extent from constraints imposed by other modules in general, and by context in particular.

Despite some important discoveries, the ARPA speech project was not generally accounted a great success and the pendulum of fashion soon began to swing back towards systems that attended mainly to the acoustic signal. Now, the tendency is once again in the first direction. The major funding agencies, notably DARPA, are increasingly insistent that speech research must be conducted within a larger language-processing framework. Likewise, computational linguists are encouraged to work in a framework that includes speech processing.

There are several reasons for the current tendency towards integration. First is the intuition that language is a whole and that, while one must often divide in order to conquer, little good can come from treating language and speech as unrelated. It is part of linguistic dogma that languages are primarily spoken, and only derivatively written. Prosody carries important information and to ignore it is to make a difficult problem gratuitously harder.

Second, independently of any judgements one might make on the outcome of the ARPA speech project, the recommendation that it should be conducted in a larger linguistic framework was surely correct. That we supply deficiencies in what we hear by various kinds of inference from context is not controversial. Such errors of judgement as were made on this score in the course of the project surely concerned not the principle but the degree of emphasis to be placed here or there.

Third, there is a larger and more slow moving pendulum that is now swinging back to its position of some thirty years ago. In the fifties and early sixties, it was hoped that much of what was important for understanding language would prove to be based on emergent properties of text and that computers opened the possibility of discovering these properties by statistical techniques. The fact that few really interesting properties of language were in fact unearthed in this way was explained by some on the grounds the machines were not powerful enough and not enough textual material was available to apply them to. Others explained it by claiming that the key properties of langauge are not emergent properties of texts so that no amount of processing could be expected to being them to light. The members of the first camp now have more powerful machines and amounts of text are available in machine readable form that are comparable in size to what a child encounters by the time it is generally said to have acquired language. They believe that the use of such statistical techniques as Hidden Markov Models have been the secret of such success as has been achieved in speech recognition and so are encouraged in the view that similar approaches will also be successful in language processing as a whole.

In summary, the belief is rapidly gaining ground that speech and language are two sides of the same coin and little good can come from continuing to treat them as unrelated. However, this does not constitute an argument that the fortunes of each should be forced to ride on the success of practical applications of the other. To an extent, Verbmobil does just this and the danger that it represents for research in Germany should not be underestimated. On the other hand, Verbmobil is designed to disappear rapidly, or to be radically redesigned if it does not demonstrate significant progress in a very short time so that, if intelligently managed, its overall effect should be positive.

It will emerge from the discussions in later chapters that we see Verbmobil as starting with a far more serious handicap than the conjunction of speech recognition and translation constitutes. In our view, the present conception of Verbmobil has a flaw that has attended the great majority of translation projects. All indications are that Verbmobil is founded on the proposition that translation is an application of a technology whose underlying science is linguistics. In fact, every professional translator is keenly aware that a great deal more than linguistic knowledge is required for the job. If we tend to think of it as primarily linguistic, it is probably because linguistic capabilities are what most obviously set off translators from other people. That crucial knowledge that a translator must have, but that humans generally take for granted, is easily overlooked by the layman, and has almost always been overlooked even by professional researchers in the field. This is easy enough to understand because it is shared by most humans, expecially when they have a largely common culture.

Finally, it should be understood from the outset that Verbmobil is not to be viewed as a single research project, to be carried out in one laboratory, or even under the control of a single director with the power to shift resources and to enforce compatibility. It is a funding program that will invite proposals from university departments and industrial laboratories, and enter into contracts with those that are thought most promising. This is not the most efficient or most focussed paradigm for the conduct of research that one could imagine, but we understand that it is the only one that was open to the ministry.

In the report on which this book is based, a number of specific recommendations were made, which are set out in detail in 4. The following paragraphs summarize them briefly.

We recommended that there should, in fact, be two research prototypes, and possibly two demonstators. The first, which we refer to as "Product one," should provide for face-to-face conversation between a pair of participants, each speaking a different language, on

extremely limited subject matter, and in circumstances in which the conversational aims of the participants would be known in advance. Product two should provide for the interpretation of occasional words and phrases in a conversation between a pair of participants who communicate mainly in a common language. The topic would also be restricted, but perhaps not so narrowly as for Product one. Product two should provide translation on demand for participants who have a passive knowledge of a language of which neither is a fluent speaker, but in which most of their conversation will be conducted. In the course of an utterance, a participant can press a button to signal that he is now talking in his native language, and that what he says should be translated into the common language. It might also be possible for the listener to make use of Product two by pressing a button when a particular word or phrase is unclear. It was recommended that all work on these initial products should be done with an eye to the fully fledged eventual Verbmobil, so that the scientific foundation of the technology should never be compromised in the interests of achieving some functionality in the short run.

The recommendation that was most strongly urged concerned the subject matter of the conversations that the demonstrator and research prototypes for Verbmobil should be expected to support. This goes to the heart of the question of what should count as success or failure within the time frame that we were invited to consider. The principal reasons for limiting the subject matter are:

1. To limit the size of the dictionary,
2. To limit the number of things that can be referred to,
3. To encourage repetition of phenomena,
4. To facilitate collection of a corpus of examples.

In addition, it is important that the domain chosen should be such as to encourage cooperative, rather than adversarial interaction. To the extent possible, the domain should be such as to enable experimenters to manipulate the tasks given to subjects so as to control the likelihood of the conversations having certain properties that interest them. During the first year of the program, possible task domains should be studied and a suitable one chosen. Among the specific domains we discuss in 4.3.1, the most attractive to us were the *Contract Negotiation* and *Design Negotiation* tasks, because they reduce to a minimum the problem of giving Verbmobil access to the objects that the participants are talking about independently of the access provided by the conversation itself.

In the early stages of the program, Product one should be carefully

simulated using professional interpreters and a set of data collected that would then be usable by all groups involved in the research and development. The results should then be processed so as to form a data base containing the expected inputs and outputs at each of the major interfaces in the system. A preliminary version of the eventual systems, called "breadboards" should be built with dummy programs in place of the major modules. The dummy programs will produce "correct" outputs for examples in the experimental data by simply delivering what the data base says they should.

The main reason for collecting a body of data will be to facilitate simulations of some components of Verbmobil so as to provide a realistic environment for experimentations with others. The breadboard is a complete system, built early in the history of the project, which can simulate the performance of the final Verbmobil by drawing on the data in appropriate ways.

As well as the two principal products, experiments could profitably be made with a number of minor variants of Verbmobil for the strength that they would give to the main effort, for their intrinsic value, and for their value as insurance against unforeseen flaws in the original conception. Product two provides for interpretation among people who are able to carry out much of their conversation in a common language, but need help from time to time. Others involve providing the system with information about the conversation that would be hard to derive just by listening. There might, for example, be one based on a notion we refer to as the *electronic blackboard.* We imagine this as an icon-based graphical interface that conversants could use to clarify what is being referred to in a conversation. The idea is simply that a person should be able to create an icon on the screen when referring to something for the first time, and to point to that icon when referring back to the object. The machine could use this to clarify the intended reference.

Alternatively, suppose that, in an adjacent room, there was a group of people whose job it was to answer questions put to them by Verbmobil. In the degenerate case, they would usurp all the functions of the machine but, in more interesting cases, they would fill more limited roles, supplying only certain of the deficiencies of the machine. This would not only make it possible to experiment earlier with Verbmobil in use but could give more pointed information on the properties required of parts of the system that were yet to be completed or that were in need of further improvement.

Our report to BMFT took the view that the design, programming, and documentation of Verbmobil should be uncompromising, never sacrificing elegance or theoretical motivation to performance or conve-

nience. In particular, a regime such as the one based on tasks that are scheduled on an agenda should be adopted and used uniformly to control the operation of the main processes, almost essentially all of which should be nondeterministic. An agenda-based system also allows modules to be separated from one another as dictated by the underlying theory while at the same time allowing freedom in the way the space of solutions to the overall problem is searched. We also took the view that, while the linguistic formalisms used in Verbmobil are undoubtedly important, the common tendency of investing great importance in particular notations in the belief that notation confers scientific respectability should be resisted.

We were strongly of the opinion that the Verbmobil program should undertake empirical investigations of translation and interpreting as done by humans, as well as investigations of the properties of dialog in situations such as those for which Verbmobil is intended. Remarkably little is known about the process of translation as done by humans and remarkably little interest has been shown in such matters by those working on machine translation systems. We do not claim that the particular methods that humans use would necessarily be the only right ones for a machine to use, but only that the designers of machine systems would surely have much to learn from them. In particular, we are impressed by the extent to which translation is a matter of compromise among competing criteria, but we see this reflected little in the work that has been done on automatic systems. We therefore explicitly recommend that Verbmobil face squarely the fact that translation is inescapably a matter of compromise and adopt an approach in the spirit of "translation by negotiation", an approach which we elaborate on later in the text.

We strongly advocated pursuing policies that would tend to increase the performance of Verbmobil in the light of experience. These include techniques like *translation by example*, where the device maintains and uses its own long-term memory, and statistical information supplied to the system from the outside.

Many of the recommendations of the report underying this book were taken up by BMFT in drawing up their detailed plans for Verbmobil. Others, notably those concerning the domain of discourse and the two principal products were not. BMFT received a first round of proposals from companies and university departments hoping to participate in the research in the Fall of 1992 with the intention of letting contracts early in 1993. The first phase of the research, which will run for four years, is intended to result in a so-called *demonstrator*. This will not be a single device, and probably not even a single computer

program. It will probably consist of a small set of devices and programs each of which could serve as a major component in an early version of Verbmobil and giving good reason to suppose that it would interact in the intended ways with other components of the system. On the basis of the demonstrator, a judgement will be made as to whether the program as a whole should continue.

2

Machine Translation

2.1 Why is Machine Translation Hard?

Why is translation so hard? The answer to this question turns on the fact that language is "situated." This is an important notion which we will explore first. We will then go on to show how it bears on fundamental questions concerning translation. We begin with some examples.

2.1.1 Situated Language

2.1.1.1 Feynman's Safe Lock

In the following passage, Richard Feynman explains how the combination lock on a safe works.[1] We shall concentrate our attention on the first paragraph, but we include the remainder for the benefit of a reader who might wish to continue our analysis. It also provides some examples of how later context can serve to clarify what goes before.

> There are three discs on a single shaft, one behind the other; each has a notch in a different place. The idea is to line up the notches so that when you turn the wheel to ten, the little friction drive will draw the bolt down into the slot generated by the notches of the three discs.
> Now, to turn the discs, there's a pin sticking up from the first disc at the same radius. Within one turn of the combination wheel, you've picked up the first disc.
> On the back of the first disc there's a pin at the same radius as a pin on the front of the second disc, so by the time you've spun the combination wheel around twice, you've picked up the second disc as well.

[1] See Feynman 1985, p. 139.

Keep turning the wheel, and a pin on the back of the second disc will catch a pin on the front of the third disc, which you now set into the proper position with the first number of the combination. Now you have to turn the combination wheel the other way one full turn to catch the second disc from the other side, and then continue to the second number of the combination to set the second disc. Again you reverse direction and set the first disc to its proper place. Now the notches are lined up, and by turning the wheel to ten you open the cabinet.

As an example of applying an old language to a new subject, without benefit of blackboard, this is surely a gem. It is therefore surprising to find how much of what one must come to know in order to understand is not made explicit in this text. Here is a short commentary on the first paragraph.

"There are three disks on a single shaft." In what sense are they "on" the shaft? Balancing flat on the shaft? No; clearly each disk has a hole in its center through which the shaft passes. Actually, you have to read on a bit to be sure of this. Leaving it vague for the time being seems to do no harm.

The disks are "behind" one another. But, which face of a disk is its front and which its back? The answer is that it doesn't matter. If there is any reason to prefer "behind" to, say, "beside," it is possibly to invite the reader to imagine the shaft as parallel to the line of vision, so that the nearest of the disks would obscure the other two.

"Each has a notch in a different place." Each what? Well, the only thing we have been told about is disks, so he probably means each of them. But this notch; could it be in just any different place—say, projecting inwards from the hole through which the shaft passes? No. Somehow it is natural to begin one's consideration of a disk from the outer rim, and somewhere here is where the notch will have to be. And, what does it mean for the places where the notches are to be different—or to be the same?

"The idea is to ...." What idea? This is the first we have heard of an idea.

In what sense are the notches to be "lined up"? This is where imagining the shaft to be parallel to the line of sight comes in, because it turns out that, when they are lined up in the intended sense, the closer notches would reveal not the disk behind, but the notch behind. With them lined up, you would see right through, and that gives you the right idea.

What does it mean to turn the wheel (what wheel?) to ten? The intended addressees of this explanation are not supposed to know how a safe lock works, but they would be in no position to read this if they had never seen such a lock or noticed that it is operated by a single wheel, with numbers inscribed on it. There is also a mark beside the wheel and lining up a number with the mark counts as turning the wheel to that number.

Now, take my word for it: the "little friction drive" was not introduced earlier. Feynman speaks of it as a familiar thing, but this is the first we have heard of it. When we are being cooperative, however, and Feynman wants a little friction drive, we give him a little friction drive!

Finally, consider the fascinating phrase "the slot generated by the notches." First, this is an unconventional use of the term "generate," at least to the extent that what is generated is open space. Second, the phrase does a lot to tie together our understanding of what has gone before. For this phrase to refer successfully to something in your mental model of what Feynman is talking about, you must have the holes in the center of the disks, the disks arranged like wheels on the shaft, and the notches at their outer edges.

Feynman's achievement in this short passage is something of a *tour de force*, and the substantial investment that the reader must make to understand it is no more than is required for most conversation and requires no unusual skill or intelligence. But, though skillfully constructed, the text is technical and by no means poetic. The points we hope to make with it are these:

1. The meaning that a word, a phrase, or a sentence conveys is determined not just by itself, but by other parts of the text, both preceding and following.
2. The meaning of a text as a whole is not determined by the words, phrases and sentences that make it up, but by the situation in which it is used.

2.1.1.2 "Open"

As a further example that goes particularly to the latter point, consider a notice on the door of a store, bearing the single word "Open." This is the conventional way of announcing to the public that the store is, at that moment, doing business. A person wishing to purchase something that the store had for sale could walk in now and expect to be served. On the other hand, if the word "Open' appeared on a large banner over a newly constructed building, it should probably be taken

to mean something similar, but not identical. It probably means, not that the enterprise housed in the building is being carried on at this very moment, but that the construction had reached a point where, on the days and at the times when one would expect such an enterprise to be functioning, it in fact would be. If the notice on the store door were displayed on Sunday or in the evening, when the store was in fact locked up, we would say it was misleading. Not so for the banner. It is surely inappropriate to think of these as two different meanings of the word "open" that should have there own entries in the dictionary. The meanings are basically the same and a reasonable person just does not expect a large banner, that requires ladders for its installation, to be taken down every evening. In many other languages, the same word would be appropriate in both situations, but not in German, where the door of the store would have to say "Offen," and the banner on the building "Neu Eröffnet."

The word "open" is also to be found on the side of the top of a milk carton where it can be opened most easily. It is not telling us that the carton is open or urging us to open it, but drawing the attention of anyone wishing to gain access to the carton's contents to the part of it most likely to repay their attention. The word "open" flashing on the control panel of a micro-wave oven at the end of the cooking time tells us to open the door, but makes no suggestion about how to go about it. The outside doors of a train carriage in Europe are often operated by a handle that can be set in two different positions. These are often marked on the door with the words "open" and "closed." Unlike the sign on the door of the store, these are not conveying conflicting messages about the current state of the door, but telling us how to tell whether the door is open or closed, given the position of its handle.

Speakers of English, even non-natives who do not understand all the subtleties of the language, can probably be counted upon to interpret these signs in the intended way, not because of what they have learnt about English or from information they would expect to find in a dictionary, but because each of them is the interpretation that makes sense in the given situation. It would not make sense to print a message on a milk carton declaring it to be open. We expect the carton to be closed when it comes from the store and that it will be open some time later. Nothing printed on it could possibly give useful information about its current state. Likewise with the door of the train. It cannot be open and closed at the same time. The salient question, about which it is in order to convey some information, is how the handle can be used to make it open or closed, or how we can know its status from

the position of the handle, or where to place the handle in order to put the door in a desired state.

The meaning of a word or phrase depends on the situation in which it is used, but it also depends on the situations in which the same word has been used in the past. That is, the meaning of a word evolves as different circumstances arise, and choices are made as to what word will cover each circumstance. The word "open" has a basic meaning of 'having no confining barrier,' but it has been extended to cover a wide variety of specific cases, while at the same time *not* covering other seemingly similar cases. For example, an "open golf tournament" is one that anyone can enter, an "open morning" is one free of appointments, an "open question" is one that is undecided, an "open market" is one without excessive tariffs or regulations, an "open job" is one that has not been filled, an "open football player" is one who is not being defended by an opponent, and an "open set" in mathematics is one that is continuous in a certain technical sense.

However, note that while we say "his eyes are open" when they are uncovered and ready to receive stimulus, we do not say "his nose is open" when he is ready to smell, nor "his feet are open" when his socks are removed. We speak of "empty glasses" and "free-range chickens" when "open" might just as well have been used. The following table shows that most uses of "open" in English translate to "öffnen" or "freie" in German, but there does not seem to be a clear way to predict which will be used. There are also examples of words other than "open" that translate to "offen."

English	German
in store door	Offen
on new building	Neu eröffnet
open door	Tür öffnen
open golf tourney	Golfspiel eröffnen
open question	offene Frage
open eyes	offene Menge
open job	freie Stelle
open morning	freier Morgen
open football player	freie Fussballspieler
loose ice	offenes Eis
blank endorsement	offenes Giro
private firm	offene Handelsgesellschaft
unfortified town	offene Stadt
blank cheque	offener Wechsel
to unbutton a coat	einen Mantel öffnen

This example stresses the fact that the relation between the basic meaning of a word and its circumstances of use is very complex. Different languages and different cultures appeal to different circumstances of use, so the mapping from one language to another is far from direct.

A word or phrase in one language can seem to contradict the meaning of its translation in another language, as in the case of the English "health insurance," which translates into German as "Krankenversicherung" (sickness insurance). The relation between the two parts of the compound is based on the different *case relations* between them. The insurance is *for* health, and *against* sickness. Though this example makes for an interesting anecdote, it is not particularly problematic because the phrases are fixed and presumably recorded in the lexicon.[2] But other cases arise where this is not so. A notorious example is to be found in pairs of languages that have different conventions for answering yes-no questions that are posed in the negative. So, if an attorney asks a defendant, in Spanish "You didn't leave your apartment all evening" The reply "sí" should properly be translated as "no," and "no" as "yes." It would be perverse to ascribe the difference to different meanings of the Spanish "sí" and "no," just as it would to ascribe them to different meanings of English "yes" and "no." What differs is clearly what is taken as being assented to or denied—the negated or the unnegated proposition in the question. Burk-Seligson 1990 points out the not surprising fact that this distinction causes great difficulties when the testimony of a witness is being interpreted, and reports the following interchange:

> DEFENSE ATTORNEY: So are you saying that you wouldn't have told the border patrol officer?
> INTERPRETER: Excuse me, sir. I have to tell you that you're using the negative all the time and his answer really doesn't mean much when you're using the negative form of questioning because when he answers "no" it actually comes out "yes." If you say "Wouldn't do this" or "Wouldn't do that. Yes, I wouldn't." You see what I mean? You're using the negative and it's confusing him tremendously.

[2]While one might not want to list *open job* as an idiomatic phrase in the lexicon, some stipulation is required to capture the fact that *open job* is good English and *free job* is not. That stipulation will have much the same effect as listing *open job* as a noun

Language	Word	Gloss
English	validate	
	cancel	
	stamp	
	punch	
French	valider	validate
	composter	punch
	obliterer	cancel
German	entwerten	invalidate
Italian	validare	validate
	invalidare	invalidate
	cancellare	cancel

FIGURE 1 Words for Validating Bus Tickets

2.1.1.3 Bus Tickets

Another striking example is provided by those places all over Europe, where quantities of bus and tram tickets can be purchased in advance and then used one by one for individual journeys. In the case of Spanish "sí" and "no," one might claim that it was not so much the situation of use that was in play as differing, but fairly static linguistic convention. Here, the case will be different.

Bus and tram companies have found it necessary to provide a means for determining when particular tickets have been used, and this usually takes the form of punching or marking it in a characteristic way using a machine provided for that purpose in the vehicle or at the bus stops. A comparable operation, often referred to with similar language is performed on postage stamps at the post office so that it will not be possible to reuse them. A list of words that are used to refer to this operation in different languages are shown in Figure 1.

The most striking thing about this list is that it contains words which, in almost any other context, would have to be taken as opposites of one another. Different parts of Italy differ in whether they use "validare" or "invalidare," and, whereas "valider" is the usual term in French Switzerland, "entwerten," which on most counts means the opposite, is the standard term in German Switzerland.

The choice between a word meaning "validate" and one meaning "invalidate" has to do with a mental model that is being invoked when the term is used. According to one model, one purchases a set of new, valid tickets each of which will retain its validity up to the moment when it is used for a particular journey. That it has been used is

recorded on the ticket in a way that an inspector can verify and, since the ticket cannot be used for any subsequent journeys, it loses its validity. It is illegal to use a valid ticket to justify one's presence on the bus. Otherwise there would be nothing to stop you using the same ticket for another trip later.

The other model is based on the notion that the tickets are only potentially valid when they are purchased. When one begins a journey, one chooses one of these tickets and uses the machine to make it valid, a status that it will retain only so long as the journey is in progress. After that, it becomes invalid.

Notice that the words in the above list also reflect other models. The ones behind "validate" and "invalidate," are quite similar in that they are associated with an abstract property of a ticket, namely that of validity. Other terms refer to this notion only indirectly. Directly, they refer to the physical effect that the machine has on the ticket, namely punching or stamping.

Consider now the situation in which rules for the use of bus tickets and warnings to those who would transgress them are to be translated from German into English. How should we treat the word "entwerten," all other things being equal? "Invalidate" is presumably not an option—well behaved English speakers do not travel on invalid tickets. The next nearest term would be "cancel," but that is much more at home in the world of postage stamps and sounds sufficiently innovative when used of a bus ticket to suggest that something unusual is going on. No such suggestion is appropriate. So we are left with "validate," which we can support on the basis of a model of the process that justifies an apparent contradiction, but a model to which one must have access if one is to use the term. Otherwise, we must choose something like "punch" or "stamp," which sound perhaps most natural of all in English.

The trouble with "punch" and "stamp" is that we cannot make a choice between them without knowing more about what the machine actually does than the original text gives us evidence for. The process referred to with the word "entwerten" could be realized by stamping, or punching, or possibly in some other way. Under most circumstances, a professional translator would probably choose "punch," because that is what happens to bus tickets in much of the English speaking world and therefore makes for the smoothest reading. But if we were imagining a technical translation, one about the rules that must be obeyed in Switzerland. In this situation, something like "validate" or "cancel" would have to be used.

One might attempt to dismiss the paradox that words meaning

"validate" and "invalidate" can refer to the same aspect of the same situation by claiming that we are in fact dealing with special technical meanings of these words which should each have its own subentry in any complete dictionary. However, while it might indeed be helpful for a dictionary to draw attention to the paradox, it is by no means clear if special meanings are involved. According to the view that that there are special meanings, the German "entwerten" presumably means:

(1) a. to invalidate, or cause to be no longer valid.
b. to punch or stamp a ticket to indicate that the journey whose payment it guarantees has started.

The trouble with this is that it suggests that what happens to the ticket has nothing to do with the notion of validity that figures in the first definition. The right solution seems to be to admit that the words have the meanings they seemed to have before the question of bus tickets arose, and that these meanings can be used in two different metaphors about the proper use of bus tickets.

2.1.1.4 An Example from Computer Programming

The words we have been discussing were clearly chosen as especially striking examples of situated language. But one should not take this as meaning that language contains only occasional instances of situated usage. All language is situated. It is true, of essentially any utterance, that it could be lifted from the context in which it was used and placed in another in which, necessarily, it would function differently. Taken out of all context, it would have to be regarded, at best, as vague. It is important to realize that indeterminacy, far from being a deficit, is one of the great strengths of human language. It is this property that makes language as adaptable as it is to the great variety of unexpected tasks that people put it to. Furthermore, indeterminacy is not a property of sloppily used language, or of artistic uses such as poetry. Consider, for example, the following, which is the first sentence of section 1.5 of *C: The Programming Language* (Kernighan and Ritchie 1978):

(2) We are now going to consider a family of related programs for processing character data.

Should "are going" be taken as an indication of future tense, or as a claim that we are on route to some place? What kind of programs might we be talking about that makes it reasonable to speak of a family of them? Being a family, we should expect the programs to be "related," so is this word simply redundant? What kind of of data are character data? Are they data about the characters of specific people? Are they data tending to bear on the characters of people?

These questions are largely settled once we situate the sentence in the context of a programming text.

The sentence from Kernighan and Richie could just as well have been:

(3) We are now going to consider a family of related programs for treating character data.

Suppose, however, that you were told that this was a sentence from a book on social psychology. Then the programs might have been treatment programs and the data might indeed have concerned the characters of human subjects.

It is precisely the indeterminacy of the words and phrases in the sentence that make it possible for Kernighan and Richie to refer so easily to a new notion. We know that the discussion is about programs and that when things are in families, there is more than one of them. This is about all of the meaning of the word "family" that is being invoked. We are told to expect that the programs in the family will be related, and this information, though vague, is substantive. It says that they will have something in common that will be interesting from the point of view of the topic under discussion.

2.1.1.5 Discussion

In the words of Barwise and Perry 1983 p. 37, "meaning underdetermines interpretation." A language makes available to its users, words and sentences that are flexible, or vague enough, so that they can fit a variety of situations. When placed in those situations, on the other hand, they acquire a precision that no grammar or lexicon could possibly have provided for.

There is nothing mysterious about the notion of a linguistic system that acquires flexibility through indeterminacy. The mystery lies in the mechanisms that remove this indeterminacy when the utterance is placed in context. Words and phrases refer to particular individuals, events, sets, and so forth, by attributing certain properties to them. These properties are shared by many individuals, events, or sets in the world at large, but are typically unique to the intended referent in the particular context. Each utterance also changes the context against which the interpretation of following utterances will take place. This mechanism is poorly understood. The process by which hearers merge information from an utterance with information from the context is sometimes referred to as the "resolution problem."

As we have said, this indeterminacy is not something to decry and to attempt to avoid because, for all the problems it may bring, it is also one of the main sources of the remarkable power and flexibility of

human communication in general and ordinary language in particular. If I perceive that the chandelier is about to fall on your head, I can simply shout "Look out!," leaving it to you to search the situation you are in for something to look out for. A more explicit message might actually defeat my purpose. By involving the active participation of the listener as well as the speaker in this way, language becomes a highly efficient and rapidly adaptable tool.

While less determined linguistic usage is doubtless often less precise, it is also often easier to understand. Much legal language is inaccessible to lay people, not so much because of the special vocabulary that is used as because the requirement of precision puts trivialities on the same footing as the main point. Herein lies the second reason to rejoice in situated, and therefore imprecise, language. The first was that it makes the language adaptable to unforeseen situations, and the second is it achieves great efficiency by allowing its users to say only what is most important. Shouting "look out" is a good illustration of this also and our commentary of the Feynman fragment at the beginning of this section points to example after example.

The notion that we say only what is important is the key idea underlying Roger Schank's notion of *scripts* and its generalizations. He invites consideration of sequences such as the following:

(4) I went to "La Tour Eifel" last night and had a wonderful dinner. But I didn't have enough money for the check.

We take it that the "La Tour Eifel" is a restaurant and that I had dinner at that restaurant, even though the first sentence does not actually say that the wonderful dinner that I had was actually consumed there. But the main point comes in the second sentence. Presumably I went into the restaurant, was shown to a table, read the menu, selected something from it, had it served to me in due course, possibly ordered desert following a similar procedure, had the check delivered to me, and then discovered my insolvency. None of these events is thought worthy of mention because they are common parts of going to a restaurant and will be understood as having taken place without being mentioned. The only things that are mentioned are those that depart from what would normally be expected. A cooperative listener will assume these details and be grateful not to have to suffer through a recital of them. The main reason why legal text is often so tedious and insistent on rehearsing every detail, however pedestrian, is that the readers of the text cannot all be assumed to be cooperative.

The sociologist Basil Bernstein, much maligned for the conclusions he drew about specific social groups, distinguished between two com-

municative modes which he termed "restricted" and "elaborated" code. Restricted code, in Bernstein's sense, involves highly efficient communication that employs compressed, "simplified" language and relies heavily on constantly evolving contextual information to establish references, and interpretation in general. Restricted code is highly dependent on context. Elaborated code, by contrast, involves more verbose description, more intricate grammatical devices, explicit identification of referents, and less reliance on contextual information. So, for example, elaborated code is more suitable for situations in which the interlocutors lack a common ground and accessible shared assumptions, or where they stand in an adversarial relationship to one another.

2.1.2 Translation Mismatches

The main point of our extended discussion of situated language is that much of what an utterance or a text conveys is often present in it only by implication. There is another side to that coin, namely that much of what is present in the utterance or the text is there, not because it is essential to the message, but because it is required by the language or the culture. Thus sentences in context are both under and overspecified with respect to what they convey.

But the challenge of machine translation does not end with the task of understanding. Beyond understanding what is conveyed, there is the problem of constructing a sentence in the target language which conveys as closely as possible what was conveyed by the source. This is a problem which most humans do *not* find easy. Translating takes not only knowledge but ingenuity.

In this section we try to get some handle on why translation is difficult by discussing a series of examples of what we shall call *translation mismatches*. Translation mismatches are cases where the resources of the target language force a translator to add or delete information not present in the source.

2.1.2.1 The Semantic Grid

In translation, it is often difficult, if not impossible, to find a word or phrase in the target language that specifies just what is specified by a particular word in the source, and leaves just the same other things underspecified. So, for example, there is no French word that covers that same range of properties as the English word "chair." "Chaise" and "fauteuil" are both candidates, but both are more specific about whether the chair has arms, and whether there is padding. The French noun "porte" leaves unspecified the information necessary to distinguish between a door and a gate to an English speaker. It is very

difficult to say just what the properties are that distinguish a door from a gate, but no one doubts that there are some, and there is no word that is neutral between them. In both cases translating between English and French confronts the translator with a *translation mismatch.* A mismatch requires a translator to add or delete information; typically, resolving the mismatch involves consulting the context to determine what the correct information to add is, or what information may be safely deleted.

Sometimes, the additional information crucial for translation can be quite subtle. Consider, once again, the distinction between the French words "chaise" and "fauteuil," both of which would often be translated as "chair." Now consider the following pair of English sentences

(5) a. I found your gloves in that chair.
b. I found your gloves on that chair.

All other things being equal, it would probably be a good bet to translate the first use of "chair" as "fauteuil" and the second as "chaise" because the arms that a chair must have to be a "fauteuil" give it some of the properties of a container, and therefore something that an object of appropriate size could be "in." A "chaise" has only an open surface on which to support other objects, and they will therefore be found "on" it.

It is much as though each language placed a grid down on reality and agreed to call by one name whatever fell within a given square of the grid. But, since the grid is different for each language, information is required, over and above what is actually provided in order to determine in which square of the overlapping grid the intended meaning lies. The area covered by the English word "fish" covers two sections of the Spanish grid, one corresponding to the word "pez" and the other to the word "pescado" (see Barnett et al. 1991a). The latter refers to a dead fish that is intended to be eaten, and would be the right translation in a sentence like "We had fish for dinner last night," whereas "pez" would be appropriate in "She has a tank with all sorts of fish in her front room." German divides the territory occupied by the English word "go" into two major areas, one occupied by "gehen" and the other by "fahren"; the former is for going on foot and the latter for going in a vehicle. So, we have

(6) a. Ich gehe in die Küche.
b. I am going into the kitchen.

(7) a. Ich fahre mit dem Zug.
b. I am going by train.

(8) a. Der Wagen fährt schnell.

b. The car goes fast.

Russian partitions the territory of "go" in a much more complicated fashion, the precise details of which are beyond the scope of this study. The verbs "xod'it'" and "idt'i" are used when the moving object proceeds under its own power, and so would be appropriate for (6) and (8). "Yesdit'" and "yexat'" refer to traveling, as in (7). The point is that there is no way to translate a simple sentence like "I am going into the kitchen" into German or Russian without making explicit information that is only implicit, or indeterminate, in the English.

The designers of translation systems can take some solace in the fact that, though it is often extremely difficult to see how the information that is missing from the utterances we have been considering can be supplied, at least the situations themselves fall into well known categories recognized in the grammars of the languages. But it is not always so. It can happen that a piece of the conceptual map that is covered by some square in the grid of one language is not covered at all by the other one. The extremely common German adverb "beziehungsweise," for example, has no equivalent in English or French. It means "according to x," where "x" is left unspecified. But in English, the phrase "according to" requires that the object be specified, so something must be provided. Consider the German sentence

(9) Aus der Ankunfthalle müssen Sie durch die rote, beziehungsweise grüne, Tür gehen.

In English, we might say

(10) You must leave the arrival hall by the red or the green door.

or, more precisely,

(11) You must leave the arrival hall by the appropriate door, the red or the green one.

To preserve the adverb—though only in the form of a prepositional phrase, one might say something like

(12) You must leave the arrival hall by the red or the green door, as appropriate.

More difficult would be the German adverb "voraussichtlich" as in

(13) Der Zug wird voraussichtlich zehn minuten spät ankommen.

"Vorausichtlich" comes from the verb "voraussehen," meaning *to predict*, or *to expect*, but the meaning is not *predictably*, but rather *according to prediction* or *expectation*. The above sentence would be fairly well translated as

(14) a. The train is expected to be ten minutes late.
or
b. The train will probably be about ten minutes late.

In the second of these translations, we have done something that is logically quite complex but entirely natural to a human translator, namely to replace the notion of expectation by that of probability and approximations, enshrined in the words "probably" and "about" respectively.

2.1.2.2 Function Words and Affixes

A notorious place in which translation mismatches arise is in function words, prefixes, suffixes, and the like which, as well as having independent meanings, must be present in order for an utterance to be grammatical. So, for example, common, count nouns in the Western European languages usually are required to be accompanied by articles. When the article is missing it is semantically significant, as in "there's cat all over the driveway," where "cat" takes on a very different meaning than the usual phrase "the cat." But articles are not required in Russian, Chinese, Japanese, and many other languages. Translating from these languages into English is greatly complicated by this fact because it is very difficult to state just what the information is that the articles carry. The same European languages require every noun to be singular or plural, which neither Japanese nor Chinese does. The European languages insist that every finite verb be in the present or the past tense, information that is not usually explicit in Chinese. Many English verbs demonstrate a distinction between progressive and nonprogressive forms which encodes information not available in many other languages. This distinction can also be made in Italian and Spanish, but it is not obligatory. In these languages, the simple form corresponds to both English forms, the progressive being used at the speaker's discretion when the extra information is considered particularly pertinent. Another set of distinctions that cuts across these has to be made in Russian.

To say that certain information must be provided only because it is a grammatical convention of the language that it be there is not to say that it is not important, as anyone knows who has heard a Finn say something like "I was just talking to my mother, and he says you can come with us to Helsinki." Finnish does not require the distinction of sex to be made in pronouns and making it in English is a difficult habit for Finns to acquire. In this example, the reference is to something previously recognized in the same sentence so that it is not difficult to sort out. Where the pronoun refers to something further away, the

confusion that results from getting it wrong can be a great deal more severe.

It can happen that the lacunae that there seem to be in a language, by comparison with another, are systematic and quite pervasive. In this connection, pronouns are often cited because the grammars of some languages allow complete omission of a noun phrase where, in other languages, it would be replaced by a pronoun. This is the case with subject pronouns in many languages such as Italian and Spanish, so that we get

(15) È arrivato.

in Italian, meaning "he/she/it arrived." In Japanese, pronouns are omitted in all positions, generally leaving a large number of matters to be resolved before a translation is possible into a language like English or German.

There are situations where pronouns are routinely omitted in English, notably in special dialects or jargons. Consider the famous, if dated injunction:

(16) Do not bend, staple, or fold.

The object of the three verbs is the very card on which the sentence is printed. Cookbook recipes often omit Noun Phrases and other material in places where omissions wouldn't be tolerated in ordinary speech:

Mix butter, sugar and flour and place in a metal bowl.

The second clause here, "place in a metal bowl," is an example of a special form of English found in recipes. Normally, transitive verbs like "place" do not allow their direct objects to be omitted, as can be seen from the oddness of the following sentence:

John mixed the butter, the sugar and the flour and placed in a metal bowl.

2.1.3 What Counts as a Translation

In order to program a computer to perform a specific task, one must have a precise idea of what that task is. In this section we address the question of what a translation is, not to propose a definition, but to point out some of the difficulties of formulating one.

There are many situations in which a human translator exercises a considerable amount of ingenuity and originality in constructing a translation, sometimes adding and deleting information that does not seem crucial for the text on hand. Here are some more simple examples:

(17) a. English: He asked where he should stand.
b. French: Il a demandé où il devait se mettre.

This is another example of a translation mismatch. The translation of "where he should stand" into French is, literally, "where he should put himself." The notion of standing is omitted because the translator can count on its being effortlessly reconstructed by the reader, given the context and the situated nature of language. To include it in the French would be awkward.

(18) a. English: He gets up early.
b. French: Il est matinal.

This translation would not be appropriate in a description of people who worked the night shift, where an early riser might get up in the early afternoon. And it would be equally wrong in a discussion of when the members of a group of astronomers arrive at the telescope on top of a hill to start the evening's observations. That might be in a text something like this:

(19) Most of them are on the hill a little after sunset. He gets up early and prepares the film.

It is clear that a translation is some transformation of the source sentence into the target language which preserves certain properties. What exactly must be preserved is difficult to say. Many of the examples we have discussed suggest that it is not the meaning, in any but the loosest sense of that word. What we will prefer to say is that a good translation is one that preserves to the extent possible, the *intention* of the original. In other words, it preserves the intended effect on the recipient. Consider once again the sentences in (17). It is quite clear that the French and English sentence do not have exactly the same meaning. The French sentence talks about placing oneself, and the English about standing. There are situations for which the French sentence is appropriate but where a better English rendering would be "He asked where he should sit," "He asked where he should park," or "He asked where he should put his name," to mention but a few possibilities.

2.1.4 Ambiguity

Akin to indeterminacy, or underspecification, is the notion of ambiguity, which is also endemic to ordinary language. While it is not the same thing, it is difficult to imagine how a linguistic system that was based on underspecification could nevertheless fail to be hospitable to ambiguity.

2.1.4.1 The Lexicon

We have seen examples where the grid over part of the semantic landscape is coarser in one language than in another so that, when translating from the former into the latter, finer distinctions have to be made. We have also seen that this can be anything but a simple mechanical matter. The territory that one language covers with a simple vocabulary item can be left uncovered in the other language so that it takes a creative act on the part of a translator to render that vocabulary item. In terms of the semantic grid, we have an ambiguity when an item in one language covers two or more pieces of disconnected territory—it has two essentially unrelated meanings. Such cases are legion and, indeed, they are among the first things that a person with a passing knowledge of more than one language thinks of when invited to reflect on the difficulty of translation.

The English verb "book" means something like "reserve" in "he booked a room for the night," but also something like "cite" (American) in "the policeman booked him for speeding." The first of these would be "reserver" in French. The second would be harder to deal with; Collins dictionary suggests "mettre un procès-verbal à." The French word "temps" means both "time" and "weather" in English. Needless to say, the metaphor of the semantic grid is imperfect. In particular, the notion of squares on the grid being adjacent to one another is impressionistic at best, so that it often makes little sense to try to distinguish between ambiguity and different levels of specificity in a pair of languages. If we think of German "Fenster" and "Schalter" as being adjacent, then we would say that German requires greater specificity when translating the English "window"; otherwise we might wish to say that "window" is ambiguous. However, the question really turns not so much on adjacency in some notional grid as on whether speakers of the language in question see the word or phrase as having two meanings, itself no easy matter to decide. In the case of "book" there seem to be two clearly distinct meanings; in the case of "window," it is less obvious. From the point of view of translation, the distinction is perhaps not important; what is important is that two or more different renderings are required in some other language, and the criteria that distinguish them can be arbitrarily subtle. From now one, we shall use the term "ambiguity" to cover different levels of specification also.

Research in machine translation has recognized lexical ambiguity as an important problem from the start. Two kinds of solution have been proposed, one global in the sense that it takes properties of the whole

text into account, and the other local in the sense that it considers other words in the immediate neighborhood of the ambiguous one.

2.1.4.2 Lexical Fields

The global approach rests on the observation that the word meanings that are invoked in a text correlate with the overall subject matter of that text. In a text on computer science, the word "register" is much more likely to refer to part of a computer's hardware than to a book in which objects, people, or events are recorded. In the same text, a "file" is more likely to be a collection of data than a wood- or metal-working tool. In computer science, a "tree" is probably an abstract structure, and thus "une arborescence" in French, rather than a large plant, and thus "un arbre." An "arrow" is more likely to be a line ending in a v-shaped head, rather than a weapon, and thus "une arrête" rather than "une flêche".

However, this approach rests on probabilities, often impressionistically assessed at best. A word or phrase can be ambiguous even within a given field and could easily be used with more than one of its meanings with in the same text. Still in computer science, consider the sentence

(20) All of these expressions are typed.

Does it means that the expressions all have types or that they are all entered through the keyboard? In music, consider

(21) All the scores contain some examples of this.

Are the scores the transcriptions of complete musical pieces, or lines of music within such transcriptions (staves or systems)? When fields are combined in one text, obvious difficulties follow. What would "signature" mean in a work on music publishing: a key signature or a set of pages to be stitched together?

2.1.4.3 Selectional Restrictions

The second approach to resolving lexical ambiguities calls on what linguists since Chomsky (1957) call *selectional restrictions.* The approach rests on the observation that words enter into certain syntactic relations only with words and phrases that have particular semantic properties. To die, one must first be alive. The verb "die" therefore only occurs with subjects that denote living things. The word "number" could clearly occur in a computer-science context meaning, say, a real number, as in

(22) This number is assigned to the variable.

or meaning a quantity, as in

(23) The routine makes use of a large number of variables.

These must be treated as different meanings because, in French, German, and many other languages, they are translated differently. The question is how to distinguish the meanings from one another. The selectional restriction approach is to say that "assign," in computer science, takes as its object numbers of the first kind ("numero" in French). On the other hand, in the context "number of x," where "x" is plural, the second kind of number is in question ("nombre" in French). This would leave the question unresolved in a context like the following

(24) This routine computes a large number.

Selectional restrictions work to the extent that the language is being used very literally. When the word "die" is being used literally, its subject must be animate, but it also appears in sentences like the following:

(25) a. The program died before it reached this routine.
b. The bill will probably die in the Senate.
c. The movement died in the late thirties.

While surely metaphorical, these are not particularly poetic usages and could easily occur in quite technical contexts.

What appears to be happening in cases like these is somethings similar to *coercion* in a programming language. In programming, it is often permissible to specify that an operation should be applied to a piece of data even if that piece of data is not of the kind that the operator is designed for. The requirement is only that the data be capable of being converted, or *coerced,* to a corresponding value of another type. So, the expression `2+"4"`, which calls for a number to be added to a string of characters, may be allowable because the character string happens to name a number and, in this case, can be coerced to the value 4.

A sentence like "The program died before it reached this routine" depends for its understanding on the willingness of the reader to think of the program as having some of the attributes of life. In other words, it is not so much that the verb "die" requires a living subject as that it invests its subject with life. This is common enough with programs, as exemplified by:

(26) a. The program was still alive after the machine was supposed to have crashed.
b. The program expects a comma at the end of that line.

c. The program wants to read more data even though it has reached an end of file.
d. You can kill the program by typing control-C.
e. The programs talk to one another through a pipe.

For some further discussion of selectional restrictions, see section 2.5.1.2.

2.1.4.4 Collocations

Related to ambiguities that can be resolved on local criteria are *collocations*, in which a word with a particular meaning is supplanted by another word because of other words in the environment. So, for example, we have the phrase "to allay fear," in which "allay" means something like "neutralize." But "allay" is the colloquial word to use in the specific case where the object is fear. On the other hand, when the object is "hunger," the right verb to use is "assuage." A fairly large number of verbs (sometimes called "light verbs") mean essentially "do" when put together with appropriate objects. The following is only a very partial list:

verb	object
make	a U-turn, an attempt
give	a cry, a shout, a moan
take	a walk, a bath, breakfast
have	a bath, breakfast
perform	an operation
turn	a somersault
commit	a crime, perjury, suicide

There is no reason to expect another language to use different verbs in all these cases. More to the point, there is no reason to attempt to relate this special meaning to the other ones that these verbs also have.

2.1.4.5 Reference

So far, we have been considering indeterminacies and ambiguities in the meanings of words. Equally important are indeterminacies and ambiguities in what words refer to. Consider the sentence

(27) John got himself into this mess, and John must get himself out of it.

It is about a person called John who is referred to in the sentence in four separate places, twice by name, and twice by the word "himself." It is theoretically possible that two people are being referred to, one in the first half and one in the second, but since no information is provided to distinguish two people called John from one another, it is

reasonable to suppose that they are the same. There is no doubt that the first instance of "himself" refers to the first instance of "John," and likewise for the second instance because, if it were somebody else, the grammatical conventions of English would have required the use of "him" rather then "himself."

Essentially the same meaning could have been conveyed by

(28) John got himself into this mess, and he must get himself out of it.

This time, however, there is more potential for ambiguity because the word "he" could refer to someone other than John. Now suppose that this second sentence were a translation from Finnish, which does not make a gender distinction in its pronouns. It would also have been possible to translate the Finnish sentence as

(29) John got himself into this mess, and she must get herself out of it.

Pragmatically, you may say, "he" is overwhelmingly more likely. But the feeling that it is more likely comes not from any linguistic properties that it has, but from our incidental knowledge that the person most in need of extracting him or herself from a mess is the person that previously got into it. However, it is by no means clear how knowledge like this could be made accessible to a machine translation program. Situations like this in which some arbitrary piece of knowledge about the world is required to decide the likely referent of a pronoun are extremely common. One is to be found in the following pair of sentences.

(30) a. The police refused the students a permit because they feared violence.

b. The police refused the students a permit because they advocated violence.

In all likelihood, "they" is intended to refer to the police in the first sentence, but to the students in the second. The knowledge needed to determine this is common sense knowledge of the modern world such as comes form reading the newspapers. But it is crucial knowledge for translation because if, for example, "police" and "students" have different genders in the target language, then the translation of "they" could well be different.

In general, it is not sufficient to know what a pronoun is intended to refer to, but also what noun was used for that referent when it was introduced into the text. The reason is that, in general, the properties that are reflected in a pronoun are those of another word and not inherent properties of the object referred to. Consider the following:

(31) The shoes are very cheap. However, they are not for sale in America.

Two possible translations into French might be the following:

(32) Les souliers sont très bon marchés. Cependent, ils ne sont pas en vente en Amérique.

(33) Les chaussures sont très bon marchés. Cependent, elles ne sont pas en vente en Amérique.

The point is simply that the two common French words for shoes have different genders, and different pronouns are therefore required.

2.1.4.6 Syntax

The kinds of ambiguity that spring immediately to the mind of a linguist have to do with grammar, and particularly, syntax. They arise because the rules of the grammar allow the words in a sentence to be collected into phrases in more than one way. Consider the following pair of sentences.

(34) a. Attach the amplifier to the output terminal with the red wire.
b. Attach the amplifier to the output terminal with the red dot.

They could presumably be paraphrased as follows:

(35) a. Use the red wire to attach the amplifier to the output terminal.
b. Attach the amplifier to the output terminal that has the red dot.

However, as far as the grammar alone is concerned, each sentence could be given both interpretations. In fact, these sentences would be provided with five interpretations by many grammars. (We have provided each with an approximate paraphrase).

(1) Attach
the amplifier
to
the output terminal
with
the red wire
[Using the red wire, attach to the output terminal,
the amplifier]

(2) Attach
the amplifier
to
the output terminal
with
the red wire
[Attach to the output terminal that has the red wire,
the amplifier]

(3) Attach
the amplifier
to
the output terminal
with
the red wire
[Attach the amplifier that is to the output terminal
and that has the red wire]

(4) Attach
the amplifier
to
the output terminal
with
the red wire
[Attach the amplifier that is to the output terminal, with
the red wire]

(5) Attach
the amplifier
to
the output terminal
with
the red wire
[Attach the amplifier that is to the output terminal that has
the red wire]

As more prepositional phrases are added, the number of syntactic structures that the sentence has goes up exponentially in accordance with a series called the *Catalan numbers*. For this example, the relation between the number of prepositional phrases and the number of interpretations of the sentence as a whole begins to rise as in the following table:

PPs	Interpretations	PPs	Interpretations
0	1	5	132
1	2	6	429
2	5	7	1430
3	14	8	4862
4	42	9	16796

Accordingly, the following sentence has 429 structures, because it has 6 prepositional phrases:

(36) Attach input *of* the amplifier *to* the output terminal *of* the oscillator *in* the rack *with* the red wire *in* the plastic envelope.

Such sentences are by no means extraordinary, especially in technical texts. Furthermore, there are other constructions that contribute to the combinatorial explosion. One that is particularly favored by technical writers is the so-called *reduced relative* as in

(37) Attach the amplifier using the wire provided.

Here the word "provided" can be thought of as an abbreviated form of the relative clause "that has been provided"; hence the term. In this case the two interpretations are one in which the wire is provided, and the other in which the amplifier using the wire is provided.

Several other kinds of syntactic ambiguity are quite prevalent in texts that would be prime candidates for translation by machine. Consider, for example,

(38) Recursive functions and procedures are treated differently.
[Are all procedures, or only recursive ones treated differently?]

(39) The doctor found the patient lying on her side.
[Who was lying on her side? The doctor or the patient?]

(40) The aim is to determine how satisfied clients appear.
[How clients that are satisfied appear, or to what extent clients appear to be satisfied?]

It has been argued that some syntactic ambiguities, notably those involving prepositional phrases, are not as big a problem for machine translation as linguists tend to expect because the ambiguity can often be carried over into the target language. So, for example, the sentence

(41) Our knowledge of the universe comes from studying electromagnetic radiation emitted by heavenly bodies.

can be rendered into French as

(42) Notre connaissance de l'Univers provient de l'étude des rayonnements électromagnétiques émis par les corps célestes.

translating the prepositional phrases, as it were, in isolation, and simply concatenating them to form the translation. The trouble is that this approach does not always work, and in any case, it is generally only possible to recognize that it will work after the fact. In other words, one discovers that the intersection of the translations that can be based on the various different syntactic structures is not empty. This is not to say that this might not be a safe way to select a translation, but it does not imply that there will be any saving of labor.

More to the point, it is not always possible to arrive at an acceptable translation in this way. More typical are examples like the following

(43) He saw the girl with penetrating eyes.

where the phrase "with penetrating eyes" would be translated "de ses yeux pénétrants" if it means that he saw her with his penetrating eyes, and "aux yeux pénétrants" if the penetrating eyes were hers.

In this summary of the properties of language that make translation as difficult as it is, we have alluded occasionally to languages like Chinese and Japanese, but have drawn most of our examples from European languages because these are the ones we expect our readers to be most familiar with. But it goes without saying that, while the kinds of translation difficulties may be much the same, their severity will be much greater when the source and target languages are only distantly related, or not related at all. Translation is immensely difficult between any pair of languages, and overcoming the difficulty turns on solving the resolution problem.

2.2 Machine Translation Systems

2.2.1 Translation Quality

Research on Machine Translation came to a virtual standstill in the United States, and was drastically cut back in the rest of the world when the National Academy of Sciences published the report of the Automatic Language Processing Advisory Committee in 1966 (National Research Council 1966). Simply stated, the committee concluded that the research currently under way was unlikely to result in practically viable systems in the foreseeable future and that any system based on current work could not be economically competitive with traditional translation methods.

Research on the subject gained momentum again in the eighties and in several places, notably in Japan, it was rapidly hailed as a successful technology. But there is a disturbing anomaly here, namely that the systems of the eighties differed in little of substance from those of the sixties. Indeed, one can go further. By far the most commercially

successful Machine Translation system in the eighties was *Systran*, a system that was designed in the sixties and had remained substantially unchanged since. How is the anomaly to be explained?

It is possible that the quality of the translations that modern systems produce has improved without any significant change in the underlying technology because computers are now larger and faster; programming methods have been greatly improved; and modern linguists can do a better job of writing even the old style of lexical entry and translation rule. The fact, however, is that there is no evidence of any substantial improvement in translation quality.

Another possibility is to appeal to changes in the commercial climate. It is conceivable that the need for translation has grown so fast that it has been necessary to lower standards. This claim is difficult to assess. However, we have not seen lower standards actually acknowledged and, if there is such an effect, we doubt that it is great enough to have brought about such a thoroughgoing change in perceptions.

What does seem to be true is that greater ingenuity has been invested in seeking out special situations where very weak technology, like the machine-translation technology of the sixties, can do a creditable job. The best example of this is provided by the TAUM-METEO (Chevalier et al. 1978) system, developed by the University of Montreal which is applicable only to the narrow domain of meteorological reports. A similar approach has proved successful for the translation of maintenance manuals for narrowly specified kinds of machines. However, this approach has proved fruitful only in extremely narrow universes of discourse and it has crumbled rapidly when the attempt was made to expand the domain even by quite a small amount.

Here are some examples of the best that modern systems can achieve. It will presumably be clear to the reader that such small samples can give no more than a rough impression of the capabilities of the systems that produced them. Here is an example of the raw output of the SPANAM system (Vasconcellos 1985):

> The extension of the coverage of the health services to the underserved or not served population of the countries of the region was the central goal of the Ten-Year Plan and probably that of greater scope and transcendance. Almost all the countries formulated the purpose of extending the coverage although could be appreciated a diversity of approaches for its attack, which is understandable in view of the different national policies that had acted in the configuration of the health systems of each one of the countries.

What follows are translations of one Japanese sentence by five dif-

ferent Machine Translation systems (taken from Nagao 1989a). The first rendering is by a human translator:

1. A supersonic jet is as much as 10 times faster than a car.
2. In the jet plane and the car of the supersonic speed, the speed differs by 10 times or more.
3. More than 10 times speed is different in the supersonic jet plane and the car.
4. In the jet and the car of [[untranslated]] speed of sound, equal to or more than 10 times of speed is different.
5. By jet plane of a supersonic speed and motorcar, a speed of 10 times or more is wrong.
6. Speed differs in supersonic jet machine and an car above 10 times.

Finally, here is an example of the raw output of the Fujitsu ATLAS II system of a small text after several iterations in each of which the input was edited in light of the results of the previous run (Nagao 1991). Clearly this is far superior to anything that could be achieved in practice:

> The internal expression thus obtained is converted into an internal expression of a target language. The change of the sentence structure accompanies this conversion. This is the idea of the transfer method. The conversion of the structure is unnecessary for the sentence with a simple structure. If the input sentence is analyzed in detail, the structure conversion will become unnecessary. This is the pivot approach. Because more important problems exist, the discussion about the superiority or inferiority of the transfer system and the pivot approach does not have the meaning so much (Nagao 1991).

As the examples above illustrate, the claim that a Machine Translation system has succeeded does not mean it does what a human translator does. Presumably the strongest claims of success must be based on the cost-effectiveness of a total system, including the humans that are involved in the process. For example, it is reported that translators using SPANAM output work one-and-a-half to four times as fast as unaided human translators (see Vasconcellos and León 1985a). Weidner claimed that post-editing with its system takes one-third the time of ordinary translation. The Japan Information Center of Science and Technology claims a 50% saving in abstract translation with a system based on MU. On the other hand, Ryozo Akiyama (see Akiyama 1989) of Arthur Andersen & Co. in Japan says of his company's experiences with Systran:

> At this point, honestly speaking, Systran does not save us much money, but I think that we should not be too much rushing it for now. The Machine translation world is still growing up, and our constant feedback and long-range consideration as a user will help Systran refine further and contribute to bear bigger fruit in the future.

The main reason why success is now claimed for essentially the same systems that were dismissed as failures twenty years ago therefore has to do with the environment in which they are used. The developers of Machine Translation systems are quick to point out that even the work of professional translators is edited by a *revisor* whenever results of more than minimal quality are expected. We should not therefore be surprised, or hold it in any way against Machine Translation that the output of the machine must be submitted for review to a *post-editor*. What we should pass judgement on is not the raw output of the machine, but the effectiveness of the system as a whole, including the post-editor. On this criterion, for example, the findings of JICST (the Japanese Information Center for Science and Technology) are unequivocal. Abstracts that would cost 4,000 yen to translate into English by traditional methods cost only 2,000 yen to translate by machine. Nevertheless, the raw output of the machine they use for this purpose is often incomprehensible by even the most liberal standard. What is going on is actually quite clear. The translation is being done by a person whose official title is that of "post-editor," and the work is being aided greatly by a machine generated document, known in the technical jargon of the trade as a "translation." This question of the roles of man and machine in Machine Translation is one to which we will return shortly.

One of the most important sources of confusion in discussions of modern Machine Translation systems has to do with the terms in which quality is reported. Frequently what is reported is the percentage of the translation that is "correct." For a variety of reasons, such claims can almost always be discounted altogether except possibly as a means of calibrating the source of information. The claim that a Machine Translation system achieves a success rate of, say, 70% can only be made meaningful in a very carefully established context. It is clearly based on the notion that the correctness of a translation, or individual parts of a translation, can be assessed simply and unequivocally. The foregoing discussion in this chapter should be sufficient to cast considerable doubt on any such notion. Indeed, assessing the quality of translations is regarded by experts as being hardly less thorny a prob-

lem than that of making the translations in the first place. Secondly, it is remarkable how rarely we are told in terms of what units the success rate is being measured; 70% of sentences would be more impressive than 70% of words, for example.

2.2.2 Quality and the Users of Translation

It is commonly believed that, since scientists and engineers are interested in the meat of what a text conveys and not in subtlety and innuendo, they should be tolerant of lower quality in the translations they read than, say, historians and students of *belles lettres.* Put this together with another common belief, namely that scientific writing is less ambiguous than many other forms, and you have the basis for the widely held view that it should be possible to fill the needs of scientists and engineers with translations of a lower quality than might be required for other purposes. As far as we know, none of the empirical investigation one could easily imagine carrying out on this question has been done. Furthermore, while some scientists and engineers might find merit in the argument at first sight, they need only reflect on the care with which the words of a tightly reasoned technical argument or mathematical proof have to be read to see it in a different light. If scientific and technical material is difficult to write—and few doubt that it is—it is surely because it requires great pains to make it comprehensible at all, let alone, easy to read.

Now, it has indeed proved to be the case that, when Machine Translation is seen as most successful, it is when the subject matter is scientific or technological. For example, SPANAM, a system used by the Pan American Health Organization, translates documents in the field of medicine and health. The original Systran application was a system that translated scientific documents for the US Air Force Foreign Technology Division. Xerox successfully used the Systran system for translating maintenance manuals written in a restricted version of English. After extensive tuning to the domain, the Fujitsu ATLAS II system has been applied by Mazda in translating its auto service manuals (see Saeki 1989). Hitachi's HICATS system has been tested on translating Japanese patents, and Boeing has plans for using translation of restricted language in various parts of its operation.

However the reason for these successes does not lie with the subject matter alone. Sometimes it lies with the intended audience. In an interesting set of cases, the audience does not consist of scientists and engineers, reading for specific information or to improve their understanding of their field. It consists of people who read for what is sometimes misleadingly referred to as *informational* purposes. This

includes scanning an book or article to determine whether it would be worth translating properly, but the term is most often used in reference to information specialists, librarians, intelligence analysts, and the like. These are people whose primary concern is to classify the document rather than to understand it. Not surprisingly, these requirements put less demands on the translator.

Other cases exist in which the apparant success of a Machine Translation system has been too readily ascribed to the narrow subject matter alone. The Canadian TAUM-METEO system referred to earlier, is a case in point. But here, it is not just the domain of discourse but the type and style of document—what we might refer to as the *genre*. What is at issue is the restricted nature of the choices entertained by the author. Meteorology is, after all, a vast and complex field. Weather reports, on the other hand, constitute a species of discourse with a very special function, allowing only a very limited number of choices to be made by the author. In some applications, it has even been possible to replace the original writer and the METEO system with another system that composes the reports, in both languages, directly from the instrument readings so that there is no translation involved at all.

In 1972, Sinaiko and Clare completed an evaluation[3] of a version of the LOGOS Machine Translation system which translated aircraft manuals from English into Vietnamese. The Sinaiko-Clare study is one of the more interesting Machine Translation evaluation efforts in that it shows clearly how the overall benefits of a Machine Translation system can depend heavily on who uses the output and what their purposes are. The Machine Translation outputs were submitted both to language experts and to the mechanics who used them for abstract judgments of quality. The mechanics in many cases had far more favorable reactions than the experts, doubtless because, for their purpose, technical comprehensibility was critical, while matters of naturalness and grammar were not. Not surprisingly, comprehension tests showed that human translations got higher ratings than revised Machine Translation output, which was in turn better than unrevised Machine Translation output. The most surprising result was that Vietnamese student pilots with some knowledge of English found the original English manual the most comprehensible of all. Sinaiko and Clare concluded that working to improve the clarity of the English manuals would be of the greatest benefit to the Vietnamese student pilots, with some extra benefit for English users as well.

[3]The study is summarized in Sinaiko and Klare 1972.

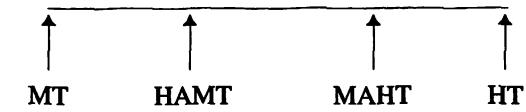

FIGURE 2 The Continuum of Human and Machine Translation

2.2.3 Human Intervention

Almost all translation systems provide some role for a human being. If the machine is in charge, turning to its human collaborator only for occasional assistance, or to make minor amendments to the final result, we speak of *Human-assisted Machine Translation* (HAMT). If the person is in charge, turning to the machine for word processing, data-base access, and the like, we speak of *Machine-assisted Human Translation* (MAHT). Clearly, a variety of possibilities exists between these extremes as depicted in Figure 2. In machine interpretation, as envisaged in the Verbmobil project, the possibilities for human intervention are presumably quite limited, and of a different kind than those that the designers of translation systems have usually contemplated.

Given some division of the labor of translation between human and machine, two questions arise. First, at what point in the process does the human intervene: is it before, during, or after the time when the text is in the hands of the machine? Second, what is the nature of the intervention? Does the human need to have the skills of a translator and is the human actually doing translation? Systems have been proposed in which the human is required to know only one of the languages involved—usually the source—and, in these cases, the human role is presumably far from that of a translator. Specialized systems have been proposed that would interact directly with the author of the document being translated, not just because that is the person best qualified to answer any questions about the intention of the text, but also because that is usually the person with the authority to actually change the text, if that proves to be the most expeditious thing to do.

Until recently, little attention was given to systems and computational tools falling towards the Machine-Assisted Human Translation end of the continuum. We discuss some examples in 2.2.4. The lack of interest in this area is remarkable in view of the fact that it is here that the greatest return of investment, especially in the short term, is presumably to be expected.

As for the time of human intervention, three paradigms have been generally recognized. The first is when there is human *pre-editing* of the source text. The best-known demonstration of the effectiveness of pre-editing is provided by Xerox's system for translating maintenance man-

uals in which a restricted version of English, designed for this particular task, is used. This made it possible for Xerox to achieve usable translations with the Systran Machine Translation system. Such restricted versions of English are sometimes called "Caterpillar English," because Caterpillar was the first firm that experimented with them. The Xerox version is called *Multinational Customized English.* Work on similar uses of restricted language has been going on at Boeing. Recasting the original in a carefully formalized version of a natural language in this way is the extremest form that pre-editing takes. More usually, it entails inserting orthographic disambiguation cues for the machine, such as markings to help with syntactic disambiguation. This is done in a number of Japanese systems, notably the one in use at JICST, where pre-editors also identify proper names that could be otherwise interpreted, and where long sentences are broken up to reduce the chance of cascading translation errors.

The second and third possiblities are post-editing by humans, already discussed at some length, and interaction between man and machine during the translation process. Only quite recently has any significant attention been devoted to this last possibility, strongly advocated in Kay 1980 and Melby 1987b. One can imagine systems that combined the three possibilities in a variety of ways.

The most common configuration of a Machine Translation system is one that requires a bilingual human, who must often also be a translator. The human can interact before, after, or during translation. The humans who interact after translation—post-editors—are generally translators and it is usually assumed that they will have access to the text of the original. This has been essentially the only way to guarantee accuracy of the translation except in cases of severly restricted input. Notice once again that we have nothing but the terminology to assure us that it is the machine that is doing the translation and the person who is editing it.

2.2.3.1 Monolingual Human

The LIDIA System (Large Internationalization of Documents through Interaction with Authors) at the University of Grenoble is an interactive system assuming a monolingual user—the author of the text to be translated—who may modify it in the course of the interaction. The N_{tran} system from UMIST (at the University of Manchester) is an English-Japanese Machine Translation system which appeals to a monolingual collaborator to resolve ambiguities. With interactive systems, more ambitious translation tasks (such as translation of business letters) can be imagined. To the extent that the system reduces re-

liance on experts, it is undoubtedly useful, and the goal is clearly to produce output that requires no editing.

One can go further in this direction. UMIST also has a project funded by British Telecom on a system that guides a monolingual English user through a menu to help them write a business letter in a foreign language. This, in effect, is translation without an original (the system described in Saito and Tomita 1986 is similar), taking the user through a negotiation which ultimately results in a business letter in both languages. For such an approach to be practical, the originator of the document must have very restricted goals (for example, filling out a purchase order). Given such a goal, the menu system can be tuned to cultural differences and applications. For example, it may produce a pre-formatted purchase-order for an English client, and a purchase letter with the correct honorifics for the Japanese clients.

Thus, one axis of intervention concerns how much the machine can be integrated into the composition of the original document. The extreme endpoint of that involvement is to let the machine compose the document itself, which is exactly what the successor to METEO is now doing.

Though obviously desirable in many practical situations, it is clearly a much greater challenge to design a system that could guarantee a high level of accuracy if the only humans it could turn to for assistance knew only the target language. However, the possibility does exist of offering a target-language monolingual a variety of candidate translations, beginning with single word suggestions, then after some initial disambiguation, leading up to phrasal alternatives, or perhaps words first, followed by phrases after some disambiguation. There are a number of disambiguating choices that a human with a little common sense and some knowledge of the domain could make that would be difficult for a Machine Translation system. Such a system would require highly motivated users—for example, someone trying to get the gist of an untranslated article in their field of expertise. In any case, one would be reluctant to describe what was being done as "translation" in a case like this; it is more like an elaborated bilingual dictionary.

2.2.4 Translator's Assistant: Machine-Assisted Human Translation

Finally, we come to systems that interact with a translator during the translation process. Such a "translator's assistant" might or might not be a Machine Translation system, depending on whether it actually tried to construct sentences in the target language, or merely offered information to a human translator who did.

In Melby 1987a some suggestions are made for the kind of work a translator's assistant might do:

> LEVEL I: text editing and a bilingual dictionary, customizable by the user. Access to databases containing translation information or examples.
> LEVEL II: a bilingual concordance; suggested translations; multi-word terms. Outputs can be sent to a shared database of translations.
> LEVEL III. a full-fledged Machine Translation system of the sort discussed in the previous sections, and a translator's assistant offering post-editing tools. One interesting possibility arises from the fact that the syntactic structure of the Machine Translation outputs is known (it had to be in order for a sentence to be generated), so that the editor supplied to the translator could be a syntactic structure-editor. Thus it ought to be possible to highlight NP's, or whole clausal constituents and perform operations on them (make them definite, make them passive, and so on).

2.3 The Historical Perspective

This is not the place to give a detailed history of Machine Translation. This has been done elsewhere, for example in Slocum 1985b, Buchman 1987, Warwick 1987, and Hutchins 1986. However, in order to understand what is proposed in Verbmobil, it is valuable to place it against the background of trends in Machine Translation over the past thirty years. Where the trends have been encouraging, we must try to ensure that Verbmobil continues them. Where they have not, we must ask ourselves what grounds we have for thinking that Verbmobil will be able to break the trend.

We were at considerable pains in the early part of the chapter to show the importance for translation of the fact that language is *situated*, and of how a proper translation almost always depends on context. For the most part, however, linguists have concerned themselves with matters that have little to do with context in this sense, concentrating their attention rather on sentences and the smaller units that make them up. The structure of discourse has been largely ignored by them. Presumably, this has been either because the approaches that have been developed in linguistics are very different from those that would be needed for the study of larger units of language, or because the structure of the larger units are determined by considerations that extend beyond linguistics.

Designers of machine translation systems have sometimes been irri-

tated by linguists for painting an increasingly depressing picture of translation because they continually come up with new forms and sources of vagueness and ambiguity in language, but rarely provide any insight into how people succeed in resolving the problems. But if we are right in suggesting that it is the situated nature of language that gives it its efficiency and adaptability, then it should not be surprising that the problem is apparent from the relatively narrow confines of linguistics, while its solution belongs to a much wider domain. This is not to say that the resolution problem has failed to attract that attention of any linguists and, it is greatly to be hoped that it will attract more within the framework of the Verbmobil program.

With the appearance of Chomsky 1957, linguistics, at least in America, entered a period of much greater formality in the way in which theories were described. Grammars came to be seen not just as ways of describing particular languages but as objects with mathematical properties. Along with this came the realization that grammars that were written according to specific conventions, and whose mathematical properties were known, could be accompanied by standard algorithms enabling them to be used in the analysis and generation of sentences. But these realizations took hold only little by little so that the people who designed the first major Machine Translation systems in the mid and late sixties either did not know the facts or thought them to be of only theoretical interest.

Formal linguistics was also largely ignored by the designers of the big Japanese systems of the eighties, even though formal approaches to the study of language had become quite widespread, by that time. There seem to have been three closely related reasons. First, the tradition of language study, by any kind of objective methods, was not well established in Japanese tradition. Second, linguists were seen as impractical people, more concerned with rare examples of quaint phenomena than with everyday usage. Third, it was the electrical engineers in Japan that took up the challenge of Machine Translation, and they saw it as an engineering problem.

It is probably not too unfair to say that there has never been an attempt to build a major Machine Translation system on the basis of theories current among professional linguists. Furthermore, the small systems that have been designed on such principles, by their very smallness and inadequacy in practical use, serve only to encourage the view that linguists have nothing of value to say on this matter. Systems built on modern principles rarely show any advantage from the user's point of view over the more *ad hoc* systems of the sixties, so that it remains the case today the Systran is by far the most successful product

in this field on the market. The reason for this is not far to seek. The early systems, lacking any sophisticated tools for analysis of any kind, treated every phenomenon as a special case. They therefore relied on huge dictionaries which were, essentially, repositories of these special cases. While more modern methods would undoubtedly make it possible to eliminate large numbers of special cases, some would inevitably remain. The collection of special cases has been underway for so long now, and the inventories of them that have been collected are so large that the systems that have access to them will inevitably appear superior until the more modern systems have been in the field for a long time. The effect is, in fact, an illusion, because the older systems have typically reached the point of diminishing returns so that substantial further improvement is in fact not a possibility for them.

There are those that doubt that what linguists know about such matters as morphology, syntax, and formal semantics is important for Machine Translation, but few persist long in this view. During the eighties, computational concerns exerted an influence on the development of linguistic theory to an unprecendented extent. To the outside observer, the principal result of this influence may seem somewhat paradoxical. It turns out that formalisms that rely heavily on procedural notions, like tranformations and rules that move words from one place in a structure to another generally make for poor computational models. One of several reasons for this is that a procedure that is written from the point of view of the producer of a sentence can only in exceptional cases be recast so that it can be used for analysis. Under the influence of computationalists, formalisms like LFG, GPSG, HPSG and various forms of categorial grammar have come into being which use the non-procedural notion of unification in a way that makes the more cumbersome procedural notions no longer necessary. This bodes well for the treatment of complexity in future Machine Translation systems and gives good reason to suppose that the point of diminishing returns will be reached later. We shall return to these matters shortly.

Both in the eighties in Japan and in the sixties in the West, it was taken largely for granted that Machine Translation systems were built, not as research prototypes, but to be used. They therefore had to contain large dictionaries and sets of grammar rules and to be efficient enough to compete with human translators. This not only meant that they had to be built by engineers, who understand these issues, but also that they could not use the nondeterministic techniques that most naturally accommodated the ambiguity that linguists recognized in language. In any case, it was argued that a Machine Translation system must produce a single rendering of everything in the text to

be translated, so that nondeterminism was inappropriate in any case. But the requirement to produce a single result does not mean that all results are equivalent, or that the need to compare potential results is somehow eliminated. This is a point that we take up again in 4.7.1.1.

Another problem with the large systems of the sixties came from the belief that the large dictionary that would inevitably be at the heart of the system could be built before, or at least, concurrently with, the program itself. There was generally little appreciation of the fact that individual entries, the kind of information in whole classes of entries, the format of the dictionary, and, in fact, just about everything about the dictionary, would have to be subject to change in the light of experience. The designers of the later Japanese systems were usually more circumspect in their approach to this problem, though there are ominous echos of the old folly in the program of the current EDR project.

Machine Translation is a manifestly complex business, and the management of complexity is a major unifying theme in computer science. Modern programming languages and practices, data-base management systems, and notions of software architecture are all developments that have come about in response to the general problem. A more specific response, within computational linguistics, has been the realization that bodies of data, such as dictionaries and grammars should be strictly separated from the programs that work on them. This matter, to which we shall return in 4.7.3, was already recognized by some computational linguists in the early sixties—the Cocke-Kasami-Younger parsing algorithm, after all, dates from 1960—but its importance was far from being generally recognized. Furthermore, most of the algorithms that Machine Translation systems contained were so particular and *ad hoc* that it is often not clear what it would have meant to separate them from the data.

From the point of view of the linguist, there is an independent reason for wishing to keep grammar and lexicon separate from the programs that interpret them, namely that they represent the *competence* of the system, whereas the program, to the extent that it is accorded any theoretical status at all, represents the *performance* component. Accordingly the programs remain constant, in principle at least, when the languages between which the translations are being made are changed.

2.4 Linguistic Issues

2.4.1 Linguistic Levels of Analysis

Every Machine Translation system is involved with linguistic structure. It must therefore incorporate some notion, however impoverished, of what linguistic structure is and how it relates, both abstractly and operationally, to actual texts. To this extent, every Machine Translation system necessarily enshrines some notion of linguistic theory. The principal argument for this rests on what is commonly referred to as the *productivity* of language.

Somehow, people are able to produce and understand sentences they have never heard before. But these utterances contain pieces that they have heard before and they know regular ways of combining these familiar pieces to create novel utterances. In other words, language has structure by virtue of which new examples can be seen as instances of old patterns. This is the basis of linguistic productivity.

The productivity of language seems to depend on various different kinds of structure. Linguists see the object of their study as divided into levels that are organized in largely independent ways. Generally recognized levels are those of phonology, morphology, syntax, and semantics.[4] In this section we discuss the basic components of analysis, and their relation to translation. We discuss the relationship of semantic and syntactic analysis, but we defer discussion of semantic representations until 2.5.2.5.

To each level of linguistic organization there corresponds a module, or major section of a *grammar*, which is a formal system describing the language. For these purposes, each module of a grammar is often written in a language specially designed for modules of that kind. But too much attention is often paid to these special languages in the mistaken belief that they invest the enterprise as a whole with scientific rigor. In 4.7.3 we urge the view that, while modularity is indeed crucial, special purpose languages are not. In what follows, we therefore use the term "formalism" to refer to a formal system with known properties and well defined interfaces to other systems with which it stands in an input/output relation, leaving aside the question of whether a special language is also involved.

[4]Sometimes pragmatics is added into this list. This is a debatable move. On the one hand pragmatics is unquestionably an important part of interpretation. On the other, not all the things that go under the heading of pragmatics—such as conversational implicature as defined in Grice 1975—seem to be particularly tied to linguistic forms. As Grice himself argues, they are perhaps better thought of as part of a general theory of communication and rational cooperation.

2.4.2 Morphology and Phonology

Morphology, or the study of how words are composed out of other words or smaller pieces, is generally seen as falling into three parts:

1. Inflectional morphology, most closely associated with grammar in the traditional sense. Roughly speaking, it is concerned with the different form that a word must take according the grammatical context that it occurs in, and with the way a word manifests properties that every member of its kind must show. The case of a noun is an example of both of these: subjects and objects require different cases, and every noun must have some case. Tense is an example of the second kind: every finite verb must have a tense, but the tense it has has semantic, but no grammatical consequences.
2. Derivational morphology which derives new words from old ones often of a different category. For example, "ness" can be added to the adjective "happy" to yield the noun "happiness."
3. Compounding[5] of two or more independent words to produce a new word, such as English "fireman" and "wayfarer," and German "Lebensversicherungsgesellschaftsangestellter" (life insurance company employee).

The importance of the morphological component is that it is often the only bridge between what a system user speaks or types in, and the system dictionary. Wordforms like "walking, happiness," and, more strikingly, "undisenfranchizability" and "Lebensversicherungsgesellschaftsangestellter" cannot be expected to occur in a dictionary. Their syntactic properties and meanings must be deduced through the use of the morphological rules of the language.

Many systems have rather rudimentary morphological processors, which do little more than strip off endings or break words into subwords from left to right, and then assign syntactic features. There are several reasons for preferring a more thoroughgoing approach to the problem:

1. Morphological analyses can be ambiguous: consider the word "unionizable," which can be derived either from "union" or from "ion." Some English words contain structural ambiguities. So consider what an "untiable knot" would be. Depending on the structure one ascribes to the word, it is presumably either an knot

[5]Compounding does not invariably result in a single *written word.* Consider the English sequence "black bird." With the stress on "bird," it is simply a phrase referring to a bird that is black. With the stress on "black," it is a compound, referring to a particular species of bird.

than cannot be tied, or a knot than can be untied. In a tutorial on Machine Translation Harold Somers gave the following examples from German: "Alleinvernehmen" as either "lone perception" or "global agreement"; "Arbeiterinformation" as either "formation of female workers" or "worker information." This means at least that the process must be nondeterministic.

2. Many languages (Finnish is a celebrated example) have very complex morphological systems, which can only be treated coherently with more systematic formal tools.
3. Speech analysis systems, like Verbmobil, need a morphological analyser that can not only accept correct analyses but also reject incorrect ones, since the speech recognizer will doubtless pass along a host of spurious tries; function words and affixes are likely to suffer the heaviest casualties during speech recognition, and the morphological component will act as an important filter in the overall system.
4. The parts that go to make up a word change in accordance with the phonological grammar of the language. Thus, the basic process of plural formation for English nouns is to add a sibilant, but the details differ according to the sound at the end of the noun. Compare "cats," "dogs," and "horses." In speech, the matter is a great deal worse, especially in languages like English and Russian, where the addition of affixes can affect where the stress falls in a word, and a change in the stress generally results in a major change in vowel quality. Consider the words "telephone," "telephony," and "telephonic."

There have been very significant advances in computational morphology and phonology in recent years, notably in so called *finite-state* morphology. This work has shown that it is possible to process the morphological part of a grammar in advance so as to put it in a form that allows remarkably efficient analysis and generation of utterances by extremely simple programs. These techniques constitute an excellent example of the value of *compilation*, or the automatic restatement of information in a relatively inexpensive step that is carried out once in advance of the main computation, the efficiency of which is thereby greatly enhanced.

2.4.3 Syntax

2.4.3.1 Phrase Structure and Dependency

Where morphology deals with how words are combined to form other words, syntax deals with how words are combined to form phrases and

sentences. The notion of a phrase is based on a very robust intuition. Consider the following examples:

(44) a. the red book
b. the red book that John sold
c. while eating his lunch one day
d. the very

(44a) is a case of a Noun Phrase. It is a Noun Phrase because the most important word in it, the word that has the most to do with determining what kind of contexts this phrase can occur in, is a Noun. The idea that all phrases—or at least most phrases—have somewhere inside them a word which mostly determines the properties of the phrase plays a role in all current syntactic theories. That word is called the head of the phrase. (44b) is also a Noun Phrase, although a more complex one; it can occur in just about all the places (44a) can. (44c) is a more complicated example of a phrase, complicated enough so that linguists differ in their claims about what the head word is, and whether it has one in the same sense as (44a); many linguists would call (44c) a subordinate clause. Few linguists would call (44d) a phrase. Although it can clearly serve as part of a phrase. Most syntacticians view phrases as being built up around their heads. Thus, if there are sub-phrases to (44a), they must be as in (45a), and not (45b):

(45) a. [the [red book]]
b. [[the red] book]

Just as important as analyzing what the pieces of phrases are is deciding what they cannot be. Thus, English is strict about requiring most adjectives to occur before the noun:

(46) * the book red

This requirement, however, does not extend to phrases which consist of adjectives followed by another phrase:

(47) a man deep in his dotage

Thus an adequate syntactic description must sort out various complex and sometimes arbitrary constraints that the language imposes. Given the kinds of examples above, the following sorts of goals emerge:

1. what the phrases of the language are
2. what the subparts of every phrase are
3. what the heads of phrases are
4. what the possible orders are for the subparts of phrases

This is a very minimal agenda for a syntactic theory which places the notion of a phrase (or *constituent*) at the center.

The family of theories known as Transformational Grammar makes a crucial addition to the above agenda argued for in Chomsky 1957 and still central even in the more modern theory of Chomsky 1981. According to this view, systematic relations between different kinds of phrase must also be accounted for. Thus, if every passive sentence has a corresponding active sentence, then the syntactic description of the language must somewhere capture this. The way in which Chomsky proposed to capture the relation was to claim that syntactic structure actually existed at a number of levels (or "strata") simultaneously; the most abstract stratum (the stratum which potentially differed the most from the surface stratum) was called *deep structure.* In earlier versions of Chomsky's theory, active and passive sentences were related by sharing a deep structure. We will call theories which posit multiple levels of syntactic description *multistratal.* Most, but not all, of the approaches which have historically descended from Chomsky's work are multistratal.

Not all syntactic theories give the same importance to the notion of a phrase. In fact, this is the source of a theoretical division that has played an important role in the history of Machine Translation. There is a grammatical tradition known as Dependency Grammar, stemming from the work of Tesnière (see Tesnière 1959), which gives the notion of head an even more central role in syntax. According to this view, the job of syntax is not so much to group words into phrases, as to establish direct relationships among the words themselves. The structure of a sentence has the form of a tree with words at the nodes and with, say, the finite verb, at the root. Of course, it is possible to induce a set of embedded phrases from such a structure because a word and the other words that fall below it in the tree can be thought of as a phrase. The diagrams in 3 show corresponding phrase-structure and dependency diagrams for the sentence "the boy gave the girl an apple."

In dependency grammar, the word under which a set of other words is collected—the counterpart of the head in many phrase-structure grammars—is called the *governor* of those other words, and they are called its *dependents.* In 3, we have followed the usual practice of showing dependents beneath their governors. We have also shown the dependents to the left or right of their governors, according to whether they come before or after them in the string. This practice is not followed by some dependency grammarians, especially those who wish to emphasize that their structures abstract away from the linear order of the string. Sometimes, different dependency reltions are recognized and the lines in the diagrams are given labels like *subject, object,* and *modifier.*

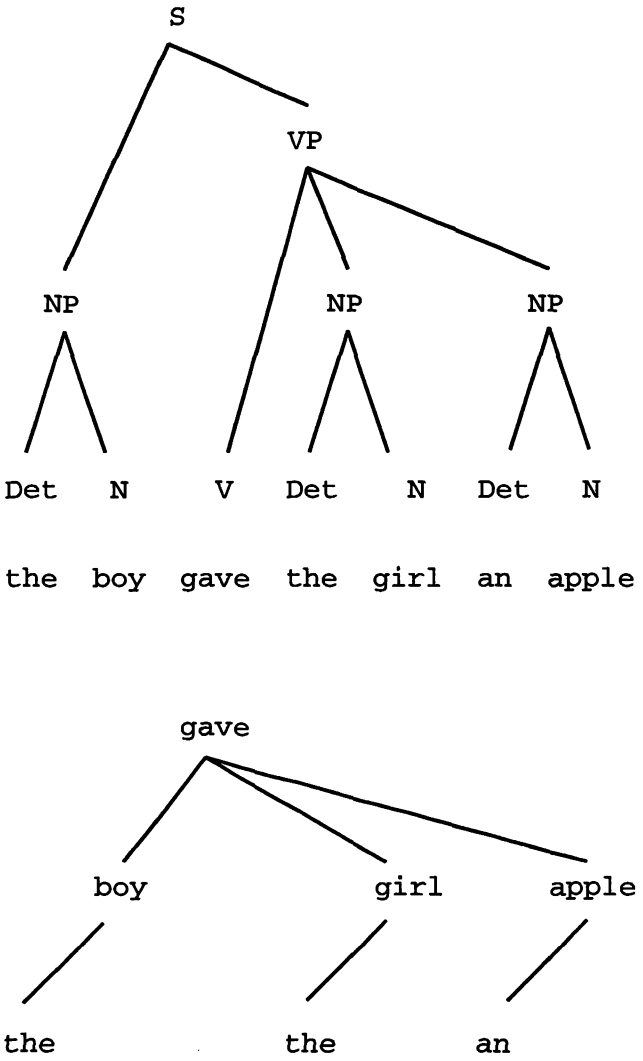

FIGURE 3 Context-free and Dependency Structures

It is no accident that dependency grammar has appealed most to students of languages with relatively free word order where the notion of a phrase appears less important because a head and its dependents may not always appear in the same order or even adjacent to one another. Thus direct and indirect objects in German may either precede or follow their head verb. Or in Polish, we find:

(48) Ciekawą przeczytalem książkę.
Interesting (accusative) read(1st-person) book (accusative)
I read an interesting book.

Here the adjective *interesting* appears separated from the head *book* on which it depends, and from which it acquires its accusative marking. Followers of Tesnière criticized early versions of Chomsky's theory because the notion of constituent was so central and because the notion of head was completely absent.

Dependency grammar has played an important role in the history of Machine Translation, largely under the influence of the GETA group at the University of Grenoble, which set the direction in this as in so many respects. Even systems that analyzed sentences in phrase-structure terms often went on to translate those phrase structures to dependency trees which they then used in the transfer phase. This was primarily because it was thought to be easier to treat dependency trees independently of the order of the words. If the relations in the dependency tree were labeled, the original order could be ignored and an entirely new one assigned in the course of generating the target text. Alternatively, the new order could be developed incrementally from the source text by a series of relatively simple steps performed on the dependency tree. It should be said that the advantages claimed for dependency structures for these purposes may be more apparent than real.

The term *case grammar* (Fillmore 1968) refers to a formalism which, though part of the phrase-structure tradition, has a strong family resemblance to dependency grammar and has also been influential in Machine Translation, especially in Japan. The similarity resides in the fact that the head of each phrase contracts a specific, named relation with the other members of the phrase. The name comes from the obvious analogy between these relations and the cases of a language like Latin or German. The appeal of the idea in Japan is easy to understand, given the fact that the order of noun phrases in a sentence is almost completely free, their roles being given by particles following the phrases which function in a similar manner to cases. The MU project, the first of the new wave of Japanese Machine Translation projects,

which began in 1982 (Nagao et al. 1985), was heavily influenced by the work of the GETA group in Grenoble but used a version of Case Grammar (with different overlapping sets of case roles for Japanese and English) to capture dependency relations. Since MU, a number of other Japanese Machine Translation systems have also incorporated case representations.

Another early approach that rests heavily on the notion of a dependency relation is relational grammar (see Perlmutter 1983). This theory is similar to Tesnière's in using syntactic dependency relations, but closer to other generative theories in being overtly multistratal and recognizing a number of levels at which dependency relations need to be captured. Relational grammar has had little influence in Machine Translation until recently. Currently IBM Japan is experimenting with this approach.

2.4.3.2 Procedural and Declarative Grammars

Transformational grammar, at least in the minds of its early advocates, constituted a strong reaction to the *structuralism* that had dominated American linguistics from the beginning. Earlier linguists had been impressed by the great variety of the native languages of America, and had taken it as their task to provide as many as possible of them with a grammar and a lexicon. In other words, they set about *describing* the languages and were thus *structural* linguists. Chomsky saw himself as engaged in a more scientific enterprise. For him, it was interesting to describe a language only if one did so in a way that was revealing for the nature of the human linguistic faculty as a whole. It was not enough that it should be *descriptively adequate*, it must be *explanatorily adequate* also. Essentially all linguists have come to espouse this view in some form or other.

The modern descendants of transformational grammar are called "Government and Binding" (GB) theory. In the world as a whole, these theories have by far the greatest number of adherents. For linguists that do not adhere to this tradition, it is probably fair to say that principal problem that motivates them is the question of how language works as a system of communication. How is it that, by making appropriate noises in one another's presence, we can succeed in causing approximately the ideas we intend to form in the mind of another? Linguists who are motivated by these considerations are sometimes referred to as functionalists because, for them, the important questions concern how language functions, and explanatory adequacy is seen in these terms.

Chomsky and his followers in the GB tradition are not functional-

ists. In fact, they appear to have little interest in language as a means of communication. For them, the motivating questions concern how children acquire their knowledge of language so effortlessly, with essentially no deliberate instruction, and with virtually no examples of unacceptable language usage. Whether this is an accurate characterization of the challenge that children face is, of course, not uncontroversial. Nevertheless, the adherents to this theory see the challenge as so daunting as to leave no alternative to the view that humans are born with with a capacity for learning language and, furthermore, that the kinds of language they are equipped to learn are of very specific kinds. The current conjecture is that, up to the choice of items in the lexicon, there are only finitely many languages. The great similarity between languages comes from the fact that the innate capacity for learning them operates in accordance with a number of very specific *principles*, and that the differences are accounted for by the different settings of a small number of *parameters*. Since the parameters are the only things, apart from the lexicon, that distinguish one language from another, there is no longer any need for language-specific grammar rules.

GB is a multi-stratal syntactic theory. A given phrase can occupy different positions in the syntactic trees that occupy each of the different strata. The possible relations among these positions are constrained by the principles of the theory and the parameters of the language. If a word or phrase appears in different places in the structures on adjacent strata, it is said to have *moved* by a device konwn as "move-α." The conditions that cause move-α to take effect are of the following general kind: If the theory says that a noun must have a case assigned to it, but it is not at a place in the structure where that can happen, then it must move to one where it can. On the other hand, if it receives case in its current position, but must move elsewhere to fill some other requirement, then it takes its case with it.

In 2.3, we remarked that theories that make explicit appeal to procedural notions tend to have undesirable properties for anyone interested in implementing them as computer programs. Adherents of GB are wont to claim that there is nothing inherently procedural about the theory; it is just that the commonest metaphor for describing it is a procedural one. The fact remains that computationally the attempt to find the nonprocedural reality behind the metaphor or an alternative metaphor without these procedural properties has so far not been completely successful.

One of the early arguments in favor of transformational grammar, the predecessor of GB, was that context-free phrase-structural grammar (CF), the best known formal theory at the time, was not powerful

enough to describe natural languages in the simple sense of not being able to distinguish the sentences from the non-sentences. Only considerably later (Pullum and Gazdar 1982) were the flaws in the argument clearly shown. First, when carefully examined, the argument turned out to be that certain linguistic phenomena would require large context-free grammars to accomodate them. Secondly, and more importantly, it was shown that the arguments turned on the particular notation that was being used for CF grammar and that a different notation put the old questions in an altogether new light. In particular, what would have been a very large grammar in the old notation could easily be of quite modest size in the new one. These observations laid the foundation for the development of GPSG (for an introduction, see Sells 1985, and for a rigorous presentation Gazdar et al. 1985). The program was to show that a CF grammar could account for the syntactic facts of natural languages and, furthermore, that it could do so in an explanatorily adequate manner. Gazdar and his colleagues presented illuminating analyses in a large number of cases.

One important device for capturing generalizations in the theory was the notion of a *metarule,* a rule for generating rules. Another consisted in the use of *features* with conventions for their propagation through the structure of a sentence. The treatment of features came to be a major issue in grammatical formalisms, an issue on which theoretical linguists, for the first time, gained substantial insight as a result of work on the computational properties of the grammars they wrote. The issue turns on the notion of *monotonicity* and of a particular operation which has this property, namely *unification.* These are the notions that made possible the construction of explanatory grammars, with appropriate power, but with none of the procedurality of GB and transformational grammar.

Informally, unification can be though of as an operation in which a pair of *descriptions* are compared with a view to determing if they might be of the same object. The proposition that the descriptions are of the same object is rejected only if there is some respect in which they are incompatible, one feature for which they have conflicting values. One says that the person's hair was blond; the other that it was black. One says that the noun is singular; the other that it is plural. If there is no incompatibility, the process yields a new description containing the union of the information in the two original ones. Continued use of the unification operation therefore tends to produce larger, and therefore increasingly more specific descriptions, thus progressively narrowing the class of objects they might refer to.

Unification is a *monotonic* operation because information that is

added to a description as a result of unification never invalidates any information that was in it before. Put another way, the objects in the class that the most specific description covers were in the set covered even by the least specific of the contributing descriptions at the outset. Somewhat more formally, a description d_1 *subsumes* a description d_2 if any object falling under d_2 also falls under d_1. We also say d_1 is more general than d_2. The result of the unification of n descriptions is the most general description which is subsumed by all n descriptions. A monotonic operation is one in which the result is subsumed by the arguments. Unification is a commutative and associative operation, so that the order in which a set of unifications leading to a given final description is carried out is immaterial. A system that is based on operations with this property is essentially declarative because the notion of procedurality turns on the ordering of the operations being important.

The first use of these notions in natural language processing was in Kay 1984. The construction of F-Structures in LFG can also be understood in terms of unification. Unification arises when a rule or principle requires that a single linguistic object be described simultaneously by two or more structures. If the descriptions are compatible unification succeeds. If they are not, unification fails, and the linguistic object is not admitted by the grammar. For a good introduction to unification, see Shieber 1986.

In its later versions, GPSG appealed explicitly to unification for the operations it performed on features. Several later theories have also depended on it very strongly, though not necessarily without having to appeal from time to time to other operations. HPSG, for example, also makes use of string combination operations and set-union in its rules. Set-theoretic operations are also a feature of LFG. Fortunately, it is also possible to interpret these operations monotonically.

Head-Driven Phrase-Structure Grammar (HPSG) (see Pollard and Sag 1987) is a direct descendant of GPSG. HPSG dispenses with the notion of a metarule, doing the same work with a combination of general principles, lexical rules (rules for producing new lexical items), and syntactic rules. It relies heavily on *subcategorization lists* associated with lexical items to determine what other items can appear in the phrases of which they are the heads and, to this extent, it has a resemblance to categorial grammar, briefly discussed below. Unification is well suited to a monostratal theory, because it allows for the integration of information of very different kinds into a single representation. In HPSG, for example, the distinction between universal principle, rule, and lex-

ical item is not made formally because all are formulated in structures subject to unification.

Lexical Functional Grammar (LFG) (see Bresnan and Kaplan 1982) is one of the earliest and most influential of the unification-based formalisms. It appeals to theoretical linguists and computationalists alike, largely because of its adherence to the notion of monotonicity. LFG assigns a number of different structures to a sentence related to one another through *projections*. The earliest versions of the theory recognized two representations. The first, called *C-structure* is a phrase, or *constituent* structure assigned by a component of the grammar that is essentially a CF grammar. The second, F-structure, is a feature-structure such as is also found in HPSG and several other members of this related set of theories. Later versions of LFG also contain other kinds of structure, notably a semantic representation. LFG has aroused considerable interest in the computational linguistic community at large, and there are a number of computational implementations. It has proved attractive as a formalism for use in Machine Translation systems, especially in Europe, probably because its mathematical and computational properties are well understood and it has changed relatively little over time. A novel approach to Machine Translation in this framework which largely eliminates the distinction between the transfer and interlingua approach is described in Kaplan et al. 1989.

Categorial grammar, first proposed by Ajdukiewicz, and first seriously recommended to linguists by the philosopher Peter Geach, carries to the extreme the notion of a grammar based in the lexicon. In categorial grammar, essentially everything is in the lexicon. It was probably the work of Montague that most clearly showed its potential for linguistic description. Ajdukiewicz's simple idea was that in the syntax, categories should be named for the categories they are seeking to combine with, and the categories that result; on the semantic side, the "seekers" are functors, the "sought" are their arguments. Results are the functor ranges. Thus, an intransitive verb would have the category S/NP, because it combines with an NP to produce an S, and semantically, it is a function which given an NP argument gives a proposition. Note how the idea of head has been transformed into that of a functor. Decomposing the atomic categories of this kind of grammar into feature structures and putting the unification operation in the place of the equality relation increases the expressive power of categorial grammar in a very natural way to give Categorial Unification Grammar (Uszkoreit 1986, the very closely related Unification Categorial Grammar (Zeevat et al. 1987), and Trace Unification Grammar (Block 1991).

Along with explorations in unification-based linguistic theories, there has been some exploration of unification-based formalisms which make no theoretical claims but provide frameworks within which theories might be cast and computations performed. Some examples are

1. FUG (Kay 1984) was the first of the unification-based formalisms
2. PATR is a grammar formalism providing facilities for context-free skeleton syntax annotated with unification equations. The structures that serve as linguistic descriptions are Directed Acyclic Graphs.
3. CLE (Core Language Engine), developed at SRI, Cambridge, is a descendant of PATR with some variations. For example, CLE restricts itself to term unification, unification with structures that have a fixed number of arguments, and CLE allows variables over sequences of categories. CLE is described in Alshawi et al. 1988.

All of these systems present themselves as formalisms that can serve for any of a number of linguistic accounts. No claims about particular linguistic analyses are part of the core package, although particular analyses are embodied in existing implentations. In particular, there is an implementation of PATR at SRI and an implementation of CLE at SRI Cambridge. Although there are various reasons why implementers of a large unification-based grammar might want to diverge from these approaches, they have addressed some of the basic issues in implementing a unification-based formalism.

Within the domain of Machine Translation there are several systems where unification has played a role.

1. Using CLE, SRI Cambridge has implemented a semantic transfer system translating between Swedish and English, described in Alshawi et al. 1988.
2. The CRITTER system described in Isabelle et al. 1988 uses a Prolog Dependency Clause Grammar, and thus exploits Prolog unification. Like the CLE system, the Machine Translation component of CRITTER uses semantic transfer.
3. MIMO-2 is a Machine Translation system at the University of Utrecht which is influenced by PATR and HPSG. It is described in Van Noord 1990a and Van Noord 1990b.
4. ELU (Environnement Linguistique d'Unification) is a unification-based formalism based on PATR and used at ISSCO for experiments in analysis, generation, and translation. An implementation of a system that translates skiing reports from German into French is described in Bouillon and Boesefeldt 1991.

Reliance on operations like unification makes for monotonicty and monotonicity makes for a relatively easy separation of data from algorithm; competence from performance. Monotonicity also makes for reversibility, that is, the possibility of using the same data, and sometimes even the same algorithms, in analysis and generation. Since all Machine Translation systems must be able to generate as well as analyze, the issue is at least an interesting one in that context. In theories which make it explicit, the transduction that a grammar effects between strings of words and some other structure is recursively enumerable, that is, effectively computable. But there is no guarantee that the reverse mapping will be. And even if it is, it may be an inherently more complex problem computationally than the analysis problem. Reversibility, then, is a property of grammars. But even when the formal property of reversibility exists, there are computational questions about the properties of the generation problem which have only recently begun to attract attention.

For a Machine Translation system grammar reversibility is obviously a desirable property, both on practical and methodological grounds. The practical appeal is that grammar writing is a difficult and complex task. It is thus desirable to build only one grammar for both analysis and generation in each language. The methodological argument is that this is linguistically correct. It is incoherent to argue that there are strings which ought to be analyzed but not generated, or vice versa.

This does not preclude a generation algorithm that does not generate all sentences of the language: this is because the generator may make deterministic choices not forced by the grammar (for example, always generating actives, never passives), analogous to the way in which a parser may be forced to make deterministic choices with respect to a grammar by enforcing agenda regimes that force certain kinds of syntactic attachment.

We said that all Machine Translation systems must generate, but this does not mean that generation has to be treated as the inverse of parsing or that the system has to contain a generation grammar. The design of systems that do not make use of such a grammar—and there have been several—rests on the assumption that the structure delivered to the generation part of the system will always be a well formed "deep structure," in the appropriate sense, in the target language. The trouble with this assumption is that it is strictly impossible to verify it.

In choosing a linguistic theory for Machine Translation applications the following criteria seem to be important:

1. whether the theory underlies a substantial amount of current research, so that there is a pool of expertise to access.
2. whether the theory has been implemented computationally, particularly with a grammar rich enough so that there has been some confrontation with the problems of complex rule interactions that plague many systems.
3. whether the theory is interpretable in a formalism that is commutative, associative, and monotonic and what effect nonmonotonic devices have on computation.
4. whether the theory is reversible.
5. whether the formalism is based on unification. There is at least a presumption in favor of a unification-based formalism, because the computational properties of unification are well-understood and unification-based formalisms encourage clear grammar-writing.
6. whether an interlingua or semantic transfer approach to Machine Translation is chosen, in which case a theory which already provides an approach to semantic interpretation has a built-in advantage.

2.5 Translation Strategy

In this section we try to present some of the basic differences in approach among Machine Translation systems. A property that all Machine Translation systems clearly share is that of mapping source language representations to a target language representations. Several questions about the mapping arise:

1. whether any linguistic analysis is attempted before the mapping;
2. how the linguistic analysis (if any) is integrated with the translation mapping;
3. what kinds of information are used to effect the mapping: linguistic, statistical or analogical, or facts about the context and the subject matter of the source.

We will argue in this section that the translation mapping in general does need to take into account linguistic information, but that it also requires nonlinguistic information. In 2.5.1, we look at three different proposals for meeting the need for non-linguistic information. In 2.5.2, we turn to systems that assume some degree of linguistic analysis, and deal with the issue of a transfer versus an interlingua approach.

2.5.1 Nonlinguistic Information

We showed in 2.1 that the underlying difficulty of Machine Translation is the situated nature of language—that sentences in context do much work that is not predictable from the forms of the sentences alone. This means that linguistic analysis can only take us part of the way towards the solution to the problem of translation.

We will look at three different proposals for meeting the need for nonlinguistic information in a translation system, which we will refer to by the following terms:

1. Analogical
2. Knowledge-Based
3. Connectionist

Analogical approaches are translations made entirely on the basis of other translations that have been made previously, either automatically or by traditional methods. Pure analogical approaches use none of the grammars, dictionaries, and the like that we have been discussing up to now. The knowledge-based approaches are those in which knowledge of the world rather than knowledge of language plays the most important role, and connectionist approaches are based on neural nets.

2.5.1.1 Analogical Approaches

Given the difficulty of formulating a theory of translation, but the relative ease of identifying examples of it, there is some appeal to an approach to novel translation tasks that is based on a corpus of known examples. In this section, we discuss two different kinds of models, first a statistical model, in which the target is found by a search for the sentence which is the most likely translation of the source, then an analogy model, in which a database of examples is used to produce new translations by analogy.

The Stochastic Approach. The statistical approach to machine translation is really as old as the field itself. It is generally believed (for instance, in Hutchins 1986) that Weaver's 1949 memorandum was crucial in arousing early interest in (as well as early scepticism about) Machine Translation. Weaver's optimism that computers could be useful in translation was based on his experience in cryptography. The memorandum recounts the anecdote of a cryptographer who deciphered a short 100 word coded text in Turkish without knowing Turkish or even that the text was in Turkish, using purely statistical analytical methods.

Weaver's memorandum recommended research in language invariants, assuming that they would be logical and statistical in nature. His

confidence that statistical properties would ultimately prove a key to translation studies can be traced to his work in popularizing Shannon's ideas on information theory, which defined information itself probabilistically ("more unlikely" equals "more informative") and provided some of the mathematical foundation for signal theory and cryptanalysis.

Although early work in Machine Translation did not make use of statistical methods, some more recent work at IBM has led to interesting results in this area. The description that follows is based largely on Brown et al. 1989.

The basic idea behind the statistical approach is to regard the occurrence of the target as *conditioned* by the the occurrence of the source; the relevant probabilistic law is Bayes's theorem:
(49)

$$\Pr(S \mid T) = \frac{\Pr(S)\Pr(T \mid S)}{\Pr(T)}$$

Read Pr(S | T) as the probability of S (the source) given T (the target). Here the problem of translation is viewed as follows: given a sentence T in the target language, find the S in the source language which is most likely to have produced T. Thus, the task is to maximize Pr(S| T), which amounts to maximizing Pr(S) Pr(T| S).[6]

Thus, the statistical translation method needs three things: (a) a model of the source language called the *language model* for computing Pr(S); (b) a *translation model* for computing Pr(T| S); (c) a *search algorithm* that searches through the source language for a sentence that maximizes Pr(S) Pr(T| S).

The language model used for Brown et al. 1989 is a bigram model. Probabilities of words in sentences are determined using only the previous word as a conditioning factor. Trigram models (using the two previous words as context) have proven quite useful in speech recognition. Models using much bigger contexts are much more difficult to compute and use. The bigram model used is constructed from the English part of the Canadian Hansard corpus,[7] which also provides the data for constructing the translation model. Constructing the model in effect means that for each pair of words $\langle w_1, w_2 \rangle$ in the corpus vocabulary, $w_1 \mid w_2$ is computed.

Constructing the translation model requires a great deal of pro-

[6]We follow the practice of the IBM group of regarding translation as *having* a target and searching for a source.

[7]the Canadian Hansard is the proceedings of the Canadian parliament, always published in both French and English; it constitues a unique source of data for work of this kind.

cessing of the bilingual corpus. We assume in what follows that French is the source language and English the target. First sentences in the corpus must be *aligned* so that there is a link between each sentence in English and its French translation sentence or sentences. Next, each word in each English sentence must be aligned with the French word or words which its translates. The possibility that it corresponds to nothing in the French sentence is also allowed for. From this, several kinds of information for words are computed: (a) conditional probabilities for single French target words given single English source words, for example the probability of "battu" given "beat" (written $\Pr(battu|beat)$); (b) for each English word w and each n from 1 to 25, the probability that w will be aligned with n words in a translation of an English sentence with w; (c) for each i, j and l from 1 to 25, computation of a distortion probability $\Pr(i \mid j, l)$, which is the probability that a word at target position i corresponds to a word at source position j in a sentence of length l. One might guess that the probability should rise the closer the correspondence of source and target position; but this is computed from the data rather than stipulated.

For an entire sentence S which is n words long and aligned sentence T, Pr(S|T) is roughly given by

(50)

$$\prod_{i=1}^{n} \Pr(\text{fertility} = \text{j} \mid \text{e}_i)\Pr(\text{fr}_1 \mid \text{e}_i)\ldots\Pr(\text{fr}_j \mid \text{e}_i)$$

Here $\Pr(\text{fertility} = j \mid e_i)$ should be read as the probability that the i-th word of the English sentence is aligned with j French words (where j may be zero). For each word e_i, the probability that it is aligned with each of the j words it is aligned with is multiplied by the probability of its being aligned with any j words. The probability for the alignment of sentences S and T is just the product of all the e_i probabilities, with the distortion probabilities factored in.

Very little is said about the search algorithm in Brown et al. 1989. The search used is suboptimal; that is, it is not guaranteed to turn up an optimal translation. Search begins with partial translations of the French, extending them word by word until there is a complete translation which is "significantly more promising" than any of the current partial or complete candidates.

The details of the search algorithm are not really that important. One of the most interesting features of the IBM approach is that once the translation and language models are built, *any* search algorithm that produces a list of candidate translations will do. The translation

model will then assign every member of the list a probability. Of course finding a list of possible translations is just what the algorithms of many conventional Machine Translation systems do. The problem many of them have is choosing the right one from that list. Thus, there is a genuine possibility of integrating the statistical-based approach here with other approaches.

There also seems to be some hope of improving the translation model by building in some linguistic sophistication. The model of Brown et al. 1989 does not allow several source words to be aligned with a single target word; in such instances the best it can do is choose one source word for alignment, and align the others with nothing, which then frees the left-out words to show up as noise. For example, the correlation of "to go" with "aller" is captured by aligning "go" with "aller," and "to" with the empty string. This will wind up assigning a fairly high probability of correlating "to" with the empty string, when in fact, if that alignment makes sense at all, it only makes sense in the presence of an infinitive verb form. Similarly, the correlations established by the translation were discovered without any morphological analysis, so that "va" and "vais" in French were treated as distinct unrelated words. There appears to be a large number of cases where some knowledge of linguistic structure could serve to better identify meaningful correlations.

We shall return to the stochastic approach after considering some alternatives.

Example-Based Approach. Researchers in Japan have been exploring a different strategy from the statistical approach pursued at IBM; it shares with the statistical approach two properties: (1) it is corpus-based, the corpus consisting of aligned translated texts; and (2) it resorts to some fairly complex computation to handle sentences that do not actually occur in the corpus. But the ways in which corpuses are used differ considerably. The Japanese approach is called *example-based* or *memory-based* translation. The idea first appeared in Nagao 1984; subsequent work developing the approach can be found in Sumita and Tsutsumi 1988, Sato and Nagao 1989, Sadler 1989a, and Sadler 1989b.[8]

Example-based translation works roughly as follows. There is a translation database consisting of paired English and Japanese dependency trees, together with correspondence links. In the case where an exact match is found, the correspondence links play no role, but in the

[8]Research on example-driven translation is being pursued at the University of Kyoto, ATR, and IBM Japan.

case where the input sentence only partially matches a found example the correspondence links tell us what part of the input we do not yet know how to translate; call that part the "novelty." For example, suppose the translation database contains the following two English sentences and their translations into Japanese:

(51) a. He buys a dog.
b. I found a pen.

Now suppose the input sentence is:

(52) He buys a pen.

What should be done is to replace the translation of "a dog" in (51a) with the translation of "a pen " for (51b). To do that we need the correspondence links between the relevant Noun Phrases in the English versions of (51a) and (51b) and their Japanese translations.

The chief research problem for example-based translation is the definition of appropriate similarity metrics. Two kinds of similarity arise: (1) similarity of elements that do not match in an input and an example (similarity of "a pen" in (52) and "a dog" in (51a)); in Sato and Nagao 1990, this is just stipulated as a lexical property; in Sumita and Iida 1991, this is computed from the topology of a concept hierarchy in a thesaurus; and (2) similarity of the environment of a translated element in an example to the environment of the same element in the input (similarity of the environment of "a pen" in (52) to the environment of "a pen" in (51b)). In Sato and Nagao 1990, because of the definition of environment, this procedure would essentially compare "found" to "buys." Issue (2) is important when the same element occurs in different contexts in the database and is paired with different translations (as might happen with different occurences of the ambiguous English word "pen"). In general, scoring similarities involves scoring similarity between two trees. This can get quite complex; differences between trees may involve distinct syntactic structures, or different words of any category, or combinations of the two. Typically, syntactic and semantic attributes are appealed to in calculating the measure. In Sumita and Iida 1991, attributes are themselves weighted based on their ability to affect translation possibilities.

A basic principle that is appealed to is that bigger matches are better. The larger the example into which substitutions are made to match the input, the better. A further complication is that similarity is ranked not just on source (or input) trees but on target trees as well: thus, in evaluating how good an example (51b) was in supplying a translation of "pen" for input (52) in template (51a), we would measure

the similarity of the translation of (51b) (minus the translation for "a pen") to the translation of (51a) (minus the translation of "a dog").

As a very simple illustration of the difficulties of matching, suppose our input was still (52) but that our example database contained

(53) a. He buys a dog.
b. He found a pen.

Now two possibilities arise: (1) replace the translation of "a dog" for (53a) with the translation of "a pen " for (53b); or (2) replace the translation of "found" for (53b) with the translation of "buys " for (53a). It is not entirely obvious which of these alternatives will yield the best general results.

An interesting application of the example-based approach is given in Sumita and Iida 1991. There the difficult problem of translating the Japanese "no" construction is attacked. The "no" construction is an widely applicable noun modification construction which sometimes translates as an English possessive, sometimes as noun plus prepositional phrase, sometimes as an adjective plus noun. Some examples from the conference-registration corpus used in Sumita and Iida 1991 follow:

Japanese	English
youka no gogo	the afternoon of the 8th
kaigi no sankaryou	the application fee for the conference
mittsu no hoteru	three hotels
toukyou no kaigi	the conference in Tokyo

If the above set serves as our example database, then it provides a good basis for handling other examples with dates, fees, numbers, and cities. This is exactly the kind of case where examples provide the right kind of information. It also seems to be a case where semantic analysis will not help very much. The meaning of the Japanese particle "no" is similar to, and certainly no more specific than the English verb "have."

An appealing feature of the example-based approach is that translations come with scores (measures of similarity to known examples) which can be regarded as confidence ratings. This could be useful to a post-editor (especially a non-expert). In the best case, the post-editor gets a translation rated 1 (the input is identical with a known example). Presumably this gives some reason for confidence (but see our discussion of nonlocal ambiguity below).

The cost of building the system is to populate a translation database by hand. Once sentences in a bilingual corpus have been aligned (probably a task doable by machine—see Denes and Mathews 1960, Brown

et al. 1991 and Gale and Church 1991), they still need to be assigned correct analyses and correspondences for the subparts (a task which machines might aid with, but which probably requires human intervention). In the scheme outlined in Sato and Nagao 1990, the system will also require at least a syntactic analysis and generation component (and thus a grammar of the source and target language) so that input sentences can be assigned dependency trees, and output structures filtered. Of course, correlating structures in the example database raises the issue of correlating the correct structures for ambiguous sentences. At least in the foreseeable future, that means that such databases can only be constructed with human intervention.

An interesting experiment with the example-driven approach can be found in Kitano 1991, which describes an attempt to integrate an example-based approach with a parallel architecture.

An appealing feature of the Sato and Nagao 1990 system is that it can be integrated with a more conventional approach; for instance, the example database might be invoked for difficult constructions (like the Japanese "no" construction) or as a failsoft mechanism (IBM Japan is experimenting with the latter possibility); in other cases a conventional transfer or interlingua approach could be pursued. Indeed, the example-component can just be viewed as part of a syntactic transfer component, since it inputs and outputs dependency trees.

A somewhat different use of an example-driven approach is found in Kosaka et al. 1988a and Kosaka et al. 1988b. In the PROTEUS system described there, a target domain is chosen (in this case, the subject matter of the FOCUS *Query Language Primer*) and detailed domain-targeted analyses are done of an English text and its translations into Japanese. The texts on which the analyses are performed are said to define a *sublanguage.*

Analysis is done by hand by linguists, and the units of the analysis are very finely grained simple (or kernel) sentence patterns; as an example, there is a pattern called CREATE-TABLE-REPORT which has as instances sentences like "Table-commands produce reports" and "Table-commands generate reports."

The analyses themselves are performed in a framework based on the work of Zellig Harris; basically an analysis is a derivation tree whose node-labels consist either of kernel sentence predicates and arguments or of operations combining kernel sentences (such as passive and conjunction). The authors argue that the syntactic representations are abstract enough so that transfer is most often lexical; they take it as a methodological imperative that transfer rules should change structure as little as possible, which leads them in some cases to chose alternate

translations not actually found in the aligned texts. There is little detailed discussion for the motivation behind this imperative by the authors. It may be that minimizing structure changing transfer makes it easier to write a control-structure that deals with input sentences that invoke more than one pattern at a time.

The chief difference between this example-based approach and the approach taken by Sato is that the handbuilt database contains patterns rather than actual sentences. Thus there is no need to invoke a similarity metric when an input sentence does not occur in the database. Either an input sentence falls under some known pattern or it does not.

The trade-off between the two approaches is in the fineness of control at the similarity metric. The Proteus approach pays a high overhead in constructing a pattern database, but in return they have a linguistically characterizable similarity metric, with very fine distinctions being made. One of the useful properties of the pattern database is that it can cope with cases of so-called *zero pronouns* in Japanese. Japanese sentences often simply omit Noun Phrases where their content can be understood (places where a corresponding English sentence might use a pronoun). The pattern database then connects the elliptical Japanese pattern with an English pattern that makes the omitted Japanese Noun Phrases explicit. This, of course, might well happen on Sato's or Sumita's and Iida's approach as well.

Our example database (53) will serve to make a general point about the limitations of analogical approaches.

We begin with example-based approach. Note that the word "pen" in (52) is ambiguous. It can mean either an enclosure for animals or a writing implement; its translations will vary accordingly. A sufficiently large corpus of examples ought to contain both translations of (52) as basic examples; thus, similarity measures won't help. The only thing that might help is modifying our database so that larger environments can be appealed to in the input and examples. In other words, the example-based approach runs into the same problem as unaided linguistic analysis in dealing with genuinely ambiguous input. It incurs a new risk in that arbitrary biasing of the example database will arbitrarily bias readings assigned to ambiguous input. Thus, whatever reading of "pen" was chosen for the translations for (53) will be the reading the system choses for "pen" in (52). The flip side of this is that a domain-specific corpus will automatically tune translation preferences to that domain.

The same point could be made for the stochastic approach: in any given corpus containing sufficient occurrences of the above example

there will probably be some statistical preference for translating it one way or the other. But unless we are in a domain that makes the choice consistently, that statistically motivated choice will be wrong a large percentage of the time. The reason of course is that the information telling us how to translate this sentence does not lie in the sentence itself; it lies in the context of its use, and it is only by examining features of the context on a case for case basis that we have any hope of finding the right answer for a given text. One can perhaps do better on both the stochastic and the example-based approach by using larger and larger stretches of text as one's analogizing unit. But then the task of building the analogizing database grows increasingly unmanageable. Moreover each increase in the size of the analogizing unit may improve performance for some number of examples, but there will always exist some residue where the disambiguating cue lies outside the analogizing unit, either further out in the text itself, or in facts about how the text is being used.

The general problem here is *nonlocality*. Analogical approaches hold some promise for relieving certain very local problems of ambiguity (the ambiguity of Japanese "no" discussed above), but they hold little promise for those problems of ambiguity that have no linguistic domains that define them. Thus, a lexical ambiguity like that of the word "pen" may be locally resolvable (as it probably is in a compound like "pig pen"), but it may not. There are no guarantees on how far away the disambiguating information is. Similarly, the antecedent of a pronoun may fall within the same sentence or it may not; and knowing the antecedent of the pronoun *it* will be crucial for translating it correctly into German.

The basic problem with analogizing approaches then is not that they cannot be improved. They clearly can. It is that improving the fidelity of the statistical or example model only promises marginal improvement in the overall performance of a system. There will always be significant problems that fall outside the system's reach.

2.5.1.2 Knowledge-Based and Inference-Based Approaches

The Need for Nonlinguistic Information. The chief point of this section is that there is often information crucial to the correct translation of a sentence which cannot be found in its form. This is a point which has already been discussed at length in 2.1, but we return to it here because it serves to raise issues about how the various kinds of information which a translation system needs to have should be organized.

Here is an example encountered by the TAUM-AVIATION system and cited in Lehrberger and Bourbeau 1988:

(54) Connect pressure and return lines to pump.

This sentence has two distinct analyses. One can be paraphrased as "connect the pressure lines and the return lines to the pump," the other as "connect the pressure and then return the lines to the pump." These have distinct translations into French:

(55) a. Relier la pression et les canalisations de retour à la pompe.
b. Relier la pression et ramener les canalisations à la pompe.

There are various ways in which a system might make the correct choice (which is a.); a very simple strategy would be to make available to the system the information that pressure is not the sort of thing which can be connected, but pressure lines can be. This is not information about the word "pressure" nor the word "pression"; it is information about the thing in the world they are used to describe, the quantity measurement pressure.

System designers have often chosen to include information of this sort in their lexical descriptions as selection restrictions (see 2.1.4.3). Needless to say, doing this in a systematic manner makes the dictionaries very expensive to build. The task of encoding the restrictions is quite costly, and in general must be done separately for each domain rather than once and for all in a general dictionary.[9] Fo example, the TAUM-AVIATION system was also tested in an electronics domain, and in that domain the selection restrictions built up for the hydraulics domain "did more harm than good" according to Isabelle and Bourbeau 1985.

Although building in lexical selection restrictions is costly, it does do useful work. One can successfully disambiguate (54) with the information the verb "connect" does not allow the noun "pressure" as its direct object. There seem to be no low-cost alternatives for doing the same work. Nevertheless, there are built in efficiency problems with this approach that may be avoidable.

The problem stems from the fact that lexical selection restrictions are tied to lexical items and not to concepts. It is therefore necessary to apply them twice, once to source structures and once to target structures. At first glance, one might think that the restrictions will do no work on target structures: if a target comes from a well-formed source, it ought to be well-formed, assuming transfer rules have been

[9]Mr. Kawasaki (personal communication) of Hitachi points out that domain tuning a general dictionary is not a monotonic process of adding constraints to a subset of the dictionary. Readings for members of that subset also have to be *removed.*

correctly written. But this does not follow, since any structure that is ambiguous relative to the target language may still yield ill-formed target structures. HICATS from Hitachi is an example of a system that applies selection restrictions to both sources and targets.

The other point to make is that, even if one does allow lexical selection restrictions, more nonlinguistic knowledge is still needed in many cases. Having once made room in a Machine Translation system for nonlinguistic knowledge, there is a clear conceptual advantage in simplifying the dictionary by moving the selectional information out of it.

Consider the following pair of sentences, modeled on an example of Jaime Carbonell's:

(56) Replace the hexagonal nut with { a. the ratchet wrench. / b. a washer.

Most linguistic systems will give the (a) and (b) versions two analyses. Thus there will be one reading of (a) on which the ratchet wrench is used as an implement to aid in the act or replacing the hexagonal nut; and another (somewhat unlikely) reading on which it substitutes for the hexagonal nut. The likely and unlikely readings are reversed in the case of (b). Again, translations into some languages will differ depending on which readings are chosen. Finding the correct readings requires having information about the conventional uses of the things being talked about. It is not a property of the word "replace" or the word "nut" or the word "wrench," or even sentence (56a) that the wrench is being used as an implement rather than a stand-in. The proof is that the other reading *is* possible: if you have just told me that our collection of metal ballast is not heavy enough to anchor our hot air balloon, I might well utter (56) with the other reading intended. Thus, the kind of information that tells us which reading is more *likely* is information about wrenches and nuts and how they tend to interact. This is information about the world.

Two things have been argued for thus far: (a) that information about the world is essential to natural language understanding in general and to Machine Translation in particular; (b) that that information ought not be confounded with dictionary-building information or grammatical information in general. Dictionaries and real world models will both be easier to build if the kinds of information they involve are kept separate.

On both these points there is now general agreement among partisans of both linguistic approaches and AI approaches.[10]

[10]Work in AI-based or *knowledge-based* Machine Translation has been pursued

Once this point is clear, the issue may be sharpened considerably: given a natural language system which represents knowledge about the world, how is it to dynamically access this information and apply it to particular cases? In other words, how is the system to *reason* from general knowledge to specific cases. We will confine our discussion of reasoning to inferencing, reasoning that involves premises and conclusions The paradigm case of inferencing is deduction, although we will use the term for two other kinds of reasoning as well, *induction* and *abduction.*

Deduction is inferencing using the laws of logic. To take a famous case:

(57) All women are mortal.
Hypatia is a woman.
Therefore Hypatia is mortal.

This sequence of sentences has the useful property that in any imaginable world, if the first two sentences are true, then the third is true. It is a valid argument. It is the task of logic to characterize valid arguments. The above argument may be formally recast in first-order logic:

(58) $\forall x[\mathrm{W}(x) \rightarrow \mathrm{M}(x)]$
$\mathrm{W}(h)$
$\mathrm{M}(h)$

The way in which validity is characterized in first-order logic is as follows. There are rules for telling whether a particular sentence is true for a particular way the world might be; the formal object corresponding intuitively to a way the world might be is called a *model.* It turns out that in any model which makes the first two sentences true, the third sentence must also be true. Characterizing validity, and thus deduc-

over a number of years (see Carbonell et al. 1978, Carbonell et al. 1981, Tomita and Carbonell 1987, Nirenburg 1987a, Nirenburg et al. 1988a, Leavitt et al. 1991, Nirenburg and Goodman 1991, Wilks 1973, Wilks 1976b, Wilks 1979). The discussion in Shann 1987 is a good review of a debate between knowledge-based and linguistics-based approaches. Although at one time there was a strong current of anti-linguistic rhetoric among followers of the AI-based approaches, and a popular view that no linguistic information needed to be explicitly represented in a natural language system, that rhetoric seems to have softened. The general view common among those pursuing the AI-based approach is that knowledge of a grammar and lexicon is among the many kinds of knowledge a Machine Translation system needs to include, and that it is good system design to separate that knowledge from knowledge of the world. This was, for example, the point of view adopted in Nirenburg 1987b and Carbonell and Tomita 1987; the latter adopts the Functional Grammar formalism of Kay 1984, very much a linguistics-based approach.

tion, turns out to be a matter of characterizing sequences of sentences that preserve truth.

A crucial step in being able to talk about truth-preservation in a language is that it be *interpreted*, that is that there be a way of assigning interpretations to the parts of any sentence so that, given a model, we can tell whether the sentence is true in that model.

Now let us return to the problem of selection restrictions and the particular facts about pressure and connecting that were relevant for (54). We might represent them logically as follows:

(59) $\forall x[\mathrm{Q}(x) \leftrightarrow \neg \mathrm{P}(x)]$
$\forall x[\mathrm{Pr}(x) \rightarrow \mathrm{Q}(x)]$
$\forall x, y[\mathrm{C}(x, y) \rightarrow \mathrm{P}(y)]$

Here, let P,Q and Pr represent the concepts of a Physical-Object, a Quantity, and a Pressure, and let C be the connecting relation. Then the first formula tells us that physical objects and quantities are disjoint: nothing can be both at the same time. The second tells us that any pressure is a quantity, and the third, that anything which is connected is a physical object. Now suppose our representation for one reading of (54) is:

(60) $\exists x[\mathrm{Pr}(p) \wedge \mathrm{C}(x, p)]$

From the fact the p is a pressure quantity and is connected, we can derive a contradiction. Thus, we can rule out the undesired reading of (54) purely by deduction.

Unfortunately there are a number of problems that deduction does not help with. The issue in example (56) had to do with how wrenches are used in *most* cases for wrenches. Unaided deduction is not much use for reasoning about the default case or what usually happens. Valid arguments are arguments that must be true in all cases, not just in most cases. What turns out to be useful in (56) is *abductive* reasoning.

Abduction is best described as follows. In deduction, from $(\forall x)p(x) \rightarrow q(x)$ and $p(A)$, one concludes $q(A)$. In induction, from $p(A)$ and $q(A)$, or more likely, from a number of instances of $p(A)$ and $q(A)$, one concludes $(\forall x)p(x) \rightarrow q(x)$. Abduction is the third possibility. From $(\forall x)p(x) \rightarrow q(x)$ and $q(A)$, one concludes $p(A)$. One can think of $q(A)$ as the observable evidence, of $(Ax)p(x) \rightarrow q(x)$ as a general principle that could explain $q(A)$'s occurrence, and of $p(A)$ as the inferred, underlying cause or explanation of $q(A)$. Of course, this mode of inference is not valid; there may be many possible such $p(A)$'s. Therefore, other criteria are needed to choose among the possibilities. One obvious criterion is the consistency of $p(A)$ with the rest of what one knows. Two other criteria are what Thagard 1978 has called consilience and sim-

plicity. Roughly, simplicity is that $p(A)$ should be as small as possible, and consilience is that $q(A)$ should be as big as possible. We want to get more bang for the buck, where $q(A)$ is bang, and $p(A)$ is buck.

We therefore need a scheme of abductive inference with three features. First, it should be possible for goal expressions to be assumable, at varying costs. Second, there should be the possibility of making assumptions at various levels of specificity. Third, there should be a way of exploiting the natural redundancy of texts to achieve greater simplicity and consilience.

We now show how abduction can help select a reading for (56). Understanding a sentence is finding a way to make it true. In the ground case, we can just assume it. But then we incur the assumability costs of the entire sentence content; if some part or all of the sentence content can be proven, or can be proven with a low cost assumption, then we save on total belief costs. The general role of cost-based abduction in resolution now becomes clear: when a sentence has two or more readings, choose the reading that can be proven with the lowest cost assumptions. There are two readings of (56); on one, the wrench is used as an instrument:

(61) $\exists e, w, n[\text{Wr}(w) \wedge \text{Hxn}(n) \wedge \text{Rep}(e) \wedge \text{Patient}(e, n) \wedge \text{Inst}(e, w)]$

On the other, it is used in place of a hexagonal nut:

(62) $\exists e, w, n[\text{Wr}(w) \wedge \text{Hxn}(n) \wedge \text{Rep}(e) \wedge \text{Patient}(e, n) \wedge \text{Rep-With}(e, w)]$

The two readings differ only in which relationship the wrench stands in to a particular event. If we make the assumption that a wrench is an instrument with lower cost than the assumption that it is the replacing object in a replacement, then we will *all other things being equal*, prefer the more natural reading.

The key point is *all other things being equal.* Adjusting our assumability costs changes the outcome. If we are in a context where there is a low cost way to prove that a particular wrench is functioning as a replacement, then the outcome will change. This is just what we want.

Problems. One problem all these approaches have in introducing world knowledge is using it: abductive and deductive reasoning, even on small domains, are expensive. Typically, finding that a particular candidate fails involves detecting an inconsistency, and inconsistency detection is *very* expensive. Other problems with the abductive approach are described in Norvig and Wilensky 1990.

Another problem with AI-based systems, and with linguistic-based systems converging in the same direction, is the most salient fact about real world knowledge: there is so much of it. Many Machine Trans-

lation researchers have conceded the importance of real world knowledge in translation, but have been content to build systems without it, simply to see how far they can get. The reason is that building up knowledge representations even of small domains is time-consuming and expensive. The general response of a committed knowledge-based translation advocate is to admit this is true and to set a high priority on (a) building tools to make knowledge representation easier; (b) evaluating what kinds of domain knowledge do the most work (that is, make the most often invoked distinctions); (c) providing means of interacting with users to access those kinds of knowledge which it is impractical to incorporate into the system. The chief challenge for this approach is to develop a methodology for covering a small domain in a reasonably perspicuous manner, and to give a feasibility demonstration: quality translation for that domain.

2.5.1.3 Connectionism

We shall have little to say about connectionism within Machine Translation here, because there is little activity in the area. One example of an experiment in connectionism within Machine Translation is the JANUS system described in Jain et al. 1991. The chief claims to connectionism for the JANUS system lie in its connectionist speech recognition component, and its connectionist parser, described in Jain 1991. We do not consider this a connectionist Machine Translation system because connectionism plays no crucial role in the actual translation process.

Rather than trying to conceive what a connectionist system would be like, we will address some issues about how connectionism might play a role in resolution, and how it might be integrated with linguistic analysis.

The greatest promise of fruitful interaction lies with examples of the following sort, noted in Reder 1983 and in Uszkoreit 1991:

(63) a. The astronomer married a star.
b. The movie director married a star.

Here, there is a clear preference for the absurd reading in (63a). Now the crucial fact is that the reading is absurd for both sentences, but somehow the mention of an astronomer in the first sentence activates an association with stars, and makes the absurd reading more accessible. Note that what seems to be going on has nothing to do with which reading is more likely. Thus an approach which relies on inference to the most likely explanation—like the abduction approach of 2.5.1.2—is likely to have trouble accounting for this preference.

One can imagine a connectionist approach with something like the following flavor: inputs are all the words in a sentence. When words are

ambiguous both senses are activated. The network has been trained in advance so that word senses with some strong association (established either by corpora studies or an association dictionary) serve to reinforce each other. Then the occurrence of "astronomer" above should serve to strengthen the activation of the astral body sense of "star."

Note that this kind of associative weighting can proceed in parallel with conventional linguistic processing. All that is required at the linguistic end is an architecture that can be sensitive to weightings of some sort when a choice needs to be made. We have already suggested that such an architecture is desirable in the case of inferencing. It does not seem unlikely that it might be desirable at the level of lexical choice as well. In such a system, sense preferences could be set dynamically by a spreading activation network.

Linguistic processing involves a large number of choices. For example, parsing is generally just the pursuit of long sequences of alternatives, most of which fail. Parsing choice-points are thus natural candidates for an analogous treatment: choices might be rigged in advance with some sort of weighting scheme. Whether that weighting scheme needs to be dynamically altered is another question to which we have no answer. In Uszkoreit 1991, the question of how to integrate a declarative grammar with weighting schemes that guide processing is addressed and a straightforward scheme is proposed. Examples like (63) are discussed and turn out to fit nicely into the scheme, irrespective of whether a spreading activation network or some other device is used to set the preferences.

In sum, we believe the greatest realm of promise for connectionist processing lies in accounting for preferences which have nothing to do with likeliness or deduced consequences—which may very well be viewed as simply the activation of certain concepts.

2.5.2 Direct, Interlingual and Transfer Methods

In this section we assume a Machine Translation system that does some linguistic analysis, and investigate three different possibilities for tranferring results of the analysis to the target language: the traditional options are the direct, transfer, and interlingual methods.

Imagine the following sort of Machine Translation system: a source language sentence is read in and a morphological analysis of the words is done, and perhaps some syntactic analysis as well. However a syntactic representation of the entire sentence is not built. Instead the system attempts to build a sentence in the target language on the basis of the morphological information and partial syntactic information. The target language sentence also does not have an explicit syntactic

representation; perhaps some morphological processing is done to compute the forms of certain words. The output usually corresponds fairly closely in word order to the source.

This procedure very crudely describes the first generation of Machine Translation systems, systems like the Georgetown system of Garvin 1967. The strategy taken is called the *direct* approach (for obvious reasons). Underlying the direct approach is the implicit claim that *no representation of the source or target at any level of analysis needs to be built.* In its extreme form that might be phrased: *no understanding of the source at any level of analysis is required.*

Contrasting with the direct approach are two approaches that began to emerge during what is known as the second generation of Machine Translation systems. These are the *interlingual* and *transfer* approaches.

The interlingual strategy is to analyze the source sentence and represent its meaning in a language which mediates between all languages: it is an *interlingua.* From that semantic representation a sentence in the target language is generated. Unlike the direct approach, the interlingual approach builds a representation as a result of analysis, and what is generated in the target depends on that representation. Thus, the interlingua approach requires at least two modules, an analysis module and a generation module. It also needs an interlingua rich enough to simultaneously represent the output of analysis and the input to generation. This means, in effect, that the interlingua must be expressive enough to capture anything expressible in any languages it mediates between. In a fully general system, that means anything expressible in any language.

Note that on the interlingual approach, once analysis is complete one can simultaneously generate translations in a number of different targets. On the direct approach, three different analysis stages would be needed.

We can define the transfer approach as a kind of compromise between the direct and interlingual approaches. In Figure 4 we see a famous diagram due to B. Vauquois which illustrates the relationship between the direct, transfer, and interlingua approaches. The basic difference revolves around whether there is a representation which is input to the translation mapping and how much analysis is required to reach that representation. The tradeoff is the following: in general, the more analysis is performed the easier the mapping is, because deeper analysis means more abstract representations with fewer language-particular features, but more detailed analysis also correlates with more difficult

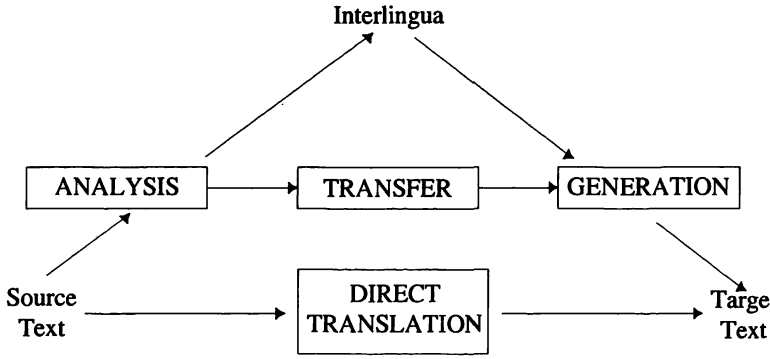

FIGURE 4 Transfer and Interlingua Pyramid Diagram

generation, because the source of generation will have features particular to the target language. So easier mapping rules are traded for more complex analysis and generation rules. On the direct approach minimal analysis is performed and no representation of the source is built; with this strategy the translation mapping is very difficult, because all the idiosyncrasies of the source must be directly aligned with all the idiosyncrasies of the target. At the cost of some analysis, the translation mapping is facilitated; we have moved up the triangle and the path of the translation-mapping is shorter. Systems known as transfer systems do at least morphological and syntactic analysis, and typically perform the transfer mapping on a relatively abstract syntactic representation, such as the dependency trees discussed in 2.4.3. However, there are systems called semantic transfer systems which perform the translation mapping on a semantic representation, that is, a representation with fewer language-particular features than are encoded in most syntactic representations. With enough analysis, theoretically, we have a representation so abstract that all differences between source and target languages have been abstracted away from. The connection between the two languages is just the identity mapping. This is the apex of the triangle, and the ideal of the interlingua approach.

We will devote only brief discussion to the direct approach. It approach has the virtue of having many years of practice behind it. The oldest systems are the best ones because the only way to improve quality is to add to the number of idiosyncratic special cases that it can handle. The quality that can be achieve with these systems tends to a fairly low asymptote to which the oldest and best systems are by now fairly close. Systems like Systran, which dates back sixties, have the problem that they are very difficult to extend without breaking

something that previously worked. One reason for this is that translation mapping rules have been stated directly in the surface forms, and thus have grown quite complicated. To get the system to perform a certain restructuring when a certain syntactic construction is observed, for example, requires writing special recognition routines for that construction, and then inspecting the interactions with all the other special recognition routines which perform independent actions of their own. There is no generation grammar filtering outputs; there is only the empirical evidence of how the system performs given all the interactions.

Perhaps the most important difficulty with direct systems is that analysis and generation algorithms have to be written afresh. There is no theory of a kind of representation which cuts across linguistic differences, and thus no algorithm for constructing one from a string or finding a string from a representation.

In what follows we discuss some of the basic pluses and minuses of the transfer and interlingual approaches, because they raise issues of genuine interest about the correct design of a Machine Translation system. But at all times Vauquois's triangle should be kept in mind; the basic lesson is that the two approaches fall on a continuum. At times it is difficult to distinguish systems which perform transfer on a more abstract level (semantic transfer) from systems which call themselves interlingual but include some sort of adjustment or replacement rules. Indeed we shall argue that a certain kind of pure interlingual scheme (what we'll call the naive interlingual scheme, which never requires adjustment or replacement of the analysis) is unworkable. In the end, the issue of interest is not whether to pursue a transfer or interlingual approach; the issue is which levels of analysis are necessary, and how to arrive at a representation suitable for generation of a target text.

2.5.2.1 The Module-Counting Argument

A commonly made system-design argument for the interlingual approach is what we shall call the *module-counting argument.* Consider an n language MT system. Whether the transfer or interlingual approach is taken, such a system must have n analysis and n generation modules. What distinguishes the two approaches is that the transfer system must also have $n(n-1)$ transfer modules.[11] For the interlingual system, adding a new language means adding an analysis and generation module. For the transfer system the same is true, but $2n$ new

[11]Systems professing to have reversible transfer rules only require half as many transfer modules, but this fact does not affect the argument.

transfer modules also have to be built. There is thus a clear economy argument for the interlingual approach.

But the module-counting argument needs to be taken with a grain of salt. For one thing, the modules that are being counted in the two approaches are not exactly the same sorts of things. This point is illustrated by Vauquois's Translation triangle: when the transfer method is used analysis can be shallower and simpler than it is in an interlingual system. Simple module counting obscures the fact that the analysis and generation modules in a transfer system will do less than the corresponding modules in an interlingual system; in fact, the whole point of the triangle is really that *for a particular language pair* transfer is a short cut; that is, *for a particular language pair*, analysis, transfer, and generation will be simpler than analysis and generation in an interlingual system for that pair, simply because similarities between the two languages can be exploited. Structures that remain the same in the two languages can opportunistically be left the same; in the interlingual system, one has to disassemble all source language input into a representation abstract enough to serve for all languages, and then reassemble them into the particular structures of the target.

Given that no one has yet designed a satisfactory translation system for even a single pair, there is a reasonable argument that the correct research strategy is to first try to build a transfer system for some candidate language pair.[12] Whether that system can then be generalized to an interlingual system is then a separate question.

An alternative sometimes pursued is to narrow the ambitions of the interlingual approach. Rather than aiming for a single language suitable for capturing any meaning expressible in any language, the system builder designs a meaning representation that can mediate between two languages, or among a family of related languages. Such a representation language is called a *pivot* language.[13] Objections to the complexity of an interlingual approach now disappear. The difficulty of building an interlingua has been circumvented, but the module-counting argument is now weakened, since adding a new language means building a new interlingua.

[12]This in fact seems to have been one reason why most of the Machine Translation community has favored the Transfer approach since the Grenoble CETA group switched from Interlingua to Transfer in 1972. An additional historical factor seems to have been the rise of transformational grammar. Transfer rules in the GETA system were viewed as tree-to-tree mappings in the style of transformations (see Vauquois and Boitet 1985).

[13]For example in the original French-Russian system at Grenoble (see Vauquois and Boitet 1985)), the interlingua adopted was only intended as a pivot.

Another factor in comparing interlingua versus transfer is the reusability of various parts of a transfer module. As we saw above, transfer may be performed on fairly abstract levels of representation; on the assumption that such representations look reasonably alike for different languages, particularly closely related languages (the basic assumption underlying the design of EUROTRA Interface Structure, see 2.1.2), transfer rules mediating between different language pairs may look alike. It is thus plausible that there could be a transfer system in which transfer rules were shared by a number of different language pairs, with minor parameterizations.

Eurotra's ⟨C,A⟩, T framework is a transfer-based approach to Machine Translation. In Arnold and des Tombes 1987, criteria for the kinds of representations which are the output of analysis and the input to generation have been identified. The criteria apply equally well to an interlingua or to the representations which constitute the input and output of transfer: they are (1) utility: translation should be easier given the analysis representation; (2) specifiability: it should be possible to write rules that map between surface strings and their appropriate representation.

The ⟨C,A⟩, T framework is a transfer framework and tries to satisfy the criteria above by imposing requirements of *constructibility*, *compositionality*, and *one-shotness* on representations and transfer operations. Constructibility means any mediating representation language has to be specifiable by a generative device which recursively enumerates the expressions. Compositionality means the translation of a larger expression must be a function of the translations of its parts. One-shotness means that a primitive translation must yield a target language representation in a single-step. The relevance of constructibility and compositionality to specifiability and utility are fairly clear. One-shotness is an effort to avoid problems that previous transfer systems have had with complex rule interactions. Strict one-shotness was felt to be difficult to achieve, so that ⟨C,A⟩, T allows for the possibility of representations intermediate between the analysis structure and the generation structure. Mappings between such intermediate representations must satisfy one-shotness, and they must satisfy constructibility. In fact, ⟨C,A⟩, T systems have had a number of such intermediate representations (see the discussion in Arnold and des Tombes 1987). The methodology and ideas in ⟨C,A⟩, T have had their followers. An interesting example is the MIMO system described in Arnold and Sadler 1990, which adopts constructibility, and compositionality, and strengthens one-shotness to exclude intermediate representations.

Still another possibility is the possibility that all transfer modules

in an n language system output and input the same kinds of representations, and thus can be composed. Then translation from language A to language C could be performed by composing the A-B transfer module with the B-C transfer module. There would be no need to build an A-C transfer module. This is in fact the strategy contemplated for ARIANE-78 and described in Boitet 1989.

A radical implementation of the interlingua approach is embodied in the ROSETTA system (at Philips), a system based on some of the ideas of Richard Montague. In a Montague-style grammar, each syntactic rule is paired with a semantic rule. In ROSETTA, an isomorphism is assumed between grammars of the source and target language. There is then a crucial addition to Montague's program: Syntactic rules paired under the grammar isomorphism must share the same semantic operation. Thus, semantic derivation trees can function as the interlingua; it is trivial to use the grammar isomorphism to map from a semantic derivation tree for the source to a syntactic tree in the target language. Obviously to achieve isomorphism, source and target grammars must be written together, and syntactic semantic distinctions that would not be made for either grammar in isolation must be made for both. The main arguments for this approach are (1) that the system is reversible; and (2) any structure that can be analyzed by definition yields an interlingual analysis that generates a well-formed target language sentence. Thus, at the cost of greatly complicating the grammar-writing process (for syntax and semantics), the central problem of how to affect the translation mapping is solved. Adhering to Montague's methodology also guarantees two senses of compositionality: (1) semantic: the interlingual analysis of an expression is a function of the interlingual analyses of its parts; (2) translational: the translation of an expression is a function of the translation of its parts. For a discussion of the basic principles, see Landsbergen 1987.

An obvious criticism of this approach is that grammars are quite hard enough to write one at a time. Larger monolingual natural processing systems already encounter serious problems of grammar complexity and rule interaction. It is thus unclear whether this strategy will scale up to handle complex grammars and languages that differ significantly in syntactic structure.

A more important criticism is that this whole line of argument carries to an extreme that fallacy of thinking of translation as a function from a source to a target text. We have kept returning to the point that translation is essentially a process in which information is lost and gained: translation is not meaning preserving. All attempts to work out the necessary compromises in a machine-based systems have

so far been very disappointing. To talk of composing such systems as one would compose mathematical functions is therefore unlikely ever to make practical sense.

2.5.2.2 Language-Pair Independence

We now discuss two assumptions of the pure interlingual approach, motivating the module-counting argument.

- **Language-Pair Independence.** Translation does not require any information specific to a *pair* of languages. All information necessary to translation is information about the analysis of the source (independent of the target) and generation of the target (independent of the source).
- **Language of Thought.** Analysis and generation are linked by a representation in *interlingua*, a language adequate for representing the semantics of any sentence in any language. That is, it must be expressive enough to capture any distinction made in any language.

Now in fact the *Language of Thought* assumption does not entail *Language-Pair Independence*, as we will see below. But *Language-Pair Independence* does appear to require some kind of language of thought. It is really *Language-Pair Independence* which underlies the module-counting argument, but we will argue below that there may be advantages to a language of thought (even in a rather uninteresting form) even when true language-pair independence is unattainable.

One objection occasionally raised to the interlingual approach is that it raises a design problem no one has yet solved, the design of an adequate language of thought. But if we take the definition of a language of thought to be simply that it be capable of expressing any concept expressible in any language, then there is a trivial way to design an interlingua, namely to take the union of all word senses in all languages; such a *polylingua* will indeed be adequate to express any meaning in any language. If there are word senses that really *are* shared by two languages, then they will be captured by a single symbol in the polylingua. Natural kind terms like English "gold" and French "or" are good candidates for words that have such senses. Many proponents of the interlingual approach would be unhappy with such a polylingua. In the next section, we try to say why.

2.5.2.3 The Naive Interlingual Scheme

We use the term *the naive interlingual scheme* for one that is based on claim that a representation can be found of the invariant in translation that will be the same for a source sentence, or text, as for its translation

in any other language. The argument against this view need not delay us long because it it follows directly form points that we have insisted on already quite strongly. The position is essentially untenable for essentially the same reason that the view of translation as meaning preserving is untenable. There may, however, be a more robust notion of interlingua, and we will explore this possibility shortly.

The simplest attack on the naive interlingual position can be mounted on the basis of translation mismatches, discussed in 2.1.2. Translation mismatches involve cases where one language encodes a concept not directly expressible in the other; the cases that make the point most directly are cases where the target language forces a choice not made in the source. Thus, for example, in translating from Japanese to English there is the recurring problem of selecting articles for Noun Phrases. Another example from Barnett et al. 1991a was discussed in 2.1.2. Spanish has two lexical items corresponding to English "fish," "pez" and "pescado," differing in that the latter denotes a fish caught for food. Presumably, the interlingual representation of a Spanish sentence containing either of these words could not be the same as the English translation using the word "fish".

Arguments of this kind have convinced some advocates of a basically interlingual scheme to add some special rules which consitute a move in the direction of transfer. But a few embarassing cases, they might argue, are not sufficient to justify the abandonment of an overall scheme with so many advantages.

Examples of transfer rules in a basically interlingual environment can be found in the Hitachi English-Japanese system (HICATS JE, described in Kaji et al. 1989). The system uses dependency graphs as its interlingua. But certain kinds of readjustment, which require a radical re-structuring of the English input, are handled by transfer-style restructuring rules. For example, in Japanese, sentences with inanimate subjects are generally quite infelicitous. Thus an English sentence like "This system has a wide range of application" has a translation like this:

(64) Kono sisutemu ha ouyouhan'i ga hiroi.
(in) this system application range (is) wide

The interlingual dependency graphs corresponding to the English and Japanese sentences are quite different: the Japanese dependency graph has "hiroi" ("wide") as its root, while the English dependency graph has "have" as its root. The HICATS-JE system uses a restructuring rule to get from one graph to the other.

Translation mismatches occur because of general structural prefer-

ences like the Animate Subject constraint of Japanese, and for a variety of reasons that we have already considered at some length. One way of looking at this kind of problem is to say that not everything in the interlingual translation of a given piece of source text will be consumed in the course of generating the corresponding piece of target text. Conversely, some additional interlingual material, not available from the source may be required to generate the translations of other pieces. The question of just how this is brought about, and how the circumstances are recognized in which such fluid matching should be allowed is an obviously complex matter.

The naive interlingual scheme does not really rule out the possibility that different languages conceptualize the same parts of the world differently; but it does rely on the hope that whenever that happens, there is a single more abstract conceptualization underlying the two divergent ones. Translation mismatches show that the real picture is more complicated. Sometimes it is not even possible when two languages share the lexical apparatus to cover the same conceptual territory. A simple example will illustrate.

Consider the Japanese verb "nomu." Often translated as *drink,* it actually has a slightly wider domain of application, as it can also be used with medicine as its object, even when that medicine is not liquid, or with small objects, such as a coin (children may sometimes "nomu" coins). The most natural translation of English "drink water" into Japanese will be "Mizu-wo nomu." A literal gloss of "nomu" might thus be "swallow."

If this informal analysis is correct, then our natural interlingual representations for English "drink" will be some (not necessarily primitive) piece of conceptual structure, call it DRINK; while our analysis of Japanese "nomu" will be some different piece of conceptual structure, call it SWALLOW. Now the key point is that SWALLOW is also the right piece of conceptual structure to associate with the English verb "swallow." This correctly predicts that "nomu" may sometimes translate "swallow." What is completely unpredictable without further principles is that "nomu" may sometimes translate "drink." That happens at least partly because Japanese lacks a lexical item meaning "ingest liquid"; since the act of drinking water includes swallowing it, the conventional way of describing a drinking situation in Japanese is simply to assert that swallowing is going on. One may truthfully assert just the same thing of just the same situations in English, but in the presence of a more conventional (and more specific) way to describe drinking water, the claim "she is swallowing water" sounds quite marked. The first context that comes to mind is drowning.

This analysis supports the view that "nomu" and "drink" do not map onto the same piece of conceptual structure, so that we are confronted with another case of translation mismatch. In going from English to Japanese, generation fails; Japanese lacks apparatus for exactly expressing the DRINK concept, and then either a complex expression in Japanese must be found, or a new interlingual expression must be found, perhaps equivalent to the source language analysis in some sense, perhaps a "best fit."

In this particular case, we might want to argue that "nomu" and "drink" do indeed map onto the same piece of conceptual structure but that what appears to be a classic case of translation mismatch is in fact traceable to a collocation in English which has no counterpart in Japanese. If this position can be upheld, the solution will be much simpler. The claim would be that "nomu," "drink," and "swallow" do map onto the same piece of conceptual structure, but that English requires different verbs with different kinds of object. In particular, "drink" is required when the object refers to a liquid. The trouble with this account is that it does not suggest why we would say of someone recently saved from drowning, that "he had obviously swallowed a lot of water," and not "he had obviously drunk a lot of water."

2.5.2.4 Translation by Negotiation

Given the arguments made in the early parts of this chapter, we should not be surprised if the naive interlingual approach does not look promising for Machine Translation. The conclusion follows from the observation that translation is not a meaning preserving function from a source to a target text. Indeed, it is probably not helpful to think of it as a function at all, but rather as a matter of compromise. Even the best of writers has no sure way of evoking in the minds of his readers exactly the notions and images he wants to evoke and, if the writer does not, then surely the translator does not either. As we have seen, the devices that serve in one language can be used in another language in a distressingly small number of cases.

We therefore contemplate a process that derives a series of interlingual representations for the source, preferably in order of decreasing plausibility. For each of these, it attempts to find target strings with interlingual representations that differ as little as possible from these. At some point, one of these is chosen as being close enough to be allowed to stand as the translation. To the extent that this process is based on a representation of the text that is partial to neither of the languages, it is interlingual. To the extent that the representation of the target string is not identical to that of the source, it is a trans-

fer process. We think of it as primarily an interlingual approach, but nothing hangs on the word.

The kind of process we have just sketched is the basis of an approach that we call "negotiated translation" which takes as primary the notion of translation as compromise. We shall come to the question of what the interlingual representations should look like in 2.5.2.5. Here we confine ourselves to illustrating the basic idea.

The overall design of a system for negotiated translation has three major parts responsible for analysis, generation, and negotiation, respectively. The negotiator mediates between the analyzer and the generator and, to that extent, it fills a role similar to that of a transfer component. For simplicity, we consider it operating on sentences one at a time, and in isolation from one another.

The analyzer delivers to the negotiator an interlingual representation of a sentence to the negotiator. The negotiator hands the interlingual representation to the generator which, on very rare occasions, may be able to find a sentence in the target language that has exactly this as one of its interlingual representations. A more likely outcome is that the generator returns a candidate translation to the negotiator together with two kinds of information which, adapting some terminology from statistics, we refer to as *errors*. They are:

1. A set of properties of the source sentence, stated in interlingual terms, that are not expressed in the candidate translation, and
2. A set of properties of the target sentence that are not expressed in the source sentence.

We call these *errors* of type 1 and 2 respectively.

Now, it is up to the negotiator to decide if the errors it receives back from the generator are acceptable. It can assume that they are the best the generator can do with the given input so that the alternative would be to modify the specification in some way. If none of the errors touches the meaning of the sentence too closely—if they are all predicates having to do with information structure or syntactic structure—then it might be a good idea to accept the translation. If there are differences in meaning, the negotiator would have to assess them in some way.

The following example is obviously contrived, but it will illustrate the point:

(65) He ate the deer.

Let us assume the the interlingual representation is something like the following:

```
E: eat(E), past(E), nonprog(E), agt(E, X), male(X),
   one(X), pat(E, D), deer(D), (D)
```

We can read this somewhat as follows:

> There is an E which is an eating in the past and not progressive, the agent of which is X. X is some definite male individual. The patient of E is D which is definite and is one or more deer.

Here are some possible translation of the sentence into German

(66) a. Er ass das Reh.
b. Er ass die Rehe.
c. Er frass das Reh.
d. Er frass die Rehe.

We take it that the generation component of our system produces each of these, but with the errors listed below:

Er ass das Reh
1. nonprog(E)
2. one(D), person(X)

Er ass die Rehe
1. nonprog(E)
2. several(D), person(X)

Er frass das Reh
1. nonprog(E)
2. one(D), beast(X)

Er frass die Rehe
1. nonprog(E)
2. several(D), beast(X)

All of the translations show an error of type 1, namely *nonprog(E)* because the German generator has no way of making any distinction that parallels that between progressive and nonprogressive verb forms in English. Let us assume that the negotiator accepts this immediately as unimportant. The translations are also distinguished on the basis of whether one, or more than one, deer is in question. The word "deer" in English happens to be one that does not show the usual distinction between singular and plural. Let us assume that the negotiator can settle this by examining its records of the preceding text and observing, say, that there has been talk of one deer, but only one. This leaves us with only (a) and (c). It turns out that German has two verbs corresponding to the English "eat" between which it distinguished on the basis of whether or not the eating is being done by a human being. We therefore get either *person(X)* or *beast(X)* as a type 2 error. The

negotiator can settle this also by examining the context to find out what "he" refers to or, failing that, what kinds of eating have been referred to previously. If that does not work, it might fall back on knowledge of how the word "eat" is most often used in the language as a whole, or in texts on topics similar to the current one.

Let us suppose that the negotiator turns up evidence that the word "he" might refer to a person of whom nothing has been said except that this person was named "Bill." If the negotiator knew that a person with the name "Bill" might well be male, it would have further evidence that the correct identification had been made. We imagine that this can also be settled through negotiation, this time with the analyzer rather than the generator. The analyzer is called upon to verify *male(X)*.

Processes like those we have just sketched are part of the process of seeking and abstract representation of a target sentence, given an abstract representation of the source sentence. In traditional terms, this makes them part of the transfer process. However, the representations to which these processes are applied are not specific to either language. Of course, we used English words as the names of the predicates in our imaginary interlingua, but this was only for mnemonic value and not because the interlingua is supposed to be oriented more towards one language than the other.

Much of what a machine translation system must do—like the inferences just exemplified—is not connected with the particulars of either of the languages involved in the translation and should therefore be conducted in a representation that is tied to neither of those languages. In our example of the operation of the negotiator, we had it apply to the analyzer or the generator for any information resting on knowledge of their languages, with the aim of making the operations as nearly generic as possible. Such independence of one component from another is clearly desirable on standard grounds of modularity. To the extent that it can be achieved, it will throw light on parts of the translation process that are very poorly understood, as well as holding out the hope that the negotiator, which could grow to become the largest component in the system, will be totally independent of the languages that the current system happens to be working with.

The picture that emerges is shown in Figure 5, a modification of Vauquois's translation triangle from Figure 4.

2.5.2.5 Semantic Representations: What's in the Interlingua

In 2.5.2.4, we argued for an interlingual approach to translation, using negotiation. At the very least we argued for a semantic representa-

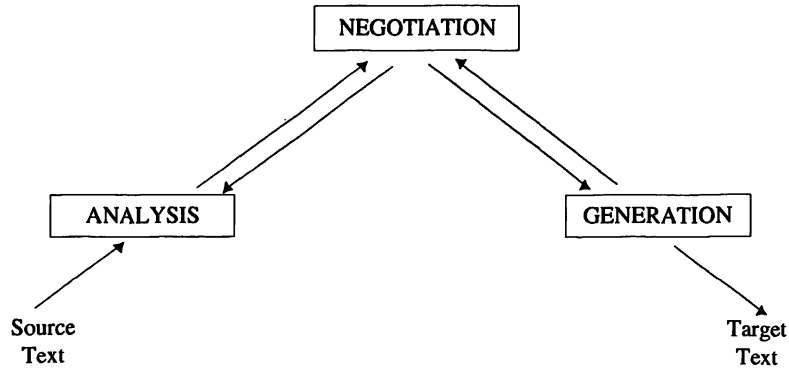

FIGURE 5 Translation by Negotiation Pyramid Diagram

tion fine-grained enough to capture every distinction in the language of interest. In 2.5.1.2, we argued that deduction and abduction would be necessary for solving resolution problems; Hobbs and Kameyama 1990 argue that deduction and abduction serve for certain kinds of translation problems as well. What this suggests is that at least some component of our interlingua be an interpreted language capable of supporting a notion of valid inference. What that suggests is some representation at least loosely based on predicate logic. For some candidates, see Kamp 1981, which introduces Discourse Representation Theory (DRT), Nerbonne and Laubsch 1991, which introduces Natural Language Logic (NLL) from the Hewlett-Packard Natural Language system, and Alshawi et al. 1988, which introduces Quasi-Logical Form (QLF) in the SRI Core Language.

One advantage of any Machine Translation system that uses such a representation is that the analysis module is then usable for a number of different natural language processing applications besides machine translation. Analysis through a logical representation is a standard strategy in application domains like query-answering and message-understanding. Given the cost and effort involved in building a Machine Translation system, this is an important consideration.

Furthermore, without making any claims about the sorts of conceptual structures different languages will carve out, a logic-oriented representation provides many of the advantages of previously proposed abstract representations in Machine Translation. Dependency relations are recaptureable by predicate-argument relations and recurrences of the same variable in different argument-positions. Semantic roles such as deep-case roles are provided in a number of logical formalisms (QLF and NLL), but language-specific syntactic properties such as whether a

particular argument has been realized as a subject are abstracted away from.

As always in a reasonably abstract representation, the question arises as to whether too much has been abstracted away from. For example, on the analysis given in most systems using logically-oriented semantics, an active and a passive sentence will have the same semantic representation. Thus information about the voice of the source will be lost during the generation phase and a passive source may very well receive an active translation.

There are a variety of replies to this objection. One is to argue that syntactic properties such as voice ought not to be translated because they are not abstract enough. Thus, the English passive is far more common than the cognate French construction, and English passive sentences are routinely translated into French using the French reflexive middle, the impersonal pronoun "on," or a change of verb. These observations do not alter the basic intuition that (67b) is a better translation than (67a)

(67) a. Une voiture a tué le médecin.
b. Le médecin a été tué par une voiture.

of

(68) The doctor was killed by a car.

What this example shows is that correct translation is something that takes more into account than just propositional content as captured by logical representations. What it does not show is that correct translation preserves as much of the source syntax as can be preserved in the target language. This is because surface syntactic similarities may mask deep differences of function. Thus, the Polish passive syntactically resembles the French passive but is even rarer, and it is almost never correct to translate an English passive with a Polish passive. Yet much of the relevant nuance can be captured by choosing the correct word order from among the relatively free Polish possibilities. The key point is that capturing analogous discourse-function seems to be far more important than mimicking the source language syntax.

We thus assume that there is something more to interlingua than a representation of propositional content. The interlingual representation must reflect information coming from a variety of different levels of abstraction in the analysis of the source and target strings. Besides propositional content, it should at the very least also reflect the "information structure" of the strings as well as possible, that is, it should say what information purports to be given, and what new. In the best of all possible worlds, it would probably not contain strictly

grammatical information, such as the distinction between active and passive. However, pending a better understanding of just what discourse function this distinction serves, we are prepared to countenance representing such things as active and passive directly.

There is a more serious objection to standard logical formalisms: rather than expressing too little, they express too much. Thus standard logical approaches assign two distinct representations to "Every man loves some woman," one expressing the claim that there is some single woman adored by every man, the other claiming that to each man a loved woman can be paired. Such ambiguities, known as quantifier-scope ambiguities, are quite common. The curious fact is that, quite often, translations preserve these ambiguities. Thus, to have to map to a representation which resolves the ambiguity, only to reintroduce it at generation time, seems wasteful. This is particularly true since resolving quantifier scope ambiguities is in general a difficult task involving reasoning about the world or the discourse or both.

A solution to this problem is to choose a logical representation which is neutral as to scope. Researchers have experimented with such representations for reasons quite independent of their utility in Machine Translation, hoping to find a representation on which some reasoning can be done before scope ambiguities are multiplied out. It is then possible, for example, to eliminate a single scope-neutral representation on the grounds that a semantic restriction has been violated. Thus, a number of readings for a multiply ambiguous sentence can be eliminated in a single stroke.

To this end, QLF includes the possibility of unscoped representations; and Poesio 1991 proposes a variant of DRT with the same effect. The situation with NLL is more complex, but it allows representations in which the scopes of indefinites and definites have not been fixed.

We now turn to the question of what predicates an interlingua might contain. We noted in 2.5.2.4 that translation mismatches made it pointless to expect all translations to have the same interlingual representations. Given that complete conceptual decomposition won't solve the problem, one might very well take as one's interlingua something like the polylingua (or union of all senses) we discussed in 2.5.2.2.

A Machine Translation system incorporating such an interlingua might work as in the following example:

(69) John held the gold in his lap.
physical-possession(john, the(x, element-79(x)), LAP)

Predicates like *physical-possession* and *element-79* are good candidates for language-neutral interlingua senses; while the concept of a lap, a

body-location that only exists in a sitting position, is peculiar to languages like English and German.

Call the representation strategy embodied in (69) a *polylingua* strategy. One may contrast the polylingua strategy with the strategy of an interlingua that attempts a decomposition into primitive conceptual components, which we will call a componential strategy. On the componential strategy, English "lap" would be decomposed into some complex of more basic concepts. This complex would require periphrastic expression in many languages, but in mapping to German it would have to be reassembled into "Schoss," a costly process since it requires the generation component to analyze and recognize a particular configuration of predicates. On the polylingua strategy, English "lap" is analyzed as LAP, a concept directly expressible as a single lexical item in only a few languages. For those languages which lack the means to express LAP, generation fails and negotiation is required. A simple model of that negotiation might invoke deduction: there is an axiom relating laps to hips and knees. That axiom fires when generation fails for LAP, enabling an approximation: since John can be assumed to be sitting, "in his lap" can probably be rendered well enough with something corresponding to "over his knees."

The polylingua strategy, then, is to do *lazy conceptual decomposition*, to invoke conceptual decomposition only when it is required to express something in the target language. It is a key advantage of the translation-by-negotiation strategy that failure in the course of generating is a run-of-the-mill event. Conceptual decomposition, when it is necessary, can thus be assimilated to negotiation.

Conceptual decomposition certainly will be required. Relating English "punch" to French "donner un coup de poing" will require knowing at least that parts of the agent's body are involved in punching.

In sum, we identify the follow properties of the interlingua:

1. It should be interpreted; that is, it should have a semantics that supports inference.
2. It should represent predicate argument relations clearly.
3. It should probably have the facilities for leaving certain kinds of standard semantic distinctions, such as quantifier scope, vague.
4. It should have facilities for representing something more than just propositional content; notions like given, new, focus, and topic may play a role. Failing that, there may be a need for representing syntactic information about the source. But such auxiliary information should be kept separate from content.
5. Conceptual decomposition should be pursued sparingly.

2.6 Current Machine Translation Systems

In this section we present a brief survey of current Machine Translation activity throughout the world. The survey is not complete, but it attempts to be representative.

2.6.1 Japan

Research in Machine Translation has grown far faster in Japan in the 1980's than anywhere else in the world. The reasons for this are varied, but there are at least three important factors.

1. Need: Japan is a translation society. Everyone everywhere needs to produce or have access to translated material, whether they are in academics, business, or a technical field. A glance at the JEIDA report (see Nagao 1989a) shows that in a survey of nearly 2,000 companies (about a fifth of which were translation companies) three-and-a-half million pages were translated in one year. 53.6% of that was from Japanese into foreign languages, with English the leader. 41% of the documents translated were manuals of some kind.
2. Firms are well-positioned for Machine Translation R&D. The proof of this is in the facts below. There were a large number of firms with resources and interest in Machine Translation research.
3. Government Sponsorship: Machine Translation in Japan got a boost with the Science and Technology Agency sponsorship of the MU project in 1982, MITI sponsorship of the Electronic Dictionary project of 1986, and the multilingual CICC project for Asian languages, and the founding of ATR, supported by the Ministry of Post and Telecommunications.

2.6.1.1 Commercial Systems

The first two firms into the field appear to have been Fujitsu and Hitachi. The Fujitsu system ATLAS I claims to have had the first prototype system in 1982, the same year that saw Science and Technology Agency funding begin for the MU project. ATLAS I is a syntactic transfer English-Japanese system.

Fujitsu (Atlas I, II) Fujitsu's work on Machine Translation centers on two systems, ATLAS I and ATLAS II. The first is for translation from English to Japanese and the second for translations from Japanese to English, though there are plans to extend it to German and Chinese. ATLAS I is described as using the "semantic transfer" method, and

ATLAS II is interlingual. ATLAS II (see Uchida 1989) is the newer and more ambitious system, and we will restrict our attention to that.

ATLAS II incorporates a special stage of *conceptual transfer.* A *world model* is used for disambiguation: conceptual structures that result from the analyses of sentences are checked to see if they can be validated in the world model; if not they are rejected and another attempt at interpretation is made. The interlingua allows for concepts that are unique to a particular language, and where possible, for those concepts to be related to more primitive concepts. The syntactic generation component accesses three knowledge sources: generation rules, co-occurrence relations, and adjacency relations. Conceptual structures are traversed by generation rules that resemble ATN parsing rules, with an arcname, condition, and action. Co-occurrence relations enforce collocational restrictions. Although conceptual well-formedness has already been guaranteed, the system allows a single concept to be expressed with more than one word: the choice of which is then a matter for the co-occurrence component. Thus, co-occurrence restrictions might distinguish between "make an attempt" and "do an attempt," for example. Adjency relations appear to check whether two morphemes can be adjacent. The system translates from Japanese to English using grammars of approximately 5,000 phrase-structure rules in both languages, with 500 transfer rules and a 70,000 entry lexicon.

Fujitsu makes use of a system of "semi-automatic pre-editing" to pretranslate clichés, simplify complex sentences, and insert brackets to clarify certain kinds of attachment ambiguities. The automatic pre-editor also attempts to restore elided phrases in certain instances by using words taken from the title. The claim is that this works 60% of the time.

Hitachi (HICATS) The HICATS system uses 3,000 English and 5,000 Japanese grammar rules and 50,000 lexical entries in each language. The HICATS JE system translated from Japanese to English. It is described as transfer-based in Kaji 1989, although transfer is done on dependency graphs with case roles, so that the results of analysis are quite abstract. One argument given for the dependency graph approach is that it "facilitates case-pattern driven" analysis, which allows for reconstruction of missing elements in Japanese. Semantic features on lexical items are used to enforce selection restrictions, as defined in 2.3. Generation utilizes a phrase-structure grammar for English with a top-down algorithm, with selection restrictions having to be checked again (since they belong to words, not their semantics). The Grammar-Description language of the system is called GDL. Dependency graphs

are nodes with attributes and labeled arcs, and rules are graph transformation rules. Attributes include features such as part of speech, inflection form, tense, definiteness, and nuance.

Toshiba (AS-TRANSACT) TRANSAC stands for Translation Accelerator. The system is written in C and is implemented on a Toshiba AS3000. It has been supplemented with a bilingual editor and Japanese and English editors. The bilingual editor is intended to provide aids for post-editing. The basic Machine Translation system consists of an ATN grammar, a semantic component which builds what appears to be a target language-specific conceptual structure, and a generator. The description of the system in Amano et al. 1989 suggests that there is no separation of the semantic analysis from transfer; in other words the conceptual structure that results from a semantic analysis is passed directly to the generation component. The system uses an ATN grammar of 10,000 rules and a lexicon with 50,000 entries.

Oki (PENSEE) This bi-directional transfer-based system uses a small grammar of about 1,000 rules in each language and a large lexicon of 90,000 entries in the Japanese to English lexicon and 60,000 entries in the English to Japanese lexicon.

NEC (PIVOT) This bi-directional interlingua-based system interacts in batch and translation mode. The interlingua represents conceptual and discourse information such as topic and focus. There is some attention to semantic issues ignored by many Machine Translation systems, such as quantifier scoping, negation, and comparison. Interestingly, the system developers complain of having problems extending the system to a third language because of a lack of a clear definition for the interlingual concepts. The grammar is based on case frames, and the lexicons have 70,000 (English to Japanese) and 90,000 (Japanese to English) entries. There is also a list of some 800,000 specialized technical terms.

Sharp (OA-110WB) The DUET bi-directional, semantic transfer system has a 1,000 rule augmented context-free grammar and a lexicon of 60,000 entries plus 40,000 technical terms. The system attempts to achieve robust parsing by dividing grammar rules into two sets—major rules which constitute the core of the grammar, and minor rules that are intended to take care of infrequent or grammatically questionable constructions. The major rules alone are used in the first pass of the parser and only if this fails are the minor rules invoked. The English-to-Japanese system is already on the market. A revised version of the system, using dependency grammar, and translating from Japanese to English is in field test.

It is Sharp's declared intention to incorporate its Machine Translation systems in hand-held machines and to sell them together with OCR equipment for English input. They also claim to have OCR for Japanese.

Sharp is working on a number of problems for the longer term. These include parallel parsing, discourse structure, the syntactic problems of coordination, ellipsis, and long sentences, as well as *layered dictionaries* and the use of thesauri. The term *layered dictionary* refers to the well known technique of looking words up first in the user's personal dictionary, then in a technical dictionary, specialized for the subject matter of the text, and finally in a general dictionary.

IBM-Japan (Jets, Shalt-I, and Shalt-2) IBM Japan is an extremely lively center of research on Machine Translation. In recent years, they have worked on three different systems, with a number of interesting innovative features.

The earliest of these, called *Jets* is a Japanese to English translation system using dependency grammar and a scheme to provide for interaction with the user. The transfer component of the system is based on relational grammar. The relational structures that are constructed are said to be largely language independent so that they take on something of the quality of an interlingua. The English generation component incroporated a *planner* which maps out the rules that will be need to generate the given sentence in advance.

The more recent systems are Shalt-I, and Shalt-2. Shalt-2, for example is an English to Japanese transfer-based system with a grammar of 200 English, 900 Japanese and 800 transfer rules, plus a 90,000 entry lexicon.

IBM-Japan is actively pursuing research of a number of different kinds. They are experimenting with a Machine Translation system that uses example-based translation. They are unusual among builders of large systems in that they produce more than one output and provide for interaction with the user to select the appropriate one. The sentence is displayed on a screen which uses a different color to set off any part of the sentence for which a different rendering is available. By pointing to this with the mouse, the user can see, and possibly select, one of the alternatives.

Mitsubishi (MELTRAN) This Japanese to English transfer-based system uses a phrase-structure grammar of about 1,000 total rules and a lexicon of 50,000 basic and 30,000 technical entries, primarily in the field of information processing. Input can be by OCR and pre-editing is suggested.

Matsushita (PAROLE) This Japanese to English semantic transfer-based system consists of 800 English and 300 transfer rules. The lexicon has 31,000 entries in each language. Matsushita is collaborating with the CMU Center for Machine Translation system. Its speech recognition system is being used in conjunction with the CMU translation system in an experiment in speech-based machine translation.

Ricoh (RMT) Ricoh's English to Japanese transfer-based system uses dependency structures. It has 2,200 augmented context-free grammar rules, 300 transfer rules and a lexicon of 30,000 entries.

Canon (Lamb) This Japanese to English transfer-based system is a small-scale research system limited to a lexicon of roughly 2,000 words.

Sanyo (SWP-7800 Translation Word Processor) This bi-directional transfer system is based on dependency structure. There are 650 augmented Context-free grammar rules, and a lexicon of 50,000 entries in each language.

Catena (Star) This English to Japanese transfer system is also based on dependency structure. There are 2,000 context-free grammar rules, and a lexicon of 20,000 entries in each language. A very high speed is claimed for translation, 15,000 words per hour or roughly a minute per page. A number of alternative translations are displayed.

2.6.1.2 Government Funded

JICST (Japanese Information Center for Science and Technology) JICST uses a Machine Translation system based on MU to translate abstracts of technical articles. The articles are pre-edited by a human, translated by machine, and finally post-edited by a human before being entered into a database. Users can then query the database to find abstracts of articles in a foreign language. The pre-editing stage involves a check by machine which makes suggestions concerning long compounds, long sentences, and ambiguous particles. The total translation process costs 4,000 Yen per abstract with a human translator and 2,000 Yen with machine translation.

EDR (Electronic Dictionary Research) One of the central goals of this MITI-funded project is an effort to build large dictionaries that will aid in the construction of Machine Translation systems. The dictionaries fall into four types:

1. word dictionaries: grammatical features
2. concept dictionaries: concepts
3. Co-occurrence dictionary: collocational information
4. Bilingual dictionaries: translations

These four dictionary-types capture distinct, complementary properties of words; for any given word, a Machine Translation system might need to consult all four dictionaries during processing. The dictionaries are written in a general format intended to be customizable to the needs of any particular system.

The concept dictionary organizes concepts into a semantic net, with hierarchical (ISA or SUBTYPE) links and case-role links capturing the relations between concepts. As an example of how fine-grained the distinctions get, we find, under the category of *action requiring someone else, concerning the vertical relationship with the other person, and not including transfer of information* concepts like *ordering*, *revenge* and *scolding*.

2.6.1.3 Research and Academic

Kyoto (MU) The University of Kyoto is where the MU project was funded and led by Professor Makoto Nagao. The system is no longer under development, but interest and research in Machine Translation continues there. Copies of many commercially available systems are kept and studied there. The idea of example-based translation was suggested by Nagao, and developed by his students in Nagao 1984 and Sato and Nagao 1990.

ATR (SL-TRANS). ATR has the only long-standing, broad coverage project on Dialogue translation. The focus is on translation of phone conversations rather than translation of face-to-face dialogue. We discuss the ATR speech translation project at greater length in 3.9.1. The translation component is a transfer module using the HPSG formalism.

2.6.2 North America

2.6.2.1 Commercial

ALPS (Automatic Language Processing Systems Limited) ALPS was founded in 1980 by members of the Brigham Young University translation group in Provo, Utah, which was active from 1970 through 1979, and focused on interactive translation systems. Originally known for its Machine Assisted Translation systems, it now appears that ALPS is branching out into other areas. According to Hutchins 1988, ALPS has purchased several translation bureaus and is marketing a medical expert system. It also markets text-processing tools, providing aids for writers including dictionaries, thesauri, and bilingual dictionaries. Its three translation related products are:

1. Transactive

2. Autoterm
3. Translation Support System

Weidner Weidner Communications Corporation. Provo, Utah. Now owned by the Japanese Company Bravis. The original product from Weidner was a translator's assistant integrated with an editor, with a dictionary that could be customized by the user.

In 1988, WCC reported a new System II which was based on LFG. Languages translated between include English, French, German, Italian, Japanese, Portuguese, and Spanish, although it is not clear that System II supports all these.

Logos Logos was founded in 1969 by Bernard E. Scott, launched with a contract to build a direct translation Machine Translation system for the US Air Force for translating American aircraft manuals into Vietnamese. The analysis phase of Logos was very similar to the early Systran system. LOGOS I did a morphological analysis and a preliminary syntactic analysis, locating Noun Phrase boundaries but not attempting to assign a structure to the entire sentence. Transfer first modified English word groups to Vietnamese-like word groups, and then did lexical substitution. LOGOS I was evaluated by Sinaiko and Clare (see Sinaiko and Klare 1972). Development on the LOGOS system continued. By 1973 an English-Russian version of LOGOS III existed, and later versions for Russian, German, French, Spanish, and Arabic were built.

In 1982 the Logos "Intelligent Translation System" was released. This product differed from previous incarnations of LOGOS in that semantic information was now used for resolving lexical and syntactic ambiguity.

The system uses a "syntactico-semantic" language called SAL (see Scott 1989), which implements "something like valency grammar." The language appears to be used to implement some semantic type hierarchy which is used for analysis and transfer and perhaps disambiguation. It is instructive to inspect some of the motivation behind it:

> It is widely recognized ... that the mind abhors complexity and that its principle means for avoiding it is abstraction, by assimilating differences to something common and therefore more abstract. In much the same way, SAL achieves semantic simplification through abstraction, through the reduction of intensional values to second-order concepts. The analysis performed by the LOGOS model however was also influenced by another characteristic of the psychological model, namely, the fact that the human mind makes no use of algorithms as it "understands" a sentence

> ... This fact of the psychological model led us away from algorithmic formalisms in the direction of more purely heuristic procedures. This bias against algorithmic solutions was further strengthened by the realization that algorithms in fact work well only with formal or artificial languages and in fact cannot cope with natural language *qua* natural language. As I said, the spectre of logic saturation was one that haunted our earliest reflections on the processing of natural language in a machine environment, and that led us eventually away from the Turing Machine model, and by an extension Church's thesis, from all algorithmic formalisms. But having said this, neither is it true that the alternative to an algorithm could ever be merely a series of ad hoc procedures, as this is just another path to logic saturation, and to chaos.

LOGOS's largest customer was Nixdorf, but since the acquisition of Nixdorf by Siemens, Nixdorf has begun using METAL.

Systran The Systran system came out of the original Georgetown Machine Translation project, GAT, in particular the SERNA version which translated from Russian to English (S Russkogo Na Angliskij). In 1968, Peter Toma, one of the central figures of the GAT project, founded Latsec, which started off with a contract with the US Air Force Foreign Technology Division. The system is now owned by the Gachot company.

As the oldest Machine Translation system, with the widest community of users, Systran serves to provide a standard against which to measure the performance of latercomers. Versions of the system translate between English, German, Russian, French, Spanish, Dutch, and Portuguese. Many of these systems have been developed over several decades now and have grown quite large. Extending them is a delicate process; not only because the systems have grown quite complicated and hard to change, but because long-time users have come to value translation consistency even more than translation quality. For example, the US Air Force Technology Division keeps a database of previously translated texts, against which it measures every new system release. The standard for acceptance is ten improvements for every degradation.

Use of Systran requires extensive post-editing. Users generally are large organizations able to support in-house translation facilities with translators who are gradually trained in the use of the system. Users of the Systran system include The US Air Force Technology Division at Dayton, the Commission of European Communities, General Motors of Canada, the German National Railway, the German Nuclear Research

Center in Karlsruhe, and Xerox. Xerox achieved a considerable improvement in performance by instituting a restricted form of English called Multinational Customized English for use in the manuals to be translated.

Smart Smart Communications provides a translating system called the Smart Translator, which operates on a restricted version of English. There is a version translating between French and English which is used by the Canadian Ministry of Employment and Immigration for job postings.

SMART includes a pre-editing system, MAX, which reads a document file and produces a set of suggestions for modifying it into Multinational Customized English. The editor uses a special terminology database, a rule base of 2,500 rules for technical writing, and a knowledge base customized to the needs of the user.

2.6.2.2 Non-Profit

Pan-American Health Organization (ENGSPAN and SPANAM) The Pan-American Health Organization uses two Machine Translation systems to translate between English and Spanish, SPANAM (Spanish-English, used since 1980), and ENGSPAN (English-Spanish, used since 1985). Documents translated are technical articles in the field of medicine and health. SPANAM is a direct translation system, with procedures tied to the particular language pair it handles (for example, routines for Spanish reflexives and negation). There is very little attempt at translation, and heavy post-editing is required. ENGSPAN does a somewhat more sophisticated syntactic analysis than SPANAM, and accordingly needs a dictionary with more information.

2.6.2.3 Academic and Research

There are a number of groups and projects at Carnegie-Mellon University's Center for Machine Translation. Here is a sampling.

KANT The information here is largely based on Mitamura 1991. KANT is an interlingual Machine Translation system with an LFG grammar and parser, and an independent generator. The objectives of the system include no post-editing and no human intervention during disambiguation. There is an emphasis on providing tools to help with the costly task of building a domain model.

DIOGENES This is a natural language generation system for Machine Translation.

JANUS This speech-to-speech translation system is describe in Jain et al. 1991. Speech Recognition uses connectionist acoustic modeling (a Linked Predictive Neural Network) and stochastic modeling based on a

bigram grammar. Linguistic processing is done in two alternate ways: (1) a convention LR-parser and (2) a connectionist parser described in Jain 1991. Generation is done by GenKit, a system that compiles a generation grammar into LISP (see Tomita and Nyberg 1988). Speech synthesis was performed by commercially available text-to-speech systems, one for Japanese and one for German.

IBM Jelinek's group uses the statistical approach, as discussed in section 2.5.1.1 of this report, and in Brown et al. 1989.

A separate group, led by Michael McCord, uses a level-based approach (see McCord 1988, McCord 1989).

Microelectronics and Computer Technology Corporation The Machine Translation group at MCC is working on a Prolog-based interlingual system.

New Mexico State New Mexico State University's Center for Research in Language (CRL) has a Machine Translation group which has been most closely associated with the knowledge-based approach; a good summary of the basic emphasis is Wilks and Farwell 1990. The system described in Farwell and Wilks 1990 is called ULTRA.

ISI The Information Sciences Institute in Los Angeles. A group led by Eduard Hovy will be working on a DARPA grant on the generation component of a Machine Translation system in collaboration with the CMU Center for Machine Translation and the New Mexico State University group.

2.6.3 Europe

Grenoble The GETA (Groupe d'Etudes pour la Traduction Automatique) project is the longest continuing Machine Translation research group in Europe and perhaps the world. Bernard Vauquois led the group until his death, and since then it has continued with Christian Boitet at the helm. From 1961 to 1971 the project was known as CETA (Centre d'Etudes pour la Traduction Automatique). The Russian-French system developed at CETA is generally recognized as the first of the big second-generation systems, systems that abandoned the Direct approach, and embraced a methodology of linguistic analysis, transfer, and separation of language-specific information from translation information. Its influence has been felt in Machine Translation systems everywhere, including EUROTRA and Japan (through the MU system). Descriptions of the work in the 70's can be found Vauquois 1975.

Since 1977 GETA has shifted its focus towards preparing systems to be transferred to industry, designing a system called ARIANE which

uses a grammar and transfer-rule formalism called ROBRA. This system, like the EUROTRA systems, still follows the syntactic-transfer methodology of the CETA system.

Siemens (METAL) METAL is the result of a collaborative agreement between Siemens and the University of Texas at Austin struck in 1978. Over the years the Austin project has closed down, but the technology has been transferred to Siemens in Germany. Pilot versions translating between English and German are in use (by Nixdorf, for example).

The Machine Translation system proper uses the ATN-parsed, syntactic transfer approach described in Slocum et al. 1987 and the fitted-parsing techniques of Jensen and Heidorn 1982 for failed parses.

Philips (ROSETTA) An interlingual system translating between English and Dutch which builds in the radical assumption that target and source language grammars are isomorphic; following Montague in Montague 1974, each syntactic rule is paired with a semantic operation.

ISSCO (ELU) ELU is a unification-based reversible translation system based on PATR which translates between German, French and Italian. It is described in Estival 1990 and Estival et al. 1990. An application involving snow reports is described in Van Noord 1990b. Transfer rules are written directly in the PATR style notation, using templates that have a source and target attribute. The MIMO-2 system at the University of Utrecht takes a similar approach.

University of Manchester (UMIST) The University of Manchester Institute of Science and Technology (UMIST) has developed N_{tran}, an interactive transfer system using LFG to translate from English to Japanese. The interaction module presupposes no knowledge of Japanese, and interaction focuses on disambiguating source language inputs (analysis resolution), and on resolving cases of translation mismatch (see 2.1).

UMIST also has a project funded by British Telecom on a system that guides a monolingual English user through a menu to help write a business letter in a foreign language.

UMIST is also working in collaboration with ATR on their dialogue translation system.

Essex MIMO is described in Arnold and Sadler 1990. This is a formalism for transfer Machine Translation that basically adopts the ⟨C,A⟩,T formalism of EUROTRA, enforcing compositionality and constructivism, but adding the requirement of reversibility.

University of Utrecht MIMO-2 is a unification-based reversible translation system which translates between English, Dutch, and Spanish. It is described in Van Noord 1990a and Van Noord 1990b. The grammar formalism is based on PATR and the approach to grammatical description is influenced by HPSG. An interesting feature of the the approach is that transfer rules are written directly in the PATR style notation, using templates that have a source and target attribute. The ELU system at ISSCO takes a similar approach.

Utrecht (DLT) The Buro voor Systeemontwikkeling (BSO) at Utrecht began work on DLT (Distributed Language Translation) in 1982 on an EEC grant, and a longer term project began in 1985. The system is described in Schubert 1987; it is an interlingual system with the unusual property that it uses a human language as its interlanguage. The language is Esperanto. Analysis is quite complex since the target of analysis is an unambiguous representation in Esperanto. In fact, DLT may be seen as a double transfer system, which transfers source language input into Esperanto, and then transfers from Esperanto to the target language. One argument given for using Esperanto as an interlingua is that it clearly has the necessary expressiveness, already being a human language. Given that a human language should be used, so the argument goes, Esperanto is especially manageable because of its regularity.

SRI Cambridge (CLE) Using the CLE (Core Language Engine) system, a small reversible semantic transfer system was built for translation between English and Swedish. Transfer was performed at the level of QLF (Quasi-Logical Form). The experiment is described in Alshawi et al. 1991b.

British Telecom A speech-based Machine Translation system has been under development since 1984. The version described in Stentiford and Steer 1987 was a small-scale experiment coupling a speech recognition system with a template-matcher and a translation phrasebook (database) of about 400 precanned phrases, with about 1,000 different words. Matching a phrase is accomplished through a search for one or more keywords in the recognizer output.

3

Speech Recognition

3.1 Why Use Speech?

For decades, the keyboard has been the primary means of input to computers. In recent years the mouse and other pointing devices have provided another input channel. Given the pervasiveness of keyboard and mouse interfaces, one might assume that typing and pointing is sufficient for all input needs. Unfortunately, this is not the case; there are still applications for which typed input is inadequate. This chapter demonstrates the usefulness of speech input, and surveys the technology used to recognize speech.

Speech is the most natural, convenient, universal, and characteristic means of communication available to the human. It is natural because it requires no special equipment or training, and convenient because it allows comfortable communication at a rate of over 170 words per minute (wpm) without encumbering the hands or eyes. Even tasks that have a non-verbal component, such as describing directions, are facilitated when speech is allowed in addition to, say, drawing maps. Speech is also universal—all healthy humans learn to speak and understand the language they are exposed to at birth—and it is characteristic of humans: no other animals speak.

Contrast this to typing, where both hands are required, training is mandatory, and even a skilled typist reaches only about 100 to 150 wpm. Typing requires more attention than speaking, and it is impossible to, say, draw a map while simultaneously typing.

Written language is derivative of speech. Written language has served to advance human culture and is important for formal documents, but speech remains the medium of choice for fast, informal communication.

In this section, we examine in more detail why speech is such a good

communication channel, and describe the capabilities and limitations of computer speech recognition systems.

3.1.1 Why Speech is Attractive

Speech is the fastest and most natural way we have to communicate words. There are at least eight factors in speech's favor:

1. Speech requires *no training.* It takes practice to type well, or to operate a joystick, mouse or other specialized equipment. It even takes training to learn to read and write. But every healthy human learns to speak.
2. Speech is *fast.* Spontaneous speech has a normal rate of 120 to 210 wpm. In contrast, handwriting communicates only 24 wpm, unskilled typists can generate 12 to 24 wpm, and skilled typists generate 100 to 150 wpm.
3. Speech *requires little attention.* It is easy to speak and attend to other tasks at the same time. In a problem-solving task that required simultaneous thinking and communicating, experienced typists were able to achieve only 18 wpm, while speech continued at the normal rate. Another experiment Ochsman and Chapanis 1974 involved a two person problem-solving task and ten possible communication channels. Using speech alone, the task was solved in an average of 16 minutes, and in 12 minutes with speech and handwriting. Of the communication combinations that did not involve speech, the fastest (handwriting plus video) took 23 minutes. Speech appears to put the lowest demand on the user's processing capabilities.
4. Speech *has few physical limitations.* A speech user can walk around, use his or her hands for another task, and focus the eyes anywhere. Contrast this to the typist, who must sit in one place in front of a keyboard, and must devote some eye attention and both hands to the typing task. Typing is associated with an increasing number of repetitive strain injuries. Speech is less affected by extreme conditions than other media. Speech can be used to communicate in darkness and around corners. For aerospace applications it is important to note that speech is not affected by weightlessness, and is less hampered by acceleration than are mechanical alternatives. For example, it takes 4.0g acceleration to cause a 10% reduction in speech recognition accuracy, but only 0.8g to cause a 10% reduction in input accuracy with push buttons or dials (Turn 1974).

5. Speech input *requires only a microphone.* This is cheaper and far less bulky than alternatives such as a keyboard and display screen. This is important for uses like an airliner cockpit that are already loaded with switches and dials, and for users who must carry their equipment with them.
6. Speech is, of course, *understandable to humans.* Thus, a user can simultaneously address an input to both the computer and another listener in the room. A transcript of the speech input could be made and read by humans at a later time. This is not true of, say, input provided by hitting buttons.
7. The *naturalness* of speech makes it widely acceptable. Many professionals—including the international business people who would be using Verbmobil—are reluctant to use keyboards and other devices, but feel comfortable using speech to communicate. Naturalness is especially important under stress. When in danger, a user is more likely to remember to speak correctly than to type or push buttons. Unfortunately, stress can change the user's tone of voice, so it is important that the speech recognition systems functioning in environments liable to induce stress be able to deal with that.
8. Speech is *compatible with the world's largest network*—the phone system. Touch tone buttons are also compatible with much of the phone system, but are much slower and offer a limited number of responses.

In summary, speech has proven itself to be the highest bandwidth communication channel we have. Its effectiveness can only be limited by faulty computer recognition technology. That is why it is worth investigating and developing that technology to its fullest.

3.1.2 Practical Uses of Speech

In the previous section we saw some of the advantages of speech, and thus we know why any face-to-face computer-mediated interaction—such as that envisioned in the Verbmobil scenario—will be more effective if it includes speech recognition technology. In this section we point out that speech recognition technology can also be useful for a variety of applications that are less ambitious than Verbmobil. Following is a list of applications, in roughly increasing order of difficulty, that could be built from the basic technology needed for Verbmobil:

1. Limited Domain Information Retrieval. Speech is the ideal input medium for communication over a telephone. Office switchboards, banks, movie theaters, as well as other offices could be

equipped to handle routine queries with speech recognition technology. If the system provides the user with very explicit prompts ("what is your account number?," "do you want to withdraw, deposit, or see balances?") then the responses will be limited and the recognition task will not be difficult.

2. Concept Spotting. Speech recognition can be used to search for keywords in any kind of voice data base, such as recording of a cockpit or courtroom dialogue. Searching for a single word is much easier than identifying all words, especially because the cost for a small number of false positives is not high.
3. Voice I/O. Speech can be used to control all kinds of computer peripherals and appliances, from the VCR and lighting in a house to the telephone or automatic map in a car to complex equipment in a hospital or airplane cockpit. This is especially useful in situations where the hands or eyes are required for other tasks. For example, a medical technician doing a blood count does not want to look away from the microscope to write down or type in a number. A doctor doing a sonogram may have both hands full and would welcome the chance to control the sonogram machine by voice. Speech recognition has been used by airline employees who use both hands to direct luggage on a conveyer belt and use voice to direct the luggage to the right destination.
4. Prosthesis. For most users, speech recognition offers convenience over other input channels such buttons and switches. But for some physically handicapped people, speech recognition is the only possible way to interact with certain devices. For the deaf, a speech typewriter could be used as an "automatic ear" to transcribe the speech of others. Speech synthesis could be employed by the mute.
5. Captioning. Simultaneous captioning of live television news shows is done by trained stenographers. The stenographic characters are then automatically transcribed into subtitles. Since the transcription is error-prone, it is conceivable that a speech recognition system could be used for this task. However, it is a much more difficult application because the domain is completely unconstrained.
6. Speech typewriter. The speech typewriter or automatic dictation machine is another difficult task with an unconstrained vocabulary. It is more difficult than captioning because the tolerance for errors is less. Of course, if a speech typewriter could actually be built it would have enormous economic impact.

7. Simultaneous Interpretation. This is the Verbmobil application: a combination of the speech recognition technology that would be necessary for real-time high-quality captioning with Machine Translation technology and perhaps speech synthesis.

3.2 The Difficulties of Speech Recognition

Experimental evidence clearly shows that speech is the best channel of communication for a wide variety of tasks. However, there are places where speech does not work so well. It does not work in very noisy environments, nor in environments where there is no air to transmit the signal, such as underwater. A nearby eavesdropper can intercept speech communication, so there is a potential loss of privacy. There is also the potential for "speech pollution"—an annoying increase in the noise level if an entire office full of workers are all talking to their machines.

Speech has not been widely used as an interface to computers because it is difficult to decode. The difficulties stem mostly from two problems: *segmentation* and *variability*.

Segmentation is the problem of deciding where one word ends and the next begins. Native speakers are so accustomed to understanding spoken language that this is easy, that there are short pauses between words that serve to separate them. In fact, while there are pauses between some words, spectrographic analysis shows that words often run together in normal fluent speech with no pause at all between them.

The difficulty of segmentation was demonstrated by Reddy Reddy 1976 in an informal experiment. He read the sentence "In mud eels are, in clay none are." to four subjects, and asked them to reproduce what they heard. The replies are shown below. The subjects were not able to find the boundaries between these familiar words because they were used in an unusual sentence structure with a preposed prepositional phrase.

In mud eels are,	In clay none are.
In muddies sar	In clay nanar
In my deals are	en clainanar
In my ders	en clain
In model sar	In claynanar

Another example was constructed by B. F. Skinner, who showed that the first sentence below, if spoken quickly, sounds very much like the second:

Anna Mary candy lights since imp pulp lay things.
An American delights in simple play things.

Segmentation problems are not unique to English. In German, the colloquial word "duselig," spoken quickly, sounds the same as the phrase "du selig." Similarly, the words "nachtisch" (dessert) and "nachttisch" (bedside table) can sound like the phrase "nach tisch" (after having eaten).

The second problem is *variability*. There is a huge variation in how an individual word can be pronounced. There is variation by sex and age—men tend to have deeper voices than women or children, for example. Differences in geographical regions and socio-economic class produce different accents. And each person has a unique vocal tract that leads to individual differences.

But even for an individual speaker, the same word can be pronounced with varying speed, stress and pitch depending on the context. A speaker with a cold sounds different because the nasal passages are constricted, so the flow of air through them is altered.

In fluent speech the beginning and end of the word may be altered depending on the preceding and following words. This is called a *co-articulation effect*—if one word ends with the tongue in a certain position in the mouth, and the next word begins with the tongue in a different position, then the two words will tend to both have the tongue in an intermediate position. Co-articulation effects are the most difficult and pervasive of all sources of variability.

The discussion of speech knowledge begins with the *phoneme*, the smallest distinctive unit of sound in a language. For example, the sounds /bit/ and /pit/ both consist of three phonemes. They represent different words because the /b/ sound is a different phoneme from the /p/ sound, even though these sounds are in many ways similar. On the other hand, there can be great variation in the pronunciation of a word without altering the phonemes. The word /bit/ can be pronounced loud or soft, fast or slow, with rising or falling intonation, and it still consists of the three phonemes /b/, /i/ and /t/. The task of speech recognition is to somehow pick out the constant phoneme sequence from this vast sea of variability.

Any project involving speech needs some way of referring to phonemes. Figure 6 shows the phonetic alphabet used by the DARPA Speech Project. Note that this phonetic alphabet covers only the sounds of English, and thus cannot be used for all languages. There is an International Phonetic Alphabet (IPA), but it is not used here because it contains symbols that are difficult to read for the non-linguist.

vowels		consonants		consonants	
example	symbol	example	symbol	example	symbol
beat	[iy]	wit	[w]	church	[ch]
bit	[ih]	which	[wh]	judge	[jh]
bet	[ey]	reel	[r]	bottle	[el]
bat	[ae]	let	[l]	bottom	[em]
but	[ah]	met	[m]	button	[en]
bought	[ao]	net	[n]	Washington	[eng]
boat	[ow]	sing	[ng]	butter	[dx]
book	[uh]	hat	[hh]	pen	[p]
beauty	[ux]	Leheigh	[hv]	ten	[t]
bird	[er]	fat	[f]	kick	[k]
buy	[ay]	vat	[v]	bat	[b]
boy	[oy]	thick	[th]	dad	[d]
diner	[axr]	that	[dh]	get	[g]
down	[aw]	sat	[s]	you	[y]
about	[ax]	zoo	[z]		
roses	[ix]	shoe	[sh]		
cot	[aa]	measure	[zh]	(silence)	[-]

FIGURE 6 The DARPA Phonetic Alphabet for English

For the Verbmobil project it is important that a standard phonetic alphabet be adopted, and it would be convenient if that alphabet had an all-ASCII representation, like the DARPA alphabet.

Figure 7 shows a recording of two speakers saying the phrase "the moon at noon" in a normal speaking voice. Each graph shows energy plotted against time. Although it is difficult to read such diagrams, this pair displays three points.

First, even though the phrase is four words long, there is only one point in each diagram where there is a discernible pause. In each speaker it occurs just to the right of the 0.762 second mark, and it denotes the stop in the /t/ of "at."

The second point is that there are obvious differences between these two speakers; the challenge of speech recognition is to find the commonalities and ignore the differences.

The third point is that even though the vowel /u/ appears in both "moon" and in "noon," it does not mean that the same pattern is repeated, even for an individual speaker. It turns out that a coarticulation effect known as fronting alters the shape of the /u/ in "noon," as compared to "moon."

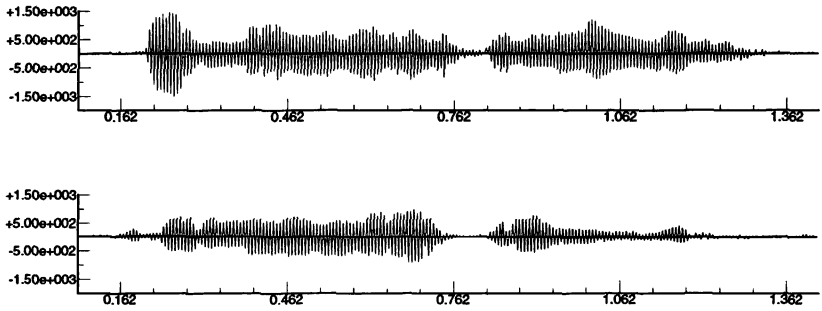

FIGURE 7 Two Speakers saying "the moon at noon"

3.2.1 Constraining the Task to Make Recognition Easier

Recognizing arbitrary speech is difficult for a computer, but there are a number of factors that can be controlled to make the recognition task easier. By imposing limitations on the speaker, recognition can be improved, although the range of information that can be conveniently communicated will go down. Whether this is significant or not will depend on the application. The five factors that can be most easily controlled are:

3.2.1.1 Kind of Speech

One way to make recognition easier is to insist that the user pause between each word. This is called *isolated word* speech and is to be distinguished from *continuous* speech. An intermediate category, *connected* speech was used to refer to systems that used isolated word techniques to achieve a limited ability to recognize words strung together. This term is no longer in widespread use.

The pause between words helps for three reasons. First, it makes the segmentation problem trivial: each word is segmented from the next by silence. If the subjects in Reddy's "In mud eels are" experiment had been listening to words spoken in isolation, they would have had no trouble in understanding them.

Second, isolated words also help with the variability problem. In continuous speech, the beginning of a word is pronounced differently depending on the end of the previous word. For example, when the word "what" is followed by a word beginning with a vowel, the final /t/ will often change to a flap. That is, "what if" will be pronounced [w ah dx ih f] instead of [w ah t - ih f]. But isolated words are pronounced in

a much more consistent fashion each time; spoken in isolation, "what" will always come out as [w ah t].

Third, the pause makes the response time faster, because there is less computation to do. If the vocabulary size is n words, then an isolated word system has to consider only n possibilities, while a continuous speech system will, in the worst-case, consider k^n possibilities, where k is the number of segments. In practice dynamic programming and beam search techniques make the search much faster than the exponential worst-case, but isolated word systems will still be faster.

The main drawback of isolated word speech is that it is unnatural, and therefore difficult to produce and annoying to listen to. It is also slow. Isolated words are spoken at about 70 wpm, or 40% of the 170 wpm rate for continuous speech. Finally, the variability imposed by continuous speech is not all bad. For example, the differences in stress that mark the prosody of continuous speech can be helpful in syntactic disambiguation. But prosody information is lost in isolated word speech.

In conclusion, while isolated speech recognition may be useful for some limited applications, the Verbmobil project will require continuous speech, and therefore we recommend against doing any research on isolated word speech.

3.2.1.2 Speaker-Dependence

Each speaker pronounces words differently, because of his or her unique vocal tract, and because of dialect differences. A system that is tuned to recognize the speech of a particular speaker is called a *speaker-dependent* system. In applications like Verbmobil it makes sense to take advantage of speaker-dependent training. But in other applications, such as a telephone information system, speaker independence is required, because it would be inconvenient for the user to spend time training the system. DECIPHER, which is representative of modern HMM-based systems, takes 2 or 3 hours to train on a new user.

Even when there is sufficient time and resources to train a system on an individual speaker, speaker independent systems will have a certain advantage in that much more data can be used—all the data from every speaker who ever read samples of the domain language, as opposed to samples from one speaker only. Currently, the advantage of more data is not enough to offset the specificity of a speaker dependent system, and they tend to have error rates that are 1/2 to 1/3 of speaker independent systems.

The best system would make use of both speaker dependent and speaker independent sources. That is, it would start with a general

model of speech applicable to any speaker, and then *adapt* the model to each individual speaker. This is what humans are able to do—it can be difficult to understand someone with a strong accent at first, but it becomes easier with practice.

3.2.1.3 Signal Quality and Noise

Speech is rendered more difficult to recognize if there is additional noise in the speech signal. Noise can be introduced either by the environment (such as on a loud factory floor), by other speakers talking simultaneously, or by the communication channel itself, as in a poor telephone connection. Current progress in microphones is reducing some of these difficulties. One technique is to use a second microphone to measure ambient noise, and subtract that from the voice signal. Another technique is to use an array of directional microphones to tune in to one individual speaker in an auditorium full of people. In some cases, such as over a telephone, the system has no control over the original signal, and it is not possible simply to switch to a better microphone. The telephone bandwidth is limited to about 3,400 Hz, while higher quality audio systems are in the 10,000 to 20,000 Hz range. There is very little advantage to maintaining bandwidth above 10,000 Hz.

It may be possible to use additional information besides the actual sound signal. For example, a camera focused on the lips could be helpful. Similarly, jaw movement is a good clue for detecting syllable boundaries and vowel height, so sensors giving the jaw location might be helpful.

3.2.1.4 Vocabulary Size

In general, the more words there are to choose from, the harder it is to distinguish one word from another. However, the word count alone is not enough to fully describe the difficulty of a task. For example, recognizing the twenty-six English letters is quite difficult. Nine letters (BCDEGPTVZ) have the same vowel (/i/), and thus differ only in the initial segment. The groups MN, YI, FSX and AJK are also easily confused with each other. On the other hand, the ten digit names in English are all dissimilar from each other. Thus, the percent correct for the 36 word letter+digit task is likely to be higher than for the 26 word letter task.

Vocabularies of up to about 100 words are considered small, and can be handled quite well by many commercial systems. Most of the current state-of-the-art systems handle a vocabulary of about 1,000 words. Larger vocabularies are handled only by a few research efforts, such as INRS's 86,000 word phonemic-based recognizer and IBM's 20,000 word TANGORA system.

3.2.1.5 Task and Language Constraints

Some systems have a large overall vocabulary, but interact with the user in such a way that the number of words that can reasonably appear in a given context is limited. This can happen because of the task: if the system has just asked "How many do you want?" it can expect an answer that is a number or perhaps a phrase like "a few," but it is not likely that the user will say "purple." Limitations can also occur for syntactic reasons: following the word "the" the system can expect a proper noun or perhaps an adjective, but not a verb. These are not really problems in speech recognition, but rather in language understanding. It is important to make a distinction between the low-level speech recognition and higher-level processing, while still accepting that there can be interaction between levels.

Although there are many possible architectures for interfacing the lower- and higher-level processing, most speech recognition systems have used higher-level processes to limit the possible words that can appear at a given point. For example, in an airline reservation task, a full syntactic/semantic analysis might determine that the only words that can follow the string "Show me the flights from Boston to ..." are names of locations, such as cities or airports. A less restrictive grammar might only consider the context provided by the last word, "to," and determine that the following word could be any noun or determiner.

The term *perplexity* is used as an information-theoretic measure of the average number of words that can appear at any point. If all words are equally likely at any point, then a vocabulary of, say, 100 words will have perplexity 100. But with the same 100 word vocabulary, if there is always one word that is 99% likely while the other words are .01% likely, then the perplexity is only 1.1. The logarithm of perplexity is called the *entropy* of the language.

It has been observed Kubala et al. 1988 that for a given system and a given test data set, the error rate is roughly proportional to the square root of the perplexity. Suppose the constant of proportionality were 1 for some system. Then for a 1,000 word vocabulary with no higher-level grammar support we would expect about $\sqrt{1,000} \approx 31.6\%$ error. A typical word-pair grammar would give a perplexity of 60 and an error rate of $\sqrt{60} \approx 7.7\%$. A tightly-constrained grammar might have a perplexity of 20 and an error rate of $\sqrt{20} \approx 4.5\%$. In fact, these error rates are very close to the values obtained by the best speaker-independent continuous speech systems.

It is clear that a good restrictive grammar can help speech recognition enormously. However, an overly restrictive grammar causes more

harm than good, because it will either force the user to use awkward, unnatural utterances, or it will reject perfectly sensible utterances.

3.3 The Technology of Speech Recognition

Language understanding is a difficult task, largely because of the uncertainty in the message being conveyed. There is uncertainty as to the intended meaning, uncertainty as to the words that make up the meaning, and uncertainty as to the sounds that make up the words. Speech recognition researchers deal with this uncertainty using two sources of information.

The *acoustic model* extracts information from the sound signal. It is designed to answer the question: given that the speaker wanted to pronounce a particular word, what is the probability that this sound signal would be produced.

The *language model* uses information about the words that have previously been spoken to determine what words are likely to be spoken next.

There has been considerable debate about the best structure for an acoustic model, and even more debate about the language model. But the underlying signal processing that extracts digital information from the sound signal is less controversial. While there are still changes and refinements being made, a wide number of systems share a common set of techniques for signal processing.

3.3.1 Signal Processing

Any sound (including speech) is an analog energy source. As such, it is not immediately compatible with digital computers. Therefore, the first step is to convert analog to digital by *sampling* and *quantizing* the sound wave. A typical sampling rate is 16 KHz, meaning that the energy of the speech signal is measured 16,000 times per second. Certain speech sounds have bandwidth up to 10 KHz, although most of the information is below 3 KHz. Shannon's sampling theorem states that the signal can be reconstructed when the sampling rate is twice the highest frequency of the signal, so sampling rates of 6 to 20 KHz have been used. Regardless of the rate chosen, a low-pass filter is applied first to eliminate high frequency aberrations.

The resulting sampled digital representation requires a lot of storage for a small amount of information. For example, if the digitization is done with 32-bit quantization at 16 KHz, then a minute of speech takes up 230 Mbytes, more than the capacity of an average personal computer. Furthermore, the straightforward representation of a sound wave masks its meaning as a unit of speech, in that sound waves repre-

senting the same word will not necessarily have similar digital representations. Therefore, we need a compact and meaningful representation of speech.

Feature extraction is a data reduction technique that results in an encoding of the original sound wave that does not lose much information, and hopefully will be amenable to further computation. The first step in feature extraction is to block the samples into longer units called *frames*. A frame size of 20 msec means that each frame contains 320 samples. If frames were disjoint this would yield 50 frames per second, but typically frames are made to partially overlap.

One simple way to extract features is to run each frame through a bank of *filters*, each designed to capture the energy in a particular frequency range. A typical use would divide the frequency spectrum from 100 to 8,000 Hz into 16 bins and measure the energy in each bin. With 16 bins/frame and 320 samples/frame, this gives a data reduction ratio of 20/1. Assuming the original energy levels were 32 bit numbers, this works out to 25,600 bits/sec. The resulting signal has lost information; something closer to 60,000 to 200,000 bits/sec are required to enable a high-quality resynthesis of the original signal. However, if all we are interested in is recognizing the words, this reduced amount of information is sufficient.

Speech has a wide dynamic range, but the resolution needed to distinguish high-frequency sounds is much less than for low-frequency sounds. Thus, it makes sense to use a logarithmic encoding of energy that requires fewer bits but gives less precision in the high-frequency ranges. Some systems use a *mel frequency* filtering scheme where the spacing of each bin is linear at low frequencies (up to 1,000 Hz) and logarithmic at higher frequencies. This arrangement reflects the response of the human ear Schroeder 1977.

The output of the filtering process is a 16-element vector of energy levels for each frame. Dealing with 16 numbers can be unwieldy, both because of the space required to store the numbers and because certain algorithms work better on discrete symbols rather than continuous numbers. Both problems can be addressed by *vector quantization*. The idea is to build a *vector codebook* of prototype vectors that are representative of a wide variety of speech. A typical implementation might use 256 vectors. Then when each new frame is encountered, it is mapped onto the index of the codebook vector that matches it most closely. That way each frame is represented by a single byte (one of 256 numbers). This yields a data reduction ratio of 64/1, and 1280/1 for the combination of sampling and vector quantization. It has been

shown Rabiner et al. 1984 that vector quantization introduces little or no inaccuracy, if the codebook is well-chosen.

Filtering is a reasonable way to summarize the information in any sound signal. However by using properties of speech that set it apart from other sounds, it may be possible to do better. In order to do that, we have to understand how speech sounds are produced.

3.3.2 Properties of Speech

The human vocal tract can be thought of as an acoustic tube between the vocal cords and the lips, with another tube, the nasal tract, that can be connected or separated by movement of the velum. Sound is produced when air is propelled by the lungs through this tract. The total sound can be thought of as a combination of the air source and a transfer function imposed by the vocal tract. The source can consist of either periodic or aperiodic vibrations, or both.

Periodic vibration of the vocal cords produces the vowels and sonorant consonants, which are characterized by a narrowing of the vocal tract. They are further divided into the nasal consonants /m/, /n/ and /ng/ (where air passes through the nasal cavity) and the oral consonants /l/, /r/, /y/ and /w/. Sounds that contain this periodic vibration are called voiced sounds; all others are called unvoiced.

Aperiodic sounds are produced by a full or partial obstruction in the vocal tract. Voiceless stops, such as /ch/, /p/, /t/ and /k/ are produced by closing the vocal tract and then suddenly releasing the built-up air pressure. They are sometimes referred to as plosives. Voiceless fricatives are produced by the turbulent flow of air past a restriction somewhere in the vocal tract, such as in /s/, /f/ and /th/. In an /f/ the restriction is between teeth and lips, while in an /s/ it is between the tongue and palate.

Some sounds are made by combining the periodic and aperiodic sources to yield voiced stops, such as /j/, /b/, /d/ and /g/, or voiced fricatives, such as /z/, /v/ and /dh/.

The rate at which the vocal cords vibrate in voiced sounds is called the *fundamental frequency*. Spectrograms will show an energy peak at this frequency. Because the vocal tract is roughly a tube which is open at one end closed at the other, resonances will be set up inside the tube. Resonant overtones or *formants* give rise to energy peaks at higher frequencies. The exact frequency depends on the shape of the individual vocal tract, but typically the fundamental frequency ranges from 50 Hz to 250 Hz in men, and up to 500 Hz in women. The F1 formant is at about 500 Hz, F2 at 1,500 Hz and F3 at 2,500 Hz. Higher

frequency formants help identify the voice of the speaker, but do not contribute to the understanding of speech.

The production of speech sounds can be modeled by a combination of two mathematical functions, one representing the regular, periodic voiced sounds and the other representing the random unvoiced sounds. These two functions are then influenced by a transfer function representing the shape of the vocal tract to produce the final result: fundamentals plus overtones. Mathematically, the periodic and random signals are added and then convolved with the transfer function representing the vocal tract to yield the signal. By applying a Fourier transform and taking the logarithm, the convolution of these two sources can be separated into two additive components in a form known as the *cepstrum* of the input. The cepstrum makes it easy to pick out the fundamental frequency and to distinguish voiced from unvoiced segments.

Linear predictive coding or LPC is a family of related spectral representations which attempt to model speech by estimating parameters of this underlying model. The basic idea is that speech is relatively constant from one millisecond to the next. Therefore it can be approximated well by a linear combination of the past few samples. A vector of LPC coefficients represents the most probable coefficients that would have generated the signal, under the assumption that they were generated by a process corresponding to the model. To the extent that the model is accurate, this will give accurate, usable numbers.

LPC coefficients are often vector quantized, just like the raw frequency measures. In addition, many systems compute the *delta cepstrum*, the time derivative of the cepstral vectors. For example, the DECIPHER system uses four features for each frame: the mel-cepstra, the derivative of the mel-cepstra, the raw energy and the derivative of the energy. Each of the four features is vector quantized by a separate 256-element codebook.

While LPC is an attempt to model the gross features of speech (periodic and aperiodic energy sources), there have also been attempts to create finer-grained models. For example, Wu et al. 1987 describe an acoustic simulation that models the length and cross-sectional area of the vocal tract. Such models are more commonly used for speech synthesis than recognition, but there is some promise that high-quality recognition could be accomplished using a detailed model as a guide.

3.3.3 The Acoustic Model

The acoustic model is responsible for extracting information from the sampled, feature-encoded signal. Linguists have a great deal of experi-

ence in describing how words are constructed from phonemes, and how phonemes can vary in different contexts. Early speech recognition systems attempted to formalize this knowledge. Unfortunately, the task proved to be very difficult, and modern systems rely less on linguistic rules and more on statistical patterns.

While there is wide-spread agreement on at least the basics of signal processing, there is no single consensus on acoustic modeling. The differences of opinion will be explained under the four approaches in the next section.

3.3.4 The Language Model

The language model is responsible for extracting information from the words that have been previously spoken and the context of the utterance. That is, given a sequence of words that have been previously recognized, it determines which words are likely to occur next.

In this chapter we concentrate on the acoustic model, since the remainder of the report is primarily concerned with the language model. However, it is important to keep in mind that the two models must communicate. This can be done in several ways. Some applications use a strict artificial grammar as their language model, and thus are able to use it as a generator of possible words, from which the acoustic model can choose the most likely. Other systems use the acoustic model as a generator of the n best words, and use the language model as a filter to choose among these. Still other systems integrate the two models more thoroughly, and choose the most likely word taking both sources of information into account simultaneously.

3.4 History and Taxonomy of Approaches

Speech recognition systems can be classified into four basic approaches:

1. The template-based approach.
2. The knowledge-based approach.
3. The stochastic-based approach.
4. The connectionist approach.

The earliest systems (1970-75) adhered to a *template-based* approach, where each word was stored as a template in the same form as the actual speech input. Recognizing the input then consists of finding the template that matches most closely. The simplest systems are really doing *sound recognition*, as there is little about the templates or the operations on them that is unique to speech. More sophisticated systems make use of speech-specific features like LPC, and have a means for dealing with time-variation.

The *knowledge-based* approach (1975-present) admits more complex models that take into account specific facts about the difference between speech and arbitrary sounds. Unfortunately, the current state of knowledge is insufficient to do high-quality speech recognition from first principles.

There is a parallel between speech recognition and the broader problem of Machine Translation: much of speech recognition can be done just by looking at one word at a time, using the acoustic model alone. However, occasionally the speech signal is ambiguous, and we need to invoke the language model to get clues about the syntax, semantics or pragmatics of the utterance that will help disambiguate the signal. It is particularly hard to recognize function words like "it," "to," "a," and "the." They are short, many of them are similar to each other, and they are often spoken quickly and joined with adjacent words. But the language model can make good predictions about when an article is expected as opposed to when a preposition is expected, and can sometimes make predictions about which preposition is expected.

The *stochastic-based* approach (1980-present) uses various probabilistic techniques to determine the most likely interpretation. The stochastic approach has two advantages over the template and knowledge-based approaches: it provides a concrete model of our ignorance of the true task, and it offers an algorithm to learn improved parameters from individual examples rather than from general rules. The most popular stochastic method is the *hidden Markov model* or HMM, which will be explained below. The best systems in place today use HMMs augmented with some knowledge sources. It has been found to be more effective to add knowledge to a stochastic system than to add *ad hoc* stochastic features to a knowledge-based approach.

Finally, the *connectionist* approach (1985-present) also learns from examples. Because connectionist models are less constrained, they are more difficult to train, and current connectionist systems are not up to par with the best stochastic systems. However, the relative freedom of the connectionist models may turn out to be an advantage in the long run: if improved learning algorithms can be found, they have the potential of performing better than stochastic systems that are forced to make very limiting assumptions. Most current connectionist systems are actually hybrids that use HMMs or similar techniques for much of the task, and employ connectionist nets for subtasks such as estimating parameters.

The four approaches will now be discussed in turn in more detail.

3.4.1 Template Based Approaches

Template-based approaches are adequate when the task involves isolated words with a small vocabulary. Their appeal is their simplicity: each word is represented in a canonical, straightforward form as a template. But this simplicity is also the biggest drawback in more complex applications. The templates can be easily matched against inputs, but there is no way to break a template down into its components, or to consider how a template might be altered by its neighboring words. Since the templates typically represent words rather than shorter units, it is difficult to deal with connected speech, where the word boundaries are not obvious.

The template-based approach consists of three steps:

1. Extract a set of features from the speech signal.
2. Compare this feature set to each of the stored templates.
3. Choose the template that is closest to the feature set.

One reason why the approach survived as long as it did is that each step is independent of the others. Thus, when a new algorithm for, say, choosing the closest match is developed, it can be used by template based systems without having to alter the other two components.

3.4.1.1 Feature Extraction

A number of different features have been used by template-based systems. (Many of them were discussed in section 3.3.1.) The energy and zero crossings in different frequency bins provide a simple measure. Linear predictive coding (LPC) is also used by many systems.

3.4.1.2 Template Similarity Measurement

The second step in the template matching model is to compare the features of the input with the features of each stored template. A naive comparison would just compare the features in the first frame of the template with the first frame of the input, and so on. For each frame the difference between the input and template features would be compared, and the overall distance would be computed as the sum of the individual differences (or perhaps the sum of the squares, or some other aggregate function).

Unfortunately, matching corresponding frames directly is a poor way to match because there can be wide variation in speaking speed. The simplest way to align the input with the template is *linear alignment:* if the template takes up .5 seconds and the input takes only .25 seconds, then match every slice of the input against every second slice of the template. In general a linear alignment algorithm matches first against first and last against last, and interpolates linearly in between.

Linear alignment works well only when speech is uniformly speeded up (or slowed down) in comparison to the reference template. In practice this just does not occur, so linear alignment is not practical.

Dynamic time warping (DTW) is a more sophisticated alignment technique that can still find a good match even when, say, the first half of a word is spoken slowly and the second half is spoken quickly. This is not uncommon in speech, as one way to emphasize a word is to draw out a stressed syllable, such as the first one in "unbelievable."

The idea behind dynamic time warping is to consider all reasonable alignments of the input with the template. A "reasonable" alignment is defined as one that is monotonic, and does not have too steep a slope at any point. A monotonic alignment is one that does not go backwards—if frame 100 in the input is mapped to 210 in the template, then frame 101 cannot be mapped to anything before 210. The slope requirement prevents too many frames from being skipped. Again, if 100 maps to 210, then 101 can map to 211 or 212, but it shouldn't be allowed to map to 400, since that would clearly correspond to leaving out a sound, not just speaking quickly.

Dynamic time warping uses dynamic programming techniques to efficiently find the alignment with the highest similarity, without having to explicitly compute all possible alignments. In effect, only the alignments that could possibly be better than the best partial alignment found so far are considered. There are algorithms to do this very quickly. The DTW approach was introduced by Sakoe and Chiba 1971.

One drawback of DTW is that it maps both long and short segments onto the same target, thereby losing duration information. At times this duration information can be useful in determining the identity of neighboring segments, and it can also carry important prosody information. DTW cannot recover this information.

3.4.1.3 Decision Making

The third and final step of the template-based approach is to choose the best match (or matches) from among the candidate templates. The obvious solution is simply to choose the template with the smallest total difference from the input. This is in fact what is done in many systems.

However, it can be worthwhile to consider more than one of the top-scoring templates. This is particularly true when a particular word has more than one template associated with it. Imagine a speaker-independent system that has ten templates for each word, corresponding to ten different speakers. If the top-ranked template is for word x but the second, third and fourth-ranked templates are all variants of the word y, then y may be the better choice. We shall see that it

may be best not to decide on a single word at all, but rather to pass along a probability distribution from which another program (such as a parser) can select the words it thinks are most likely, taking other knowledge into account.

NTT (see Itakura 1975) was one of the first to use dynamic time warping. LPC was used to produce one template for each of 200 words. The system achieved 97.3% accuracy in isolated word, speaker-dependent recognition with perplexity 200.

To adapt the template-based approach to speaker-independent applications, it is necessary to have multiple templates. Rabiner et al. 1979 describes a system where 100 different speakers produced a template for each word. These were then clustered together by statistical techniques to obtain a smaller number of templates that span the range of speaker variation. The system achieved speaker-independent accuracy of 79% on a 39-word vocabulary.

Sakoe 1979 showed that the DTW technique could be extended from isolated words to a limited form of connected words. His speaker-dependent system could recognize strings of up to four connected digits with 99.6% accuracy.

3.4.2 Knowledge-Based Approaches

The *knowledge-based approach* is an attempt to use the experience of expert engineers, linguists and phonologists to form a better model of speech and thus a better recognition algorithm. Knowledge-based approaches tend to be the most complicated. Instead of having a single uniform representation of each word, knowledge-based approaches admit a vast array of different representations concerning phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. The difficulty in the knowledge-based approach is three-fold. First, there are things about the speech process that remain unknown. Second, even in areas that are understood, it is difficult to get experts to make their knowledge explicit in a form that can be put into a computer model. Third, even when there is a good model of expert knowledge for one area (such as syntax) it is difficult to combine this with knowledge from other areas.

In theory, the knowledge-based approach subsumes the others, since any kind of representation can be considered a source of knowledge. But in practice the systems that have relied primarily on expert knowledge have performed worse than stochastic pattern recognition systems relying primarily on a uniform representation whose parameters are tuned by training data. This reflects the incomplete state of current knowledge of speech and the recognition process. However, just be-

cause the best systems in existence today use rather shallow knowledge does not mean that knowledge sources should be ignored. To the contrary, there is evidence that today's systems will not scale up to larger vocabularies without significant additional knowledge.

It just so happened that the most successful knowledge-based systems of the early ARPA speech understanding project were the CMU systems that produced a single best-choice of words as their output. The BBN group produced a lattice of words with their probabilities as their output. Although BBN did not fare as well in word recognition percentage, their approach can often lead to better overall performance on the part of an integrated speech/language system. For example, even if the BBN system got 0% words correct, if the true word in each case were the system's second or third pick, and if the difference in probabilities were small, then the language modeling component may be able to recover from the speech systems shortcomings, and come up with a final interpretation that is 100% correct.

Until the mid-1970's, the best speech recognition systems identified phonemes with only 60% accuracy. This poor performance led to speculation that perhaps it was impossible to recognize phonemes without using higher-level knowledge as an aid. The ARPA speech recognition project of the 1970's stressed knowledge-based high-level processing over low-level signal processing. As a result, the systems created under this project advanced the state of the art in natural language processing, and were relatively good at making predictions in a constrained domain. However, they were not particularly strong at the task of actually recognizing individual words.

Among knowledge-based systems, perhaps the most well-known is HEARSAY-II, developed at Carnegie Mellon University under the ARPA speech project (see Lesser et al. 1975 and Erman and Lesser 1980). It was based on the assumption that low-level signal processing was going to be inherently error-prone. Thus, a significant effort was made to find a flexible control structure to combine evidence from different knowledge sources. The result was the *blackboard* architecture. In this architecture there is a single shared global variable, known as the blackboard. It is partitioned into several levels representing phrases, words, syllables, and phones. The system operates by a series of hypotheses and test steps: First a hypothesis that, for example, the word "today" occurred during the 100 to 600 msec time frame. Second, this hypothesis would be considered in relation to other hypotheses at the same level ("total" occurred from 100 to 500 msec) or at other levels (/p/ occurred from 100 to 120 msec). Constraints among and between lev-

els would be propagated until the best overall interpretation is found. HEARSAY was successful in that it came close to meeting the stated goals of the ARPA project, but it did not lead to a useful recognition system, and was very reliant on the low perplexity of the task at hand.

A series of experiments by Ron Cole (see Zue and Cole 1979) challenged the assumptions behind HEARSAY by testing the ability of a trained researcher, Victor Zue, to read spectrograms. He was able to identify between 81% and 93% of the phonemes by visually inspecting the spectrograms. The test materials included nonsense sentences such as "Wake jungle gasoline sudden bright," so Zue could not have been relying on syntactic or semantic knowledge. This demonstration served as an existence proof that quality phoneme-based recognition could be done without resorting to the kinds of syntactic and semantic knowledge that HEARSAY required. However, it does not mean that the task is easy. So far, no other human has been able to duplicate Zue's virtuoso performance, although others have had partial success. For example, Ron Cole (see Cole et al. 1986b) was able to read spectrograms of the 36-word alphadigit vocabulary with 98% accuracy.

There are many knowledge sources that can be used to improve recognition. We will start with the *phonotactic constraints*, which limit the classes of phonemes that can co-occur. Phonemes can be classified into broad categories based on their manner of articulation. For example, the phonemes /p,t,k,b,d,g/ all belong to the class of stops, which will be denoted [STOP]. Now suppose that we could identify the class of each phoneme without determining its exact identity. In the subset of English represented by the Merriam Pocket Dictionary, almost one-third of all words are uniquely identified just by the sequence of classes. The average equivalence class size is about 30, and the maximum size is about 200. In other words, this broad classification cuts the perplexity from 20,000 down to about 30 in the average case, and 200 in the worse case. Similar results have been shown in studies of German, Italian, French and Swedish. As another example, the following is a list of the words that match the class sequence

(1) [STRONG-FRICATIVE] [STOP] [HIGH-FRONT-VOWEL] [STOP]

skate	spade	stake
skid	speak	state
skip	speed	steed
skit	spit	

From this we see that the only possible initial consonant is /s/, so it would be wasteful to spend more computational resources on deter-

mining the exact identity of that consonant. Similarly, out of all the possible stops, the only possible ones in English are the unaspirated voiceless stops /k/, /p/ and /t/, so computation should concentrate on distinguishing between them.

The stress pattern of polysyllabic words can be used as another knowledge source to reduce the applicable vocabulary. For example, the word "classify" can be analyzed as a stressed syllable followed by a reduced one and an unstressed one. When both articulation class and stress patterns are taken into account, the average size of the equivalence class for polysyllabic words is only 8. This is a remarkable reduction, but note that monosyllabic words are the most common, and that it is possible to pronounce a word with a non-standard stress pattern to achieve certain stylistic effects.

Cole's FEATURE system was based on the lessons he learned from his intensive study of spectrograms. The system picked out features of the input such as the frequencies of the first three formants, the duration of aperiodic energy before and after vowels, and the ratio of high frequency to low frequency energy. A Bayesian classification method was then used to construct the optimal decision tree based on the selected features. The system achieved 89% accuracy on the alphabet task.

3.4.3 Stochastic-Based Approaches

The *stochastic-based approach* can be seen as an extension of the template-based approach using more powerful mathematical tools. The underlying idea stems from an admission that we do not have a perfect model of speech. Therefore, we would like to have a model that is flexible enough so that we can recover from the ill effects of any faulty assumptions. The stochastic-based models are all parametric models where the parameters can be estimated by comparing the performance of the system on known test data with the desired results. A good stochastic model is one with a number of parameters that is small enough so that they can be estimated from training data in a reasonable amount of time, and large enough to cover the variation from the training data in the actual inputs to the system.

The most commonly used mathematical formalism is the *hidden Markov model* or HMM. A Markov model is a network of states where each state has two probability distributions associated with it: the *emission probabilities* determine what symbol to emit (output), and the *transition probabilities* determine what state to go to next. Over a series of time steps a Markov model will emit a series of symbols and change from one state to another. It is called a hidden Markov

model because the state of the system remains unknown to the outside observer; only the emitted symbols can be observed.

A Markov model is a kind of finite-state model, because there are only a finite number of states in the whole system, and at each point the only thing that determines what to emit and what state to go to next is the identity of the current state. In other words, Markov models are memoryless—if the system is in state s_3 at time step 3, then the next output and the state at time step 4 depend solely on the properties of state s_3, and can not depend on the state of the system or its output at time 2, 1, or any earlier time. Thus, Markov models are very restrictive—there can be no complex interaction between outputs that are separated by more than one time step, unless they are explicitly modeled by distinct intermediate states.

Figure 8 shows a hidden Markov model for a single phone. It consists of three states corresponding to the onset, middle and offset of the phone. The possible transitions to new states are shown with arrows. The sum of the probabilities on the arrows should sum to 1, but the numbers are not shown. In addition there should a probability distribution of outputs associated with each state. These outputs represent the speech signal features in a given frame, or more often vector quantizations of those features. Notice that each state has a loop: a state transition to itself. This allows a state to stretch across more than one time frame.

Figure 9 shows a model for the word "have." This model is depicted so that emitted symbols are phonemes, but this should be considered schematic for a topology where each circle is replaced by a complete phone model, as in Figure 8. The word model is complex because it admits four different pronunciations for "have": [hh ax v], [hh ae v], [ax v] and [ae v]. The latter two pronunciations occur when the word is spoken quickly in an unstressed position.

The next step is to construct a model of the whole vocabulary. Figure 10 shows how a continuous speech recognizer can be built from a list of separate word models. This configuration has no grammar constraints; if syntactic rules are known then a more restrictive model can be built, although it must still be finite-state. That means that non-local constraints such as subject-predicate agreement cannot be handled within the HMM formalism, but must be modeled instead by a more powerful formalism such as a unification-based grammar.

Instead of using dynamic time warping, HMM systems account for the variability in the length of phonemes with loops in the Markov model. Some systems also employ variable-frame-rate compression. This technique collapses strings of identical frames into a single frame.

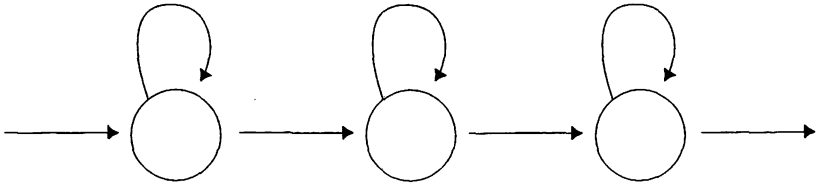

FIGURE 8 Hidden Markov Model for a Single Phoneme

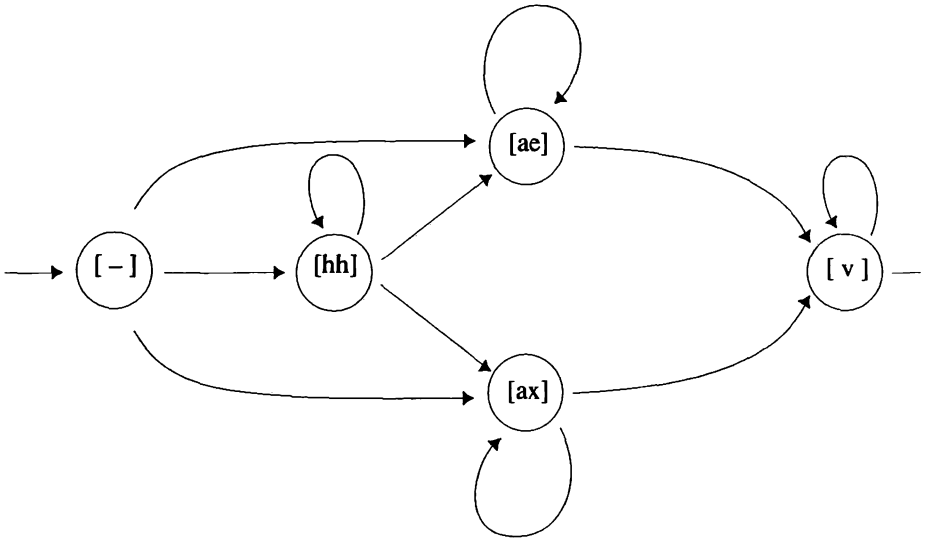

FIGURE 9 Hidden Markov Model for the word "have"

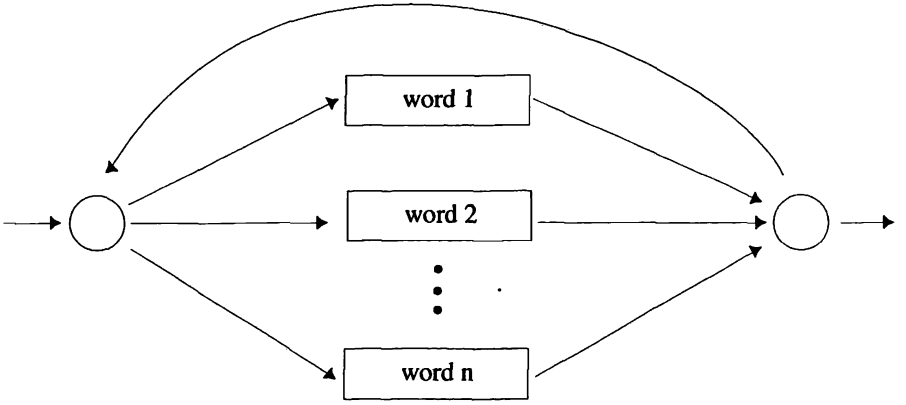

FIGURE 10 Hidden Markov Model for a Large Vocabulary

Of course, the chance that two frames will be exactly identical is minuscule unless the system is using vector quantization.

Once the topology of a HMM model has been determined, the next step is to determine good values for the model's parameters (the emission and transition probabilities). Setting these parameters by hand would be a tedious task, and the chances of arriving at a good model by trial and error would be small. Fortunately, there is an algorithm for automatically arriving at a good set of parameter settings, given only a sequence of training sentences. This algorithm is known as the *forward-backward algorithm.* It is not guaranteed to find the best possible parameter settings, but it is guaranteed to improve the value of an initial setting (or at least make it no worse). The algorithm can be applied iteratively until no more improvement is made. The best part about the forward-backward algorithm is that it in effect does segmentation automatically. By that we mean that the training data needs to consist only of a set of pairs of spoken sentences (in quantized form) with their transcriptions. The algorithm will determine which part of the speech signal corresponds to which part of the transcription. It is relatively easy to acquire a large corpus of training data in this format. Other formalisms require data where a human laboriously labels each segment of the speech signal with a phoneme label. Such a corpus is much more difficult to obtain. See Baum 1972 for more details on the forward-backward algorithm.

It is the effectiveness of the forward-backward algorithm that has led to the wide-spread adoption of HMM models. Researchers have been willing to go through great contortions to fit all speech knowledge into the HMM framework so that the forward-backward algorithm can be used to pick good parameter settings. One of the most difficult problems is choosing the proper unit of speech. In Figure 10 we saw a model of continuous speech that consists of a network of words, where each word is a network of phones, and each phone is a network of states. If every parameter on every state were considered separately then a huge amount of training data would be required to set all the parameters accurately. On the other hand, if the parameters for a given phone are required to be the same in all occurences of the phone, then there will be far fewer parameters. This means that very little training data will be required, but that there will be no allowance for co-articulation effects. In practice most HMM systems fit somewhere between these two extremes. Some systems use a bigram model, where each pair of phones is required to have the same parameters in every word it occurs in. That is, the [f] in "flat" would have the same parameters as the

[f] in "flop" because they are both followed by [l], but it would have different parameters from the [f] in "fat." Other systems use trigram models where the parameters of a phone depend on the preceding two phones. There are a lot of triphones—for a language with 50 phones there are 125,000 possibilities, of which roughly half will be realized in a large lexicon. To deal with this large number of combinations without requiring huge sets of training data it is desirable to share information across phones. The so-called generalized phone model does just that. It starts with separate models for each triphone, but then uses a clustering technique to merge information (that is, share training data) for similar phones. For example, the clustering algorithm might determine that [b] and [p] behave similarly in a certain environment, and therefore should be merged.

After the training data has been used to determine a good set of parameter settings, we are finally ready to use the model to recognize a sentence. The *Viterbi algorithm* is used for this problem. Given a speech signal, the Viterbi algorithm produces the state sequence that has the highest probability of being the one that generated the observed signal. It does this by a kind of dynamic programming technique similar to DTW. While this is quite efficient for isolated words and short sentences in small vocabulary models, it can be very time consuming for larger models. In that case, a *beam search* is usually used. A beam search keeps only the best few partial interpretations at each time step, discarding the rest. Most of the time it will be discarding improbable interpretations, thereby saving time at no expense. Sometimes, however, it will discard an interpretation that initially looked bad, but would have turned out to be the best if the full Viterbi algorithm had been used. In general there is a time/accuracy trade-off in choosing the width of the beam. See Viterbi 1967 for more details on the Viterbi algorithm.

HMM systems, like most others, perform better when they can use speaker-dependent information. In practice, it has been shown that networks trained on speaker-dependent data perform 2-3 times better than the same system using speaker-independent training. The difference is much greater for speakers who are further from the norm, such as non-native speakers. Some applications, such as telephone information services, will always be speaker-independent. But for other applications, including Verbmobil, it is quite reasonable to train the system for each user. Modern phone-based systems can be trained with about 20–40 well-chosen sentences, followed by the forward-backward algorithm computation. This computation can take several hours on

large models. Word-based systems would require the new user to recite each word in the lexicon at least once, which adds up to many hours or days of training time for a large lexicon. That is one reason why word-based systems are no longer popular.

A cross between speaker-dependent and speaker-independent processing is offered by the new adaptive systems that have been implemented in recent years. In such a system the parameters of the model depend partly on training data supplied by the current user, and partly on a large data bank of training examples from many speakers. Brown et al. 1983 introduce the use of Bayesian techniques to perform this calculation, and Shikano et al. 1986 shows how adaptation can be achieved through vector quantization by mapping vectors of the input speaker to corresponding vectors of the reference speaker (obtained from the training data). On the more practical side, Murveit et al. 1991 and Kubala et al. 1991 describe the implementation of adaptation techniques in current systems. The difference between Bayesian learning and maximum likelihood learning is that the former requires an estimation of the prior probabilities for each parameter. These priors can come from a speaker-independent model which is then updated by Bayesian learning, or they can come from some other knowledge source.

The basic idea behind Markov models and the basic algorithms on them (such as the forward-backward and Viterbi algorithms) have remained unchanged since their inception, admitting only minor refinements. However, there has been much work involved in adding speech knowledge of various kinds to existing HMM systems. Systems that have taken this approach include SPHINX (Lee 1989), DECIPHER (Cohen 1989) and BYBLOS (Chow et al. 1987). In fact, most of the current systems are based on hidden Markov models. Current systems are covered in section 3.9.

CMU's DRAGON system (see Baker 1975b) was the first to use a uniform stochastic modeling approach. It demonstrated that the conceptual simplicity of a single uniform model more than made up for the lack of certain knowledge sources.

CMU's HARPY system (see Lowerre 1976) combined the uniform stochastic model of DRAGON with some of the knowledge sources used in HEARSAY. It was able to roughly meet the aims of the 1971 ARPA speech project.

3.4.4 Connectionist Approaches

The *connectionist* or *neural net* approach is also based on the idea of acquiring training data and using it to improve the performance of a system. Like a HMM, a connectionist or neural network is a complex arrangement of simple components. The difference is that the Markov restriction makes the HMM simple to analyze mathematically. Connectionist networks have no such restriction. This means that they can potentially provide wider coverage—every HMM is a connectionist net, but not every connectionist net is an HMM. On the other hand, it also means that it is harder to analyze the connectionist net mathematically, and thus harder to make any performance guarantees. Since there is a larger space of models to consider, training may take longer and be less effective. Most research has centered on small-vocabulary isolated word systems. As the newest technique to be tried seriously, the connectionist approach is promising but has not yet delivered a high quality system for a large vocabulary or for continuous speech.

Limitations of the connectionist architecture make it most appropriate as part of a hybrid system, rather than as a completely connectionist system. Connectionists claim that their approach can lead to fast highly-parallel implementations, and hardware that uses analog components and thus avoids some of the problems in quantization.

Given the success of HMM systems, it is worth asking why other approaches are still being investigated. The answer lies in three problems of the HMM approach: First, HMM models make an assumption of conditional independence—that each parameter is independent of the others. This assumption is unrealistic, and can be a source of error. Second, HMM models work on a frame-by-frame basis, and thus have some difficulty with integrating knowledge about segmental features that occur over a series of frames. Third, they have even more difficulty with super-segmental features. For example, if a speaker has a cold, then each segment will have a "nasal" quality. A HMM model can account for this, but in doing so it must in effect consider the probability of the speaker having a cold every frame. It would be better if this consideration were done only once.

The first two problems were addressed by Austin et al. 1991 with a hybrid system that uses a traditional HMM to propose the N-best sentence interpretations, and then using two separate scoring mechanisms, one HMM and one connectionist net, to score each one. The two scores are combined linearly to choose the overall best candidate. It is interesting that while the HMM system alone had an error rate of

9.1% and the connectionist net alone had a much worse error rate of 20.3%, together they had an error rate of only 8.5

Time-delay neural networks or TDNNs are an attempt to deal with the problem of varying duration of segments. HMM models deal with duration with loops that allow one state to span several frames, but this technique has problems due to the finite-state nature of the HMM model. TDNNs provide a more natural way of modeling duration phenomena.

Some researchers have concentrated on duplicating standard algorithms with connectionist techniques. For example, Huang et al. 1988 presents a connectionist version of the Viterbi algorithm. Huang 1991 shows how a multi-layer perceptron net can be used to estimate parameters for speaker adaptation. Nelson Morgan describes a HMM system that uses a connectionist net for one task: calculating the emission probabilities in the HMM model. The claim is that connectionist models are better than standard statistical models that make the assumption that all parameters are independent. This claim is probably true, but it is unfair to the stochastic models, since there are standard methods of estimating parameters without making independence assumptions.

3.5 Outstanding Problems

In this section we identify six fundamental problems that must be addressed. The first four are based on the analysis of Lee 1989. Current speech recognition systems would improve if they could:

1. Use more of the speech signal. Current systems digitize the speech waveform and extract certain features that are easy to use. Because the features that are extracted are minimal, these systems are in effect throwing away useful information. The challenge is to extract all the information that is acoustically important, but to keep the total number of features (and bits) small and manageable.
2. Find the right unit of speech. In small-vocabulary isolated word systems, it is easy to pick a good unit of speech: the word. But in large-vocabulary connected speech systems, the word is a poor choice for two reasons. First, there may not be enough training data to get a good model of each word. Second, words are pronounced differently depending on the preceding and following words. Thus, it is necessary to choose a smaller unit of speech. The phoneme is small enough, but a phoneme is pronounced differently depending on the context, so there has to be some

method of combining phonemes or taking context into account. Diphone, triphone and syllable models have been suggested as compromises. Research continues on this issue, with many modern systems using a combination of units.

3. Utilize known facts about acoustics, phonetics, and lexical forms. Human phonologists know more about speech than they can put into current models, because the models are primitive. We need some way of accommodating specific knowledge without hampering the ability to improve the system with training data. The problem is that some of these knowledge sources cannot be integrated with simple models like the HMM formalism, either because the resulting model would no longer be finite-state, or because it would have too many parameters to be trained effectively with the sparse data available to us. Methods of tying parameters and clustering models can be used to eliminate some of these problems.
4. Adapt to each speaker. It is common for a human to have trouble understanding the speech of someone with a strong accent, but to quickly adjust to the accent and improve recognition. Current systems are just starting to incorporate speaker adaptation. A good example is the BYBLOS system described below. The adaptation techniques in BYBLOS cut the error rate for non-native speakers from 30% to 6% after a 40 sentence training period.
5. Interface better with other modules. There has been an emphasis, particularly within the US DARPA-sponsored projects, to concentrate on word recognition percentages as the measure of a system. In reality, the only important measure is how well an overall language understanding system can perform. A speech recognition component with a lower accuracy may lead to a better overall understanding system, if it is more flexible in the way it integrates with the rest of the system, or if it produces N-best output rather than a single interpretation.
6. Improve Robustness. Current systems perform well in carefully controlled experiments, but degrade badly when the inputs are unusual (for example when the speaker has a cold or a foreign accent); when a different microphone or even a different placement of the microphone is used; or when the speaker uses an unknown word or a novel grammatical construction.

3.6 Speech Synthesis

Unlike speech recognition, which is generally considered to be a serious problem requiring much research before systems adequate to tasks like that of the Verbmobil will be available, speech synthesis is in some respects a solved problem. That is, for a wide range of applications, for a number of languages (including English, Japanese, and Swedish) imperfect but nonetheless acceptable synthesis is currently available. If we restrict our attention to the synthesis of isolated words, there is every reason to believe that the technology used by these systems can be straightforwardly adapted to other languages.

Although for some limited purposes simple concatenation of recorded speech waveforms may be used, the higher quality synthesis systems all make use of some kind of parametric synthesis, generally either formant synthesis or LPC synthesis. These techniques allow the spectral content to be manipulated independantly of amplitude and F0, and require less storage than a system that must store waveforms. Differences between systems center on which sort of parametric representation they use and on the size of the basic unit. The larger the basic unit, the fewer problems arise in accounting for local dependencies, such as coarticulation, but the number of basic units that must be stored rises quickly and soon becomes intolerable. The Verbmobil project will have to choose which particular synthesis approach to use, but all of the better ones seem practicable.

The principal defect of existing synthesis systems is widely acknowledged to be the prosody, in particular the F0 pitch and durations. This is the impression one has on listening to synthetic speech, which sounds "flat" and "mechanical," and is confirmed by studies of synthetic speech. Inadequate prosody not only makes the synthetic speech sound unnatural, but in some circumstances it distorts the information conveyed. Intonation, for example, helps to disambiguate structurally ambiguous utterances, and conveys pragmatic information, such as whether an utterance is a question or a statement. If the prosody is poor, this information may not be available.

Fairly good prosodic rules have been developed for some languages, and a reasonably good general framework for handling prosody now exists, but considerable work remains to be done in working out the details of prosody in particular languages and incorporating good prosodic rules into synthesizers.

3.7 Prosody

The term *prosody* is used in an ambiguous way, sometimes to refer to any phrase-level phonological phenomenon, and sometimes in reference to suprasegmentals, such as F0, pause duration, and duration. In the familiar languages of Western Europe these tend to be highly correlated. On the one hand, with the exception of French *liaison* there are few obvious phrasal phenomena that are not suprasegmental. On the other hand, suprasegmental information, in particular information about F0, is not required in the lexicon in most of these languages, the exceptions being Swedish, Norwegian, and some dialects of Basque.

If, however, we consider other languages, the distinction becomes clear. Non-suprasegmental phrasal rules are not difficult to find. For example, Modern Greek has a complex set of rules for resolving sequences of vowels arising across word boundary (Condoravdi 1990). Although most Western European languages happen not to make use of lexical tonal information, this is atypical—most of the world's languages are tonal, and most human beings speak tonal languages.

The bottleneck lies in two areas. First, we do not know as much as we need to know about how to predict F0, duration, and pause duration from the relevant linguistic information. Second, most existing synthesis systems lack the capacity to accept and make use of the information necessary for high quality prosody.

Doing a good job of prosody requires several things. First, given the phonological phrasing of the utterance, we must have good rules for computing the F0, durations, and pause durations. These rules must take into account a wide range of utterance types and speaker characteristics, e.g. widely varying pitch ranges. Substantial progress has been made in this area over the past 15 years, both in the form of advances in linguistic theory that provide a general theoretical framework, and in the form of detailed descriptions of phenomena in particular languages. This work has led to the development of greatly improved synthesis for some languages. Nonetheless, even for the languages on which the best research has been done, further work is necessary, and for most languages the necessary work has hardly begun. Consequently, it will be necessary for the Verbmobil project to devote some attention to the development of prosodic rules, the amount of work depending on the languages chosen.

Second, the phonological phrasing must be computed from syntactic and discourse structure information. Here again there has been enormous progress in recent years, but detailed algorithms exist only for a few languages, and probably none of them adequately take into

account the discourse structure, in part because work on discourse structure is itself rather primitive. Consequently, it will be necessary to devote research to the computation of phonological phrasing.

The second aspect of the bottleneck, namely the inability of most synthesis systems to make use of the necessary information, has two causes. One is the fact that basic research on prosody has made its way into applications only gradually. A second is that most synthesis systems have been parts of text-to-speech systems.

Text simply does not provide all of the information necessary for high-quality synthesis. In some cases this information cannot be recovered at all. In other cases it can only be recovered by means of complex inferences, some inferences being at the level of full understanding of the utterance. The difficulties include writing systems that do not provide full phonological information (e.g. the failure of English spelling to differentiate the voiced and voiceless interdental fricatives, both of which are spelled /th/, or the failure of Italian to differentiate the tense and lax mid-vowels), the lack of prosodic information, and the difficulty of obtaining the phonological phrasing. The best text-to-speech systems, such as MITalk (Allen et al. 1987), are devoted in large part to the analysis of the text, from which they obtain a crude syntactic parse, and, in the case of English, the morphological information necessary to derive a phonetic representation from the spelling. Since text-to-speech systems lack necessary information, or can obtain it only with great difficulty, the synthesizers incorporated in such systems have in general not been constructed to make use of extensive information about syntactic and discourse structure.

The fact that most synthesis systems are really text-to-speech systems means that, although we recommend making use of existing technology to the extent possible, it is inappropriate to generate text and feed it to a text-to-speech system. To begin with, it is pointless to go through all the work that the analytic component of a text-to-speech system does when the necessary information will be available from the generation system. More importantly, it is unlikely that the text-to-speech system will be able to recover all of the necessary information, or that, if it does, its synthesis component will be able to make use of it.

3.7.1 Interface

These considerations lead us to an important point about the overall system design and organization of the project. It is very important that the generation component generate linguistic representations containing the sort of information required by a synthesizer, including

syntactic and discourse structure, and that the synthesizer be designed so as to accept such input. The interface between the generation system and the synthesizer should be decided early on. While it may prove desirable to modify the interface at some point, it is essential that each of the groups working on these two modules keep in mind the needs of the other, and that it be possible to connect the generator to the synthesizer for testing.

3.7.2 Suprasentential Prosody

One area that is likely to be of some importance to the Verbmobil is that of suprasentential prosody. Most linguistic research on prosody deals with individual sentences, and to our knowledge all existing synthesis systems generate sentences in isolation. Very little is known about suprasentential phenomena, but just enough is known that we can say that they exist and may be significant not only for the naturalness of the speech but also for its correct interpretation.

A clear demonstration of the role of suprasentential prosody in interpretation is provided by the work of Silverman 1987. Silverman constructed an advertisement for windows that had three parts. The first part was a general advertisement for the windows. The second part offered some special, expensive windows. The third part offered free deliver of the windows within the Cambridge area.

In the written text the first part was always a separate paragraph, but in one version the offer of free delivery was made part of the second paragraph, while in another it was made into a separate, third paragraph. When the offer of free delivery forms a separate paragraph, the advertisement is interpreted as offering free delivery for all types of window, but when it is appended to the second paragraph, the advertisement is interpreted as offering free delivery only for the more expensive windows.

On the basis of his observations of the F0 and pause durations in recordings of the two texts, Silverman synthesized the advertisement in versions that differed only in the F0 contours and pause durations. When he played the two versions to subjects and asked them to interpret the advertisements, he found that he could reliably induce one or the other reading depending on the F0 and pause duration parameters.

This experiment demonstrates both that it is necessary to take into account paragraph level structure, and that the choices made may influence the interpretation of the text.

For many applications, a failure to introduce such effects is not critical, but for the Verbmobil, which may well be required to interpret connected discourse, suprasentential prosody is likely to be important.

3.7.3 Evaluation of Synthesis

The correctness of the output of a machine translation system or speech recognition system can fairly easily be tested simply by reading it and asking whether it corresponds to the input. Elaborate evaluation procedures are not called for. In the case of speech synthesis, however, the matter is different. The researcher who knows what the input text was may easily overestimate the intelligibility of the synthetic speech to a naive user, and he or she may become accustomed to synthetic speech and incapable of fairly evaluating its naturalness. He or she is also unlikely to detect failure to disambiguate sentences that are ambiguous without suitable prosodic information. For these reasons, it is desirable to implement explicit experimental tests of the synthesis system.

At least three sorts of test may be made. First, the basic segmental intelligibility of the system may be evaluated, e.g. does it adequately distinguish [p] from [b]? A good summary of segmental intelligibility testing may be found in Logan et al. 1989. Second, the naturalness of the system may be evaluated by asking naive subjects (unacquainted with synthetic speech) to comment on the naturalness of the speech. Their comments on the source of any perceived unnaturalness may provide useful cues for improvements. Finally, the prosody should be evaluated to see if it conveys differences in syntactic structure, scope, etc., by tests like the experiment by Silverman described above.

3.7.4 Recommendations

- For the segmental aspect of synthesis, Verbmobil should build on the existing technology. This means using formant or LPC synthesis, and where available, using the phonetic rules already developed.
- Synthesis research should concentrate on prosody, including the full range of utterance types likely to be used in the area of application, and giving attention to suprasentential domains.

3.8 Voice Conversion

An interpreter, whether machine or human, is needed when the speaker cannot speak in the hearer's language, whether because of lack of knowledge of the language or because it is impossible to speak in the languages of several hearers simultaneously. One way in which even the best human interpreter does not emulate perfectly what the speaker would do if he or she were speaking in the target language, is to duplicate the speaker's voice. In principle, a machine interpreting

system could do this. Since it is clear that, at least to some extent, hearing a voice radically different from that of the original speaker is disconcerting, there has been some interest in giving the synthesizer output the characteristics of the speaker's voice, a process known as voice conversion.

With the exception of research on the synthesis of female voices,[1] there has been very little research on this topic. One of the very few efforts that has been made is the work of Abe et al. 1990 on the mapping of LPC vector codebooks.

Two general approaches to emulation of the speaker's voice are available. The first, and easiest, is to provide a fixed set of voices, minimally a male voice and a female voice. Design options would be for the speaker to select the voice used, for the hearer to make the selection, or for the system to apply a distance measure and make the selection itself.

Much more difficult would be true voice-to-voice conversion, in which the interpreted speech would have the characteristics the speaker's own speech would have if he or she could speak the target language.

Intermediate between the two would be the use of a fixed set of voices with some parameters, such as those controlling F0 range, determined by the input speech.

In the Verbmobil application, it would be possible for each speaker to have a set of pre-determined parameters characterizing his or her speech. These parameters would be shared with the other participants' Verbmobils so that each speaker's voice would be duplicated.

It is important to recognize that it is not entirely clear what it would mean to perform voice conversion. Most discussion of the question has centered on the idea that it is necessary to adapt only to the physical characteristics of the speaker's body, such as vocal tract length, so that a good voice conversion system might be based on estimation of these parameters from the input speech and their use as control parameters in the synthesizer. However, it is clear that there are idiosyncrasies of speech that are due not to the form of the speaker's body but to the speaker's speech style, so estimation of the physical parameters of his or her vocal tract may be insufficient.

A second issue is that it may be more appropriate to translate style at a more abstract level rather than at an acoustic level. If a speaker adopts a "tough" speech style, we may wish to convey the

[1]Most synthesis research has been conducted on male voices.

same impression in the interpretation, but toughness may be conveyed by different physical parameters in the target language.[2]

It is unclear how important it is for the interpreted speech to resemble the voice of the original speaker, given that this is the one property of our ideal system that even the best human interpreter does not have. If the system is extended to multi-lateral conversations, voice conversion may be very desirable, as an aid in allowing the hearer to separate overlapping utterances by different speakers and to identify the speakers. In the bilateral system presently envisaged, this motivation for voice conversion disappears, but it may nonetheless be desirable in order to maximize the naturalness of the conversation. To take an extreme case, it would be disconcerting to engaged in a conversation with a woman and hear a man's voice from the Verbmobil.

3.8.1 Recommendations Regarding Voice Conversion

- Full voice conversion should not be a high priority for Verbmobil, but is likely to be important if the project is extended to multilateral conversations.
- Users are likely to find the system considerably more natural if the synthetic speech is matched to the speaker's voice to some extent, at least to the extent of providing a choice of male and female voices.
- Some research on the importance of full voice conversion might appropriately be undertaken as part of the Verbmobil project.
- If full voice conversion is undertaken, the possibility of pre-encoding the speaker's voice characteristics should be considered, as the computation of these parameters at the time of use may be impractical, due either to the amount of speech necessary for training or to the time necessary for the computation.

3.9 Current Speech Recognition Systems

This section gives a brief overview of the most important research and commercial speech recognition systems currently being developed throughout the world.

[2]Still another issue that arises here is that what is considered appropriate speech differs from culture to culture. An interpreter may be doing the speaker a disservice in emulating his tough speech style if such a style is inappropriate in the target language. This kind of adaptation is a service performed by good human interpreters.

3.9.1 Japan

ATR The ATR Interpreting Telephony Research Laboratory in Kyoto has one of the largest efforts in speech recognition, and is certainly the leader in speech-to-speech translation with their SL-TRANS system. As a large laboratory, they have been able to take several different approaches. The version of SL-TRANS described in Nagata and Kogure 1990 consists of six components. A HMM recognizer produces a lattice of word possibilities, which are pruned by a dependency-based filter operating at the level of words and small phrases. Next an HPSG-based active chart parser produces a lattice of possible parses, which are then rewritten using a transfer module, turned into English sentences by a rule-based generation module, and pronounced by a speech synthesizer. The overall system has been able to translate 69% of the sentences in one test set. The grammar is unusual in that it is designed especially for spoken language, and it includes pragmatic constraints such as the grammatical restrictions on honorifics.

The speech recognition component itself is more accurate than the complete translation system. It can recognize 81.6% of phrases in speaker-independent mode, and 88.4% in speaker-dependent mode (see Hanazawa et al. 1990).

ATR is also experimenting with neural networks for speech recognition. Waibel et al. 1989 and Waibel et al. 1988 explain how a Time Delay Neural Network is used to solve the difficult problem of distinguishing "B," "D" and "G." The network was able to learn by itself some features that are widely acknowledged as important for this task, such as F2-rise, F2-fall and vowel-onset. Overall, the neural net achieved 98.5% accuracy, compared to 93.7% for a HMM model.

NEC Sakoe and Iso and their colleagues have been investigating speaker-independent isolated word recognition using neural networks. In Sakoe et al. 1989, a dynamic programming technique was used to obtain the advantage of dynamic time warping while maintaining the ability to improve performance by learning. 99.3% accuracy was obtained on the Japanese digit task. A similar model by Iso and Watanabe 1990 achieves 99.8% accuracy on this task.

NTT The NTT Human Interface Lab in Tokyo is doing research on isolated word recognition. They have developed a hybrid technique that uses a HMM to match phones and DTW to match pitch contours. Words with similar phonetic patterns such as "kyūryō" and "kyūgyō" might be confused by the phone matcher, but would easily be distinguished by the pitch contour matcher, because the former word always has an accent on the first syllable, and the latter does not.

The Human Interface Lab, in conjunction with the NTT Basic Research Lab, is working on a speech typewriter. It is based on HMM technology, using syllable trigrams as the unit of speech. The system achieved 94.9% accuracy in a speaker-dependent test of 279 utterances having a perplexity of 3.9. A continuous speech recognizer was reported by Matsunaga et al. 1990. It is HMM-based and uses a two-level grammar: a phrase structure grammar for individual phrases, and a dependency grammar to join phrases together. Phrase recognition rates of 86.8% were achieved, building on a word recognition rate of 98.4% on a 216 word vocabulary.

Matsushita In 1985 the Matsushita Research Institute in Tokyo developed a speaker independent isolated word system based on LPC cepstrum coefficients and various knowledge-based sources, including modules for consonant segmentation, consonant recognition, vowel and semivowel recognition, phoneme sequence production and word matching. With a 274 word vocabulary, the system achieved 95.6% accuracy. (See Morii et al. 1985.)

Matsushita has also developed a compact (business card size) board for limited vocabulary (15 words) speaker-independent isolated word recognition. The first application of this technology is as an interface to a VCR.

Sharp The Sharp Information Systems Lab in Yamatokoriyama has worked on recognition using the syllable as the unit of speech. This is an appropriate unit for Japanese, since there are only about 100 syllables all together, and the language is clearly syllable-based. Phonological rules were used to derive the different possible realizations of each syllable, and dynamic time warping was used for recognition. Speaker-dependent recognition rates of 91–94% for a 300 word (isolated word) vocabulary were obtained.

In 1985 Sharp placed a voice-operated word processing system on the market. Since then, the emphasis has been on improving the interface.

Fujitsu The Information Processing Division of Fujitsu Labs in Kawasaki is focusing on large vocabulary recognition. A dynamic programming approach is applied to a 100,000 word lexicon (mostly names). Because this would require extremely long computation, a pre-processing step is used to quickly select the 500 or so most likely word candidates. This pre-processing step appears to perform well, as the correct word is within the top 20 choices 93.5% of the time. However, overall performance is poor, as the system picks the correct word only 55% of the time. See Kimura 1990 for details.

A separate project is developing an English text-to-speech system based on neural nets. A description of the mechanism for detecting phrase boundaries in the text is given in Yamaguchi and Matsumoto 1990.

3.9.2 North America

SRI (DECIPHER) This HMM-based system has undergone a series of refinements to emerge as one of the top-scoring speech recognition systems in the world. The front end uses 25 Mel filters spanning 100 to 6400 Hz to derive 12 Mel-cepstra coefficients per frame. These features are reduced to four for each frame (using vector quantization): the energy, energy derivative, Mel-cepstra and Mel-cepstra derivative. A set of phonological rules (see Cohen 1989) are applied to the base forms of the words to generate a more complete vocabulary averaging 40–75 pronunciations per word, which are then pruned to eliminate all but the most likely pronunciations—typically 4 per word. This pronunciation modeling separates DECIPHER from other HMM systems; see Cohen et al. 1990 for a demonstration of the benefits of this approach.

Once the phoneme-based models for each word are produced, the acoustic model is produced using a combination of biphone, triphone, phone-in-word and context-independent models that are trained using an initial hand-labelling technique augmented by bootstrapping on an unlabelled corpus.

DECIPHER uses tied-mixture HMMs, a technique that enables sharing of training data across phonemes. It was also demonstrated that separating male and female speakers into two separate models increased performance. Each utterance is scored against both the male and female model, and the interpretation with the highest score is accepted.

Currently, the researchers are experimenting with speaker adaptation. The tied-mixture HMM model can be updated to reflect characteristics of an individual speaker. One experiment reduced the error rate from 7.4% without adaptation to 6.1% for adaptation done on 20 sentences of training data on the new speaker. This improvement is not marked enough to make the speaker spend the time to recite 20 sentences, but it is possible that online adaptation could be useful.

Besides the general biphone and triphone models, frequently used words and phrases are modeled separately. For example, the phrases "what are the" and "give me" are modeled explicitly with networks that include the pronunciations "what're-the" and "gimme," respectively.

Overall, DECIPHER achieved accuracy of 95.2% on the DARPA resource management task with a bigram model of perplexity 60, and

82.4% with no grammar (perplexity 1,000). For more information, see Murveit and Weintraub 1988 and Murveit et al. 1991.

SRI is also working with the University of California at Berkeley in developing hardware for real-time speech recognition. Eight special-purpose integrated circuits have been designed, fabricated, and tested. They are used in a board to do HMM computations.

BBN (BYBLOS) Bolt, Beranick and Newman Systems and Technologies Laboratory in Cambridge, MA has had an ongoing project developing the BYBLOS speech recognizer (see Chow et al. 1987, Kubala et al. 1988, and Kubala et al. 1991). The current version of the system is similar to SRI's DECIPHER—it is a HMM-based recognizer which samples Mel cepstra and their derivatives plus energy and its derivatives. It uses tied-mixture Gaussians to obtain parameters for its diphone, triphone and phoneme context models. On the DARPA resource management task it achieved word correct scores of 96.2% and 81.2% with perplexity 60 and 1,000, respectively.

BBN was able to show considerable progress in adapting the system to speakers with unusual dialects. In a test of four non-native speakers, the word error rate of the baseline system was 31.7%. But after estimating parameters for a probabilistic spectral mapping between each of the training speakers and the test (non-native) speakers, the error rate dropped to 6.5%. A 40 sentence training period was required for this improvement (see Kubala and Schwartz 1990). In addition, the system works as well with training data from only 12 speakers as with training data from 100 speakers. This means that the computing time needed to process the data is reduced, and that it will be easier to collect new data for new domains.

The BYBLOS speech recognizer is integrated with the DELPHI natural language understanding system. BYBLOS delivers the N-best interpretations, and DELPHI then chooses the best overall parse using a statistical-based agenda, an efficient semantic processor, and a domain-dependent frame-like representation of the state of the discourse.

Another innovation is a technique for detecting words that are not in the known vocabulary list. Other systems would find the closest match, even if there were no good matches. BYBLOS can now detect that a word is unknown 70% of the time, with only 1% false alarms.

CMU Carnegie Mellon University's School of Computer Science has been doing pioneering research in speech recognition since the early 1970's. The new Center for Machine Translation adds to the list of on-campus researchers interested in speech and translation. Some of the recent CMU projects are listed below.

The ANGEL project (see Cole et al. 1986a and Cole et al. 1986b) was a good example of a knowledge-based system. It used four separate modules to identify stops, fricatives, closures, and sonorants. Ward et al. 1988 describes a parser designed to work with the lattice produced by this recognizer. There are two problems in a word-lattice parser. First, if there are many word candidates the search for the complete interpretation can be slow. Worse, if a correct word is missing from the lattice, it can be difficult to recover. The paper describes techniques for recovering from missing words. A paper by Tomita 1986 shows how the efficiency problem can be addressed.

The SPHINX project was one of the first highly-tuned HMM-based models. Lee 1989 introduced a number of innovations which led to a significant improvement in overall performance. Many of the features employed in SPHINX—Mel scale cepstra, differential coefficients, multiple vector-quantized codebooks, duration modeling, phonological rules, generalized triphone modeling—have become standard in modern HMM-based systems. On a 1,000-word speaker-independent continuous speech task SPHINX achieved word accuracy measures of 94.7% and 73.6% for perplexity 60 and 1,000, respectively.

Alex Waibel and his co-workers have recently turned to neural net models, with a particular emphasis on Time-Delay Neural Nets (TDNNs). See Hataoka and Waibel 1989. While most connectionist systems have been limited to small vocabularies, Hirai and Waibel 1989 have investigated large vocabulary systems using a hybrid connectionist/dynamic programming approach. Waibel is also known for his work on incorporating lexical stress into recognition systems (see Waibel 1986).

The ϕDMTRANS project is a speech-to-speech translation project using the translation-by-example paradigm. The name stands for "Direct Memory Translation," meaning that the translation is driven by examples that have been stored in memory. The thorough integration of phonology, syntax and semantics is an interesting feature of the system, but the speech recognition component contains no novel features. See Tomabechi et al. 1989.

A CMU project in vocabulary-independent recognition takes a long view at the future of speech recognition. Once systems are released from the lab and start to be placed in the field, the problem of building application-specific vocabularies will become the bottleneck. One way to alleviate this problem is to produce a vocabulary-independent system that is trained on the phonemic combinations of a language, but not on any particular words. See Hon and Lee 1991 for details.

IBM The Speech Recognition Group of IBM's T. J. Watson Research Center in New York is headed by Fred Jelinek. It is one of the oldest groups, having been founded in 1972. They have been concentrating on very large vocabulary real-time recognition, primarily aimed at dictation transcription. A isolated word HMM system with a 5000-word vocabulary was completed in 1985. It achieved word accuracy rate of 94.3% to 98.0%. It is described in IBM Speech Recognition Group 1985 and Jelinek 1985b.

TANGORA is a 20,000 word isolated word recognizer developed at IBM. TANGORA has a means of adapting to the subject matter of a particular document or conversation by catching the words that have been spoken most recently, on the grounds that they are likely to be repeated. With a 1,000 word cache the error rate decreases by up to 24%. This experiment is reported in Jelinek et al. 1991; an overview of TANGORA is in Averbuch 1987.

AT&T As a communications company, AT&T has had great interest in speech recognition and related technology. Rabiner et al. 1988 describe a real-time HMM system for connected digit recognition. More recent work concerns lexical access–mapping a sequence of phones to a word using a dictionary and statistical techniques. This is described in Riley and Ljolje 1991. AT&T is also interested in integrating speech recognition with syntactic (Miller and Levinson 1988) and discourse-level (Hirschberg and Pierrehumbert 1986) analysis.

MIT (Zue) The Spoken Language Systems Group of the Laboratory for Computer Science at MIT has been working on a segment-based speaker-independent continuous speech recognition system called SUMMIT. Each utterance is given separate scores by the acoustic, duration, segmentation and language models. These scores are then combined using a discriminate function which has been trained to make optimal classifications. SUMMIT is described in Phillips et al. 1991. Earlier work on multi-level segmentation is described in Glass and Zue 1988.

More recently, Zue and his colleagues have been working with neural net approaches. Leung and Zue 1988 describes the use of multi-layer perceptrons to model phoneme classification.

MIT (Lincoln Lab) Lincoln Labs is an off-campus research lab associated with MIT. They performed some of the initial research on neural net modeling for isolated word recognition (see Huang et al. 1988 and Gold and Lippmann 1988). More recently, they have developed a continuous speech recognition system based on HMMs (see Paul 1991.

ICSI The International Computer Science Institute in Berkeley, CA has done work in connectionist speech recognition. The emphasis is two-fold: to build fast connectionist hardware for recognition, and to develop connectionist algorithms for estimating the emission probabilities in a standard HMM model.

Dragon Systems This company located outside of Boston, MA markets a speaker-dependent automatic dictation system based on the technology developed in Baker's work at CMU (see Baker 1975b). The system performs in near real-time on a 486-based PC while running the 844 word mammography report task. It uses the phoneme in context HMM model

Siemens The Siemens Corporate Research and Technology Laboratory in Princeton, NJ sponsored a project in connectionist continuous digit recognition (see Lubensky 1988). This system achieved 96.9% accuracy, compared to 96.5% for a traditional pattern-matching system.

INRS The Montreal branch of INRS-Télécommunications has developed a large-vocabulary speaker-dependent dictation system. It combines the 60,000 word Merriam's dictionary with 26,000 additional words, mostly names. The system uses a vector-coded HMM approach with duration constraints. The word recognition rate of 93% is promising, and INRS intends to market the system as soon as it can be made to run in real time. They are investigating parallel hardware for this purpose. See Lennig et al. 1990.

3.9.3 Europe

The European ESPRIT program is sponsoring several speech projects.

SUNDIAL The SUNDIAL project (Peckham 1991) is concerned with information access over the telephone. The plan is to allow access to information in English, French, German and Italian, with a vocabulary of 1,000-2,000 words in each language.

The HMM is used, with experiments taking place to determine the effectiveness of continuous density versus discrete density models. Currently the former outperforms the latter by 75.8% to 67.1% word accuracy for a 1,000-word Italian vocabulary. A 31 word prototype achieved 95.6% accuracy. These are all speaker-independent results, as is appropriate for a telephone information system.

Language modeling is done using a variety of formalisms, principally: Categorial Unification Grammar in English and French; Augmented Phrase Structure Grammar in German; and dependency Grammar in Italian.

There is an emphasis on real-time response.

SUNDIAL shares similar goals with the DARPA spoken language project, but places greater emphasis on using the limited bandwidth of the telephone, and on managing the dialogue. Work has been done on the grammar of spoken language, the trade-offs between user and system initiative, and cross-language similarities and differences in dialogue strategies.

Partners include Logica Cambridge; Univ. of Surrey; CNET; CAP SESA Innovation, France; Univ. of Rennes; CSELT, Italy; Saritel, Italy; Daimler Benz; Siemens; Univ. of Erlangen. Initial applications include train and airline information services and hotel reservations.

POLYGLOT The POLYGLOT project (Boves 1991) is aimed at developing multi-lingual Speech-to-Text and Text-to-Speech conversion. It is based on a data base and statistical analysis of Dutch, English, French, German, Greek, Italian and Spanish. Markov models were used to produce grapheme-to-phoneme and phoneme-to-grapheme conversion modules.

Isolated word speaker-dependent recognition is done with up to 20,000 word vocabularies. For 2000 word vocabularies in each language, the system is able to limit the choice to one of 100 or so words with 98% accuracy. Accuracy figures on the final choice among the preselected words are not given. New speakers need only give about 40 words to train the system. The system uses 20 LPC cepstrum coefficients and 2 energy coefficients for each 10 ms frame.

Continuous speech recognition will be attempted using HMM and TDNN. The Philips Hamburg group reports a word error rate of 23.3% in a no-grammar language model.

Text-to-Speech work has produced a Language Identification Module (LIM) that identifies the language of a text sentence virtually without error. A wild-card parser is used to uncover the syntax of a sentence and thus help with the prosody.

Partners include Olivetti; Bull; Philips; Siemens; The Centre for Speech Technology Research, Edinburgh; CNRS, France; I.P.O., Eindhoven; Nijmegen Univ.; Patras Univ.; Ruhr Univ.; and Univ. Politécnica de Madrid. Language-independent software is developed in C and designed to be shared among all partners. Applications in dictation, office automation and teaching aids are planned. The three-year project runs until 1992.

SUNSTAR The SUNSTAR project concerns the integration and design of speech understanding interfaces. The aim is to build prototypes on two fields: a professional office environment and a public telephone

network environment. The project concentrates on the integration of existing technology rather than on research in new speech technology.

ARS The Adverse-Environment Recognition of Speech or ARS project addresses the problem of speech recognition in the presence of noise. Two application environments—vehicles and factories—have been chosen. The target is to develop a real-time system with a vocabulary of 100-500 words, while making advances in

- reduction of the effects of noise
- identification of robust speech features
- robust matching algorithms
- speaker adaptation
- error correction and feedback mechanisms
- development of hardware prototypes

Participants are CSELT Italy; Telcom France; Univ. of Cambridge; Univ. Politécnica de Madrid; Univ. of Keele; Logica Cambridge; Politecnico di Torino.

ROARS The Robust Analytical Speech Recognition System (ROARS) project will start with an existing real-time speaker-independent continuous speech recognition system for French, and develop a Spanish version. The second phase of the project will develop demonstrations in the air traffic control domain in both French and Spanish, integrating speech with other I/O devices such as keyboards, trackballs and screens.

Participants are Thomson-CSF/Sintra-ASM, France; CRIN; Univ. Politécnica de Valencia; and Ena Telecomunicaciones.

SPRINT The SPRINT project (see Choukri 1990) addresses Speech Processing and Recognition using Integrated Neurocomputing Techniques. It is an attempt to harness the potential of neural nets for non-linear, self-organizing, parallel computation. The main areas of investigation include determining the right features and mapping from parameters to the phonetic level and from the phonetic to the lexical level via competitive learning techniques.

VODIS This is a UK Alvey sponsored project to develop Voice Operated Database Inquiry Systems, with an emphasis on high-level knowledge (syntax and dialogue context). Young et al. 1988 describes VODIS II, a research system that links a standard DTW recognizer (supplied by Logica UK; see Peckham 1982) with a chart parser to explore alternative paths and rank alternatives. Partners on the project are British Telecom, Logica UK and Cambridge University Engineering Department.

Philips The Philips GmbH Forschungslaboratorium in Hamburg developed a variant of the forward-backward algorithm that integrates well with the higher-level syntactic processing that would normally be done by an Early or CYK parser. Using a vocabulary of 82 French words, the system was able to parse 80 test sentences, eliminating "nearly all" errors. The system was developed as part of the Siemens-Philips-IPO (Eindhoven) SPICOS project, which was sponsored by the BMFT.

CSELT The Centro Studi e Laboratori Telecommunicazioni in Torino has done work under the ESPRIT project P26: "Advanced algorithms and architectures for signal processing." A 1,000-word speaker-dependent recognizer achieved a word accuracy of 94.5% for one speaker and 89.3% for another on a task of perplexity 25. The system is described in Fissore et al. 1989.

3.9.4 Comparison of Systems

Speech recognition systems are easier to evaluate than Machine Translation and other natural language processing systems. The speech task is simpler—to identify a sequence of words—and thus it is easy to give a precise measure of success. This is not so in Machine Translation, where there are many possible translations of an input with varying degrees of quality. In the United States, many speech recognition systems have been developed under ARPA funding, and have shared a common vocabulary and task domain. That makes them even easier to compare. Project GRECO in France and the Speech Technology Assessment Group (STAG) in the UK provide similar benefits from standardization.

The following table summarizes the performance of some recent systems:

System Name	speaker indep.	contin. speech	vocab. size	perplexity	% correct	year
TANGORA	–	–	5,000	160	96.9	1985
INRS	–	–	86,000	700	92.8	1990
SPICOS	–	yes	900	74	93.0	1988
SPHINX	yes	yes	1,000	60;1,000	94.7;73.6	1989
DRAGON	yes	yes	1,000	60	94.6	1991
BYBLOS	–	yes	1,000	60;1,000	96.2;81.2	1991
DECIPHER	yes	yes	1,000	60;1,000	95.2;82.4	1991

3.10 Conclusions and Recommendations

Speech recognition technology has made strong, incremental progress over the last five years. This trend can be expected to continue, regardless of the influence of the Verbmobil project.

Speech recognition remains a difficult problem, but it is more mature than Machine Translation technology, in that there is some evidence that it is already approaching the limits of human performance. Pollack and Pickett 1963 performed an experiment where words recorded on a tape of fluent speech were spliced out and played back in isolation, without the surrounding context. Subjects were able to identify the isolated words only about half of the time. When two or three consecutive words were played, recognition was only 70%, about the ability of current speech recognition systems. This suggests that humans are not very good at identifying words in speech, but are able to understand whole sentences because they have a very good language model. On the other hand, studies of trained telephone operators show that they are still much better than computers at recognizing digit strings. This suggests that there is still much room for improvement in speech recognition technology.

We believe that the emphasis should be placed on integrating speech recognition and synthesis technology with the lexical, syntactic, semantic and pragmatic processing necessary for a fully functioning Verbmobil. This means that the goal should be to maximize overall system performance, rather than maximizing word recognition rates.

Most current speech recognition systems produce *words* as their output. In applications with a limited language model, this is appropriate. However, it is not appropriate in the Verbmobil task, which requires complex syntactic and semantic processing of the recognized words. There are two problems with defining speech recognition as word recognition. First, higher-level processing may be able to correct for inaccurate word recognition. It is easier to do this if the output is a list of N-best word candidates than if it is a single word. Second, the way in which words are pronounced can be as significant as their identity. For example, the translation of the single word "yes!" into French might be "absolutement," while the translation of "ye-es" with rising/falling intonation might be "peut-être." We recommend that the output of the speech recognizer be three-fold:

1. A word or segment lattice. Instead of generating the best guess for each word, the recognizer should keep a data structure showing all the word candidates and their probabilities. At the lowest level the recognizer will be building a segment lattice, which lists

the possible phoneme labels for each acoustic segment, but this may be too much information to retain, since it will slow down subsequent computation.

2. Prosodic Structure. Pauses and intonation serve to group words into phrases that can form a different structure from syntactic phrases. This information should be recorded in the lattice. It is especially important for speech-to-speech translation.
3. Intonational Transcription. The system should indicate where pitch accents occur in each word.

Internally the speech recognizer may be computing metrical structure and other things, but there is no need to pass these on to other modules.

The challenge is to integrate knowledge-based sources with the stochastic models that are so good at taking training data into account, in effect providing a model of our ignorance of the true nature of speech. While the HMM approach is currently leading, it may turn out that connectionist models are better at this kind of integration.

The HMM model's recognition algorithm is linear in the length of the input. This is very good news compared to natural language analysis, where parsing is $O(n^3)$ or worse, depending on the model used. Since current speech recognition systems run at 2–3 times real-time, it seems likely that a true real-time system can be produced merely by waiting a few years for the general-purpose chip manufacturers to develop hardware that is 2–3 times faster. There is no need to develop special-purpose hardware in the near term. In the long run, when the details of the system have been determined and cost and size become important for a portable Verbmobil, then it may be time to develop special-purpose hardware.

4

Recommendations

4.1 Introduction

Verbmobil is conceived as a portable, automatic simultaneous interpreter, that is, a machine that makes it possible for people speaking different languages to converse with one another. Since the machines would be portable, each participant would bring his or her own to the meeting. One participant would speak into a microphone and hear translations of what the others said, through headphones, or loudspeaker attached to their machines.

That Verbmobil should be portable is a crucial feature of its design. The notion is that people routinely involved in negotiations with people from other countries should have a device that is attuned to their mannerisms and ways of speech and, conversely, that they have adapted themselves to. Verbmobil should be a personal device for these people who come to rely on it in all their multilingual interactions, and not just when they are in a place that happens to have a large computer installation of the appropriate kind. Crucial though it is, however, we shall have little to say about this aspect of the device for two reasons. First, it is natural for prototypes to be larger and more cumbersome than final products for the simple reason that they must allow for frequent changes in response to experimentation. Secondly, the functionality that can be expected of a device of a given size is still increasing quite rapidly, and we are confident that producing a portable device with the capabilities that it will be possible to provide within the constraints envisaged in this report will not be a great challenge.

This characterization, as well as the term "simultaneous interpretation," brings to mind international conferences in which interpreters look down on large audiences and render the pronouncements of dignitaries into several languages. But, as the pages of this report should make abundantly clear, an automatic device that could provide this

kind of service is far beyond the reach of current science and technology. To try to build one today would be much as if the Wright brothers had made it their aim to build a supersonic jet aircraft before Kitty Hawk had taken off for the first time. If we evoke the image of such a device at all, it is because it serves to establish a direction for this work rather than the goal we actually expect to achieve in more than a very limited way. But we shall want to be able to argue that the actual goals that we establish for ourselves seem to be pointing us in this general direction. To help sharpen our conception of Verbmobil, we begin by considering a number of ways in which the original idea might be restricted so as to make it more attainable.

4.1.1 Verbmobil for Face-to-Face Conversations

The first move is to imagine Verbmobil operating, not at an international conference in a large auditorium, but in a small room where a very restricted group of people is engaged in face-to-face conversation. We take it that the participants will be out of the public eye and will be genuinely more concerned with reaching an understanding than with an impression they might be making on third parties. We imagine that they come to the meeting in a spirit of cooperation and they are highly motivated to reach a successful conclusion. So, diplomats and representatives of countries or companies that are not on good terms are not the best examples to think of. It would be better to think of discussions among companies that are engaged in collaborative ventures, engineers and technical people. However, it is indeed countries and companies that we must have in mind because, while many technical matters surrounding Verbmobil may be in question, its cost, for the foreseeable future, is not.

Another scenario for a device like Verbmobil is the telephone. In some parts of the world, facilities exist for engaging an interpreter for the duration of a telephone conversation who, operating from some central location, is interposed between the participants in a conversation. Setting aside obvious technical difficulties, it is not hard to imagine a machine in this role, and the image provides the motivation for some current projects in automatic interpretation, notably those at ATR in Japan and at British Telecom. In what follows, we will be comparing and contrasting telephone conversations with face-to-face dialog because this will serve to bring out some of the salient features of the latter while setting the Verbmobil proposal off against other automatic interpretation projects currently in development or under consideration.

Verbmobil, as we have said, should be thought of in the context

of face-to-face meetings of small groups that are highly motivated to achieve genuine communication. The point is only that we need people who will be appreciative of the benefits they are receiving from Vermobil and thus tolerant of its imperfections. It also has other effects. The participants in an ordinary conversation communicate in a variety of ways. Most obviously, they communicate through what they say and, given the remarkable flexibility and efficiency of ordinary language, most of the information is probably exchanged in this way. But a great deal also passes by other channels, such as gestures and body language of one kind and another, by pointing and otherwise inviting attention to objects in common view, and by sketches, written formulae, and the like.

4.1.2 Secondary Communication Channels

This makes Verbmobil different from the telephone projects in ways that could prove very important. The secondary channels—gestures, sketches, etc.—are, for the most part, unavailable on the telephone. Consequently, the speech-and-language channel must bear the full responsibility for whatever is communicated. In face-to-face conversation, some of the burden can be taken over by other channels. Experienced travelers who are used to knowing little or nothing of the language that is being used around them make major shifts of this kind quite routinely.

In the long run, this could affect Verbmobil in contrary ways. On the one hand, it could make things easier for the simple reason that, since other channels are carrying some part of the load, a smaller proportion would be left to it. On the other hand, it could be a source of difficulty because, to the extent to which Verbmobil has less responsibility for certain parts of what is being communicated, it is also denied access to some of the context of the communication for which it is responsible. Verbmobil knows only about the linguistic channel and it can easily happen that some linguistic material depends for its interpretation on information that has been conveyed by other means.

Consider an example: Suppose that Verbmobil observes the following interchange:

(1) a. Have you seen this?
b. No, but I read the first one.

To translate "the first one," it would help to know what is being referred to by "this." Suppose that, as he spoke, speaker A held up a copy of a paper with a very obvious "2" or "II" in its title, or on the cover. In the absence of information to the contrary, it is a fair bet that there is

another substantially like it, but with a "1" or "I" in that position. If the translation is going into French, and "article" has been established as the French for "paper," then "the first one" might be rendered as "la première." But, if the conversation is about books, which have been referred to as "livres," then the translation should be "le premier."

For the long-term future of Verbmobil, these effects should not be underestimated. As we have stressed repeatedly in this report, the principal difficulties of language processing by machines come from the fact that so much of what is communicated is implicit or dependent in some way on the context. The availability of parallel communication channels serves only to increase this effect, and to an extent that we have no means to calculate. However, these deficits could still be outweighed by the generally decreased burden that the device would have to carry. We cannot assess how this will fall out because little data is available at present that bears on the question.

For the short-term future of Verbmobil, the situation seems clearer. A primary design criterion of the early versions must clearly be that its dependence on contextual clues be reduced to a minimum because these are the capabilities that will be most difficult to automate. The fact that the machine will be denied access to information in other channels that would be important for this particular purpose should therefore be correspondingly less important. While the designers of Verbmobil may profit from this fact, they should keep two things in mind. The first is that the work required to give later versions of the device the access to the other communication channels that they require will be hard and time consuming, so that it should begin in parallel with the early design work. The second point is, in a sense, the obverse of this, namely that effort should not be expended on attempting to recover from the linguistic channel information that there is no reason to expect to find there.

4.1.3 Overlapping

In ordinary face-to-face conversation, there are periods during which more than one linguistic channel may be open because people interrupt one another, because they misjudge whether the other person has finished, because they try to take the next turn in the conversation at the same moment as someone else, or simply because they deem an interruption to be in order. Only some of these possibilities exist when there are only two parties to the conversation.

If the contributions of the participants in a well behaved conversation do not overlap more than they do, it is because there are conventions that govern "turn taking." An understanding of these may

be important for eventual instantiations of Verbmobil, so that it will know, not simply when one speaker has yielded the floor, but also what the logical structure of the conversation is. We do not believe that this is a matter of concern for the moment, but in eventual embodiments of the device, the pattern of turns in a conversation will have to be consistent with the pattern of questions and answers, proposals and counter-proposals, assertions and confirmations, that give the dialog its structure.

4.1.4 Repairs

In spontaneous conversation, even when speakers speak as carefully as they can, they rarely utter well-formed sentences. Speakers often make errors and correct them. They often begin phrases, abort them, and produce other phrases instead. They often introduce correction phrases like "I mean" and "you know." And they often fill their hesitations with expressions like "uh." To take an actual example from spontaneous conversation, one man produced this sentence:

> Mallet said he felt it would be a good thing if Oscar went.

But he didn't produce it in the clear. In uttering it, he aborted several phrases, as shown in italics here:

> *This Mallet said* Mallet *was* said *something about* he felt it would be a good thing *if* if Oscar went.

He abandoned *this*, for example, and replaced it with *Mallet said* ..., and so on. Also, he signalled the type of trouble he was in with "I mean" and "you know" and prefaced his entire remarks with "well" to give this:

> *Well I mean* this Mallet said Mallet it was said something about *you know* he felt it would be a good thing if if Oscar went.

Finally, he often hesitated, and he filled many of his hesitations with "uh." His actual utterance was this (where full stops mark pauses):

> Well . I mean this . uh Mallet said Mallet was uh said something about uh you know he felt it would be a good thing if u:h . if Oscar went.

In all, this speaker uttered 29 words to produce a sentence of only 13 words–an efficiency of 13/29 or 45%. Although the example may be extreme, in one large sample of spontaneous but careful speech, the average efficiency was 85%.

The challenge for Verbmobil by this 15% excess verbiage should be clear. Verbmobil must be able to identify hesitations, interjections like

"uh," and correction phrases like "you know" for what they are. Ordinarily, we would not want to translate these features. Verbmobil must also deal with self-corrections in which speakers abort one phrase and produce another. The problem is that the aborted phrases are easily mistaken for pieces of the sentence the speakers are trying to produce. Indeed, it may be necessary to translate some of the aborted phrases. Fortunately, self-corrections have a structure all their own, and studies at the Max Planck Institute for Psycholinguistics in the Netherlands, at Stanford University, and elsewhere have uncovered much of that structure (Clark & Wilkes-Gibbs, 1986; Levelt, 1989). So while these are problems that Verbmobil must be able to handle, we believe they are tractable problems. But they deserve more careful study in face-to-face dialogues, where their identification is complicated by turn taking and overlapping speech.

4.1.5 Different Speakers

Telephone conversations differ from those that are conducted face to face in that, except in the relatively rare case of conference calls, the former only involve two people. A conference call would present any interpreter, human or machine, with two additional problems, first to recognize who was talking and, second, to convey this information in the interpretation. The first of these problems, while doubtless interesting for Verbmobil in the long run, need not impede progress towards the creation of the device. For the foreseeable future, we can take it that each participant in the conversation will have a personal microphone that will be so arranged as to exclude whatever others might say. This will be required initially in any case simply to exclude background noise.

A related problem has to do with the fact that different speakers present different problems to the recognition system. As we have seen, there is general agreement among speech researchers that speaker independent systems are considerably harder to build, and subject to more recognition errors, than systems that are trained to a particular speaker. Current speaker dependent systems generally have 1/2 to 1/3 the number of word recognition errors. It is partly for this reason that the initial conception of Verbmobil provides for a separate "box" for each speaker through which his contributions to the conversation pass and which is tuned to his or her speech patterns. However, the information that enabled the system to treat each speaker's input individually need not take the form of a box—it could be a "smart card," a floppy disk, or simply a file of information created for that speaker. Furthermore, the information it contained need not be limited to facts

about voice, speech patterns, dialect, and the like, but also could contain facts about lexical and syntactic usage, even as related to the particular topic of the current discussion.

Conference calls would not only require recognition of input from different speakers, but they would present the users of the system with the problem of how to distinguish the other speakers from one another. According to the simplest design, the system would distinguish the speakers by the microphones they used and deliver interpretations in a standard voice and accent. A person listening to the interpretation would be able to tell whose contribution was being interpreted at any given moment only by remembering the voice and accent of the original or by observing visually who had spoken. The first of these could put an unwelcome burden on the hearer's memory and it assumes that the hearer would naturally come to recognize the voice and accent of a person whose language he does not know. Furthermore, these methods rest on the assumption that Verbmobil really does operate in real time. If there is any substantial lag between the original speech and the interpretation, they could easily present an intolerable burden.

The alternative is that Verbmobil should be in a position to render the interpretation of each person's contribution in a different voice, preferably one that approximates the person's true voice. This is a proposal that is being taken seriously by researchers in telephone interpretation, some of whom do not regard it as an especially difficult problem.

4.1.6 Background Noise

The technology for dealing with background noise is fairly well in hand. Microphones are readily available which, when placed close to the mouth, largely exclude ambient noise without replacing it with new noise coming from such sources as the speaker's breath impinging on it. In addition, systems exist which take in the ambient noise through a separate microphone and effectively cancel that component from what is received from the speaker.

4.1.7 Monitoring

A question that arises whenever a text is translated or interpreters used is that of giving the author, or the participants in the conversation, the confidence they need that the translation really says what they mean. An author has cause to rely on the professionalism of the translator. The author knows that the translations are routinely checked and revised by a second translator, and that the translators will consult with the author for clarification in cases of real difficulty. A person ad-

dressing a conference usually also relies on the professionalism of the interpreter, and on the fact that nothing binding will be allowed to turn simply on the result of the interpretation. All this is not to say that problems do not arise as a result of mistranslation, sometimes severe ones. Users of the initial versions of Verbmobil, however, will not be able to rely on the professionalism of the interpreter. Why should these users place their trust in the machine?

This is not a question that touches research on Verbmobil directly. But it does touch the way in which early versions will be received and therefore, to a large extent, the likelihood that the program will be allowed to continue beyond its initial phase. In the long run, users of Verbmobil will come to trust it simply on the basis of proven performance and the fact that, as with human interpreters, it will not be relied on for the most important parts of the interchange. Furthermore, as we have argued elsewhere, successful technologies do not simply fill previously existing needs, or simply do automatically what had previously be done by a person. They affect the problem as well as the solution. If Verbmobil operated essentially according to our initial conception of it, and cost under $1,000 in today's money, but made egregious errors in one utterance out of ten, there is little doubt that it would be bought in great numbers and very widely used. Owners of the device would rapidly, naturally, and almost imperceptibly adopt patterns of behavior when using the machine that would compensate for its shortcomings. They would follow important utterances with paraphrases of them and invite their interlocutor to verify that the meanings were the same, for example. For the short term, however, this problem has to be addressed, but in such a manner that it does not come to greatly affect the overall plan of research.

4.1.8 Psychological Factors

We discussed earlier the fact that face-to-face conversations make available important channels of communication other than the one that speech and language provide. The other side of this coin is that it may be easier in a face-to-face conversation to remove a sense of pressure that the telephone may engender, at least for some people. Even in the long run, the parties to a conversation that relies on Verbmobil will presumably not know one another well. Furthermore, we must assume that considerable importance attaches to their conversation, so as to justify the use of this expensive device. This means that the situation is apt to be a tense one. It is possible that the use of the telephone would serve to heighten this tension. One becomes more readily at ease with a person with whom somewhat more elaborate forms of social in-

teraction are possible than the telephone allows—if one can shake the person's hand, smile and be smiled at, borrow a sheet of paper, and the like. For many, the telephone remains a short cut, a surrogate for more effective communication, a compromise with the bustle of modern society. Among the many unknown factors associated with Verbmobil is therefore the influence on its users of the environment and the extent to which this engenders a feeling of well-being among the participants and a consequent lack of frustration with the imperfections of the interpretation service.

4.2 Overall Recommendations

The purpose of this chapter, as a whole, is to set out a plan of work leading, by the year 2000, to substantial practical products that will amply justify the cost of the Verbmobil program. We now come to the substantive part of that.

1. The Verbmobil program should aim to deliver two products by the year 2000:
   a. Product one should provide for face-to-face conversation between a pair of participants, each speaking a different language, on extremely limited subject matter, and in circumstances in which the conversational aims of the participants are known in advance.
   b. Product two should provide for the interpretation of occasional words and phrases in a conversation between a pair of participants who communicate mainly in a common language. The subject matter should be restricted, but considerably less than for product one.
2. All work on these initial products should be done with an eye to the fully fledged eventual Verbmobil, so that the scientific foundation of the technology should never be compromised in the interests of achieving some functionality in the short run.

Pursued with vigor and imagination, the Verbmobil program will result in new and important products and technological advances in a small number of years. But the full fruition of the idea will clearly require a much longer period. It is very possible that the simplest and cheapest ways of building the initial products would be such that they would not fall on a natural path towards the eventual goal and, in this case, their value would be greatly diminished. Since we clearly do not yet know how the eventual goal will be achieved, there is no way to be sure that what is done in the early stages will, in fact, be on the right track. But it is of the first importance to guard against making

major investments in short-term goals if they are not expected also to contribute in an important way to the major line of research and development.

Clearly, our recommendations cannot all have equal status. What can actually be done will depend on the economic, human, scientific and technological resources available at every stage, and what is done later will depend on the outcome of what was done before. Our overall plan, however, is clear. We will recommend that most of the work in the program be concentrated on the production of two main artifacts, which we will refer to simply as *product one* and *product two*. In addition, we will propose some studies to provide scientific support to the main efforts.

Product one is intended to come as close to the original conception of Verbmobil as we think possible by the year 2000, even with a fairly high level of investment, great ingenuity and diligence, and considerable good fortune. It is intended to differ from the original conception of Verbmobil as a complete, portable, interpretation machine mainly in that it will be usable only for conversations between pairs of people, and the subject matter and the overall aims of the conversation will be quite restricted. Product two is a much more conservative goal, which, though far from trivially achievable, should be attainable with greater surety by the same time. It will provide language assistance for a pair of people, each of whom has an imperfect knowledge of some common language, with which they need occasional help. We foresee minor variants of both products but, in any of their forms, they would be of immense practical utility. Each of them would amply justify the program in purely economic terms.

We see Verbmobil as a single, indivisible program, and these recommendations should be read with that clearly in mind. In particular, we do not see products one and two as alternatives to one another, but crucially as standing in a symbiotic relationship. Product one is a more ambitious undertaking, and therefore one that involves considerably greater risk. On the other hand, it is designed specifically to stimulate the search for answers to some of the key scientific questions that Verbmobil raises. Product two is more readily attainable, but possibly by providing answers to these questions which, in the long run, would prove to have been misleading at best. We believe that the best way to ensure the success of Verbmobil, both for the short and the long term, will be to insist that products one and two be developed on the same scientific and technological bases. In particular, a proposed solution to a problem arising in the development of product two should

be accepted only with the greatest reluctance if it does not solve the corresponding problem for product one.

It would be easy to argue that, since what we are proposing as product one is more speculative, and also more expensive, than product two, it should be dropped and all the available resources devoted to ensuring the success of the more conservative goal. We have argued against this on the grounds that product one is needed to ensure the scientific integrity of product two. While we believe that this would be sufficient justification, it is not all that there is. Product one is clearly of wider and more long term practical value and, even if it is not attained within the time frame envisaged here, something close to it surely will be attained in the not very distant future and it is clearly highly desirable that this project place itself as close as possible to the path leading to that goal. It would be unfortunate indeed if whoever was first able to build product two did not thereby place themselves in a better position to reach product one.

If fact, we do not see product one as justified by any role that it might play in relation to product two. We put product one in first place because it is on a direct path from our current state of knowledge to the one that we must be in in order to build Verbmobil as originally conceived. It is product two that plays the subservient role. We propose product two for a number of reasons. Needless to say, it will constitute a valuable outcome of the program even if product one turns out to take longer than projected. But this is not the main thing. Far more important is the role it will play in stimulating early work on aspects of Verbmobil at an early stage that would only be possible later if it were necessary to wait for early prototypes of product one to be built. To operate successfully, both products must be sensitive to the continually changing context of the conversation. This is particularly challenging in the case of face-to-face interactions because, while we expect these to be different in many ways from written text, we currently know very little about them. For this, as well as other reasons, it would be invaluable to have a device that operated in substantially the same environment as the eventual Verbmobil, providing substantially the same service from as early a stage in the program as possible. This is the scientific role of product two.

Our recommendations are intended to cover a period of about eight years and, throughout most of that time, we think that it would be inappropriate to devote much effort to the development of an embodiment of either product in hardware. We recommend that a hardware prototype of product two be built during the final three of the eight years and that product one continue throughout the period to be pur-

sued mainly on general-purpose computers. Furthermore, we think it desirable that the software development of both products be carried out using the same "breadboard" program, which we describe later.

4.3 Product One

Product One of Verbmobil should provide for face-to-face conversation between two people, in close to real time, with each participant speaking clearly and distinctly, about a very narrowly constrained topic, and with clearly defined aims to be achieved in the conversation.

During the development of Verbmobil, a great deal of experimentation will be done in which prototypes, components of the system, and people simulating the system, are exercised to collect data either on standards of performance for the system, or on current performance relative to those standards. Much therefore depends on how closely the data collected in this way mirrors the situations in which Verbmobil is eventually expected to perform. On the other hand, a well designed experimental program requires that the data, and therefore the situations from which it is collected, be artificial in certain important ways.

This section contains two main subsections. The purpose of the first of them is to motivate the kinds of severe restriction on the subject matter of the initial versions of Verbmobil that we propose from the point of view of the design process. Roughly, our argument will be that limiting the subject matter of the initial instantiations of Verbmobil will serve to focus the research effort and to greatly reduce the cost of mistakes. In the second subsection, under the title of "User Motivation," we discuss a different kind of motivation, namely one that must be provided to participants in the experiments that will be a crucial part of the Verbmobil program. The idea is to ensure that the data they provide will have the degree of realism necessary to make it really useful. This will turn out to be of great importance in choosing the particular domain to be worked on. We will argue that it will not be sufficient to instruct the subjects in the experiments to imagine themselves to be in some real situation, but that they must be motivated, much in the way that players in a game are motivated, to demonstrate the kinds of behavior expected of them. Before coming to these subsections, we discuss in somewhat more general terms the ways in which we expect early versions of Verbmobil to differ from the original conception of it.

In the fullness of time, there will come to be a Verbmobil portable interpretation machine with at least the following properties:

1. Conversations will be face-to-face, with any number of people, and any number of languages.
2. Each person will speak his or her own language, and have the contributions of others translated in real time.
3. Participants will speak in the way that is most natural to them.
4. There will be no restrictions on the subjects that can be discussed.
5. Verbmobil will be sensitive to gestures and other nonverbal means of communication.

Many of these things are clearly beyond reach in the easily foreseeable future and our first task must therefore be to set forth a more modest set of goals that might be achievable by the year 2000. To this end, we propose *product one* which will be the closest thing to the original conception that we think might be achievable in the time frame.

The description of our proposed product one results from modifying the properties we ascribed to the original conception in the following ways:

1. Face-to-face conversations between two people will be provided for using only two languages.
2. Each person will speak her or his own language, and have the contributions of others translated in close to real time.
3. Participants will speak carefully, distinctly, and in a standard dialect.
4. The subject matter and the goal to be achieved by each participant will be narrowly controlled.
5. Verbmobil will be totally insensitive to gestures and other nonverbal means of communication.

We believe that achievement of product one rests mainly on point 4.

We do not regard the restriction to pairs of participants as severe. The difference between face-to-face conversations and monologue or written text is, however, great and not generally understood. The addition of more speakers will be altogether less significant than this, serving mainly to complicate matters at a relatively low level—the placement of microphones, elimination of cross talk, distinction of speakers for the hearer, and the like. We do not believe that these matters should be allowed to divert attention from other matters that are clearly more important. Just what should count as acceptable performance can only be determined as a result of experimentation.

When we say the "Participants will speak carefully, distinctly, and in a standard dialect," we are also referring to standards that will

emerge over time. However at no time during their development should either product restrict its users to speaking with isolated words. That is, participants should never have to speak with pauses, however short, in unnatural places.

When we say that hearers should receive the interpretation of what the other participants say in "close to real time," we are concerned that that time should not be so long as to destroy the continuity of the conversation. Without this requirement, there would be no satisfactory way of judging the success of the enterprise. However, we wish to stress that performance should be used only with great reluctance to justify *ad hoc* solutions to problems when more principled ones are available.

4.3.1 The Domain Restriction

We believe that the conversations that are mediated through product one should be restricted in two ways:

1. The subject matter should be very narrowly constrained, and
2. Each participant should have a task to perform, and there should be clear criteria for determining when it has been successfully completed.

What we have to say here is also relevant to product two, although we believe that a broader domain would be appropriate there.

4.3.1.1 Design Motivation

The subject matter that is discussed using product one should be severely limited:

1. *To limit the size of the dictionary,*
2. *To limit the number of things that can be referred to,*
3. *To encourage repetition of phenomena,*
4. *To facilitate collection of a corpus of examples.*

We have discussed the importance of domain restrictions for Machine Translation systems in general in 4.3.1. We return to the question here with Verbmobil specifically in mind. Limiting the subject matter is standard procedure in artificial intelligence and computational linguistics. First, it is important to limit the size of the vocabulary that the system must be able to deal with. In many respects, what is learned from each new word that is acquired diminishes rapidly as the vocabulary increases and the point is rapidly reached at which adding new items is little more than a drain on resources. Inevitably, in so adventuresome an undertaking as is being contemplated here, there will be major changes in direction from time to time, possibly making it neces-

sary to change the system's dictionaries in fundamental ways. Clearly, the smaller those dictionaries are when that happens, the better.

It is also important to limit the ontology—the number of different kinds of things that the system must know about, and therefore have means of referring to. We believe that it will be necessary for Verbmobil, even in its initial configuration, to have quite detailed knowledge about the small part of the world to which it has been specialized. Unless the size of this is kept well under control, the designers will be reluctant to contemplate changes in direction that would affect the way this knowledge is represented and used.

By far the most important reason to restrict subject matter in the initial version of Verbmobil, and the aims of the participants, is to encourage the natural recurrence of phenomena of various kinds across different conversations. A primary source of difficulty in much of artificial intelligence and computational linguistics is that of determining whether two pieces of data represent occurrences of the same phenomenon. This is made a great deal worse by that fact that many of the phenomena of interest seem to occur naturally only in quite weak dilution. It will therefore be very much in the interests of the Verbmobil designers to achieve repeatability from one experimental situation to another.

A concomitant of this repeatability property is that it should be possible to assemble the corpus of examples of Verbmobil's expected performance early, relatively easily, and concurrently at more than one site.

We see the domain restriction as playing a particularly important role with respect to the part of Verbmobil concerned with speech input. This is because we think it very important that all phases of the work on speech recognition should be conducted in the framework of a system designed to profit from constraints of many kinds coming from outside the speech component itself. We believe that it is impossible to assess properly a speech system outside the context in which it is intended to operate and that even the best techniques that will become available while this project is going on will require strong outside constraints to work effectively. Only with severe restriction on the size of the domain with which the initial prototype works will it be possible to provide a rich enough context to treat each utterance in a principled way while, at the same time, not making too early a commitment to components of the system that have not been sufficiently exercised.

4.3.1.2 User Motivation

It is important that the domain chosen should be such as to encourage cooperative, rather than adversarial interaction. To the extent possible, the domain should be such as to make it possible for experimenters to manipulate the tasks given to subjects, so as to control the likelihood that the resulting conversations will have properties that interest them. During the first year of the program, possible task domains should be studied and a suitable one chosen. The most attractive candidates at present are the Contract Negotiation *and* Design Negotiation *tasks, because they reduce to a minimum the problem of giving Verbmobil access to the objects that the participants are talking about.*

We now go on to discuss the properties that we think would be desirable in a domain for use in the initial phases of Verbmobil experimentation, and ways of constraining the goals of the participants. We are not advocating one particular choice, because we believe that this is a subtle matter, itself worthy of the investment of some serious research effort. In fact, we propose that this be a matter of intensive study during an initial phase of the program. We appreciate that there could be more than purely technical reasons for preferring one domain to another and this is an additional reason for not advocating a specific one at this stage.

4.3.2 Purposefulness and Evaluability

We take it that language is a purposeful activity: speakers use language to help accomplish goals. Thus, for Verbmobil we would like to encourage interactions where each participant has a clear goal. By measuring how well the goals are achieved, we can assess how effective the communication process has been.

This kind of purposeful interaction is often found in games. In a game, there is usually a very limited set of clearly understood conditions that would constitute success and bring the activity to an end. The overall aims of each player are known to everyone from the start; only the particular strategy currently being pursued may be unclear, though the rules of the game generally limit the possibilities quite severely. In the game-like activities that would suit our purposes best, there would be certain things that each participant would have to achieve, preferably through verbal interaction with the other participants so that the linguistic correlates of these should be part of the transcript of every session.

4.3.2.1 Cooperation

The second property that we would like to see in the domain chosen for Verbmobil is that the conversations among the participants should not be adversarial. We are not concerned to exclude bargaining, or conversations between people with conflicting interests. We only want to minimize interactions between people with hidden agendas. We believe that it would not promote the interests of early research on Verbmobil if participants were frequently in the position of saying one thing but meaning another, providing certain information while concealing related information, and the like.

Although, as we have said, our desire to foster cooperative dialog by no means entails eliminating all elements of negotiation, it sets something of a premium on identifying a class of goal-directed activities, in which the participants are partners in a contest with nature, time, or previous attainments, rather than with one another.

4.3.2.2 Manipulability

A property that would be desirable in a domain, for our purposes, would be a certain kind of *manipulability.* This is a poorly defined property, but one which can differ markedly from one domain to another. The basic idea is that we want to be able to manipulate the tasks so as either to force subjects into particular linguistic predicaments, or to make this unlikely. This notion will become clearer from the discussion of particular domains that we are coming to now.

4.3.3 Possible Domains

We now consider a number of domains from the point of view of their suitability of experimentation with Verbmobil. We are not recommending that the designers of Verbmobil chose a domain from our list. On the contrary, we are recommending that the search for a suitable domain be an explicit topic of research during the first year of the program. We examine these examples, somewhat cursorily, in the hope that this will sharpen our view of the importance of this task and of the key factors that it involves.

The first two domains we discuss have a long history in work of this kind, the first at Edinburgh University, and the second at SRI International.

4.3.3.1 The Map Task

The map task is carried out by two people. Each has a map, and on one of them, a route has been drawn from one point to another. The task for the person with this map is to convey enough information about the route for the other person to draw it in. Success is judged by placing

a grid over both maps and counting one for each corresponding square through which both routes pass, and subtracting one for each square through which only one route passes.

The maps used in Edinburgh are not road maps. They show an imaginary island and a variety of landmarks. One of several variations of the task turns on using maps which, while quite consistent with one another, are not identical. For example, the map on which the initial route is shown may show a windmill with the route going past it. The map of the participant who is expected to draw in the route may show two windmills. The person with the reference map can almost be counted upon to reach a point when he talks about "the windmill," whereupon a negotiation starts to determine which windmill is being referred to.

The possibility of including two windmills on one map and only one on the other is an example of the kind of "manipulation" that we think would be desirable in a domain chosen for Verbmobil. Through the kinds of maps used and the features shown on them, it is possible to exercise a considerable amount of control over the vocabulary used, and by varying the number of instances of a given feature, it is possible to create particular problems in resolving reference. By manipulating these things, it should be possible to create data sets that are closely comparable with one another.

Other variations in the task are of less immediate interest for present purposes, though they give rise to effects that it might be important to control for. For example, it was found that, if the two participants in the dialog could see one another's faces (though, of course, not one another's maps), the task was completed significantly faster. The same was true if the participants were already familiar with one another.

A minor objection that might be raised to the map task is that the roles of the two participants are too asymmetrical. One has the answer to the problem and the other has to get it. We do not see this as a serious disadvantage, especially in view of the fact that the introduction of minor differences in the maps can give rise to a lively interaction and a more balanced relationship between the participants. In fact, it could easily lead to a reversal of the roles during part of the conversation.

4.3.3.2 Trucking

The Edinburgh version of the map task could be thought of as one of a family of tasks. The following variation places the participants in a more equal relationship relative to one another. It is a simplified

version of a task that was used some years ago at the Xerox Palo Alto Research Center as part of a training course in Artificial Intelligence techniques. We place it in Germany; the original was situated nowhere in particular.

Suppose that a road map of Germany is provided to each participant and that the distances between key cities are marked. Each participant is to play the role of a truck driver who is initially in a certain city, with a truck having a certain capacity and average speed. In certain cities there are given quantities of perishable merchandise addressed to other cities on the map. It is the collective responsibility of the truck drivers to see that it all gets to its destination as soon as possible. It is arranged that this cannot be done at all well by having one truck pick up each item and deliver it directly to its destination. It can only be done well if the trucks arrange to meet and exchange merchandise according to a complex plan, which they must negotiate. There are clearly indefinitely many variants of the map task, and this is one of the things that makes it an attractive paradigm.

4.3.3.3 Journeyman-Apprentice Tasks

Not greatly different from the map tasks is the journeyman-apprentice task that was the basis for the contribution of SRI International to the original ARPA speech project. Like the map task, it had several variants.

Two people are required for this task, one who will build a pump, or some such device, out of component parts that are supplied, but who does not know how to do it, and another who knows how to do the job but cannot actually touch the parts. In addition to the parts of the pump, a number of tools are available. The first person, the apprentice, must find out how to build the pump from the second person, the journeyman. Often, the apprentice was not familiar even with the names of the parts or of the tools that it would be necessary to use.

As with the map task, the basic scenario was played out in a number of different variations. In some, the journeyman could see what the apprentice was doing, but could not intervene; in others, the participants were in different rooms. In some variants, the participants communicated by ordinary speech and, in others, through a computer terminal.

There is much in common between this and the map task. This clearly also belongs to a family whose members differ in the exact nature of the task, how difficult it is, how many people are involved, both as journeymen and as experts. Like the map task, it places those

involved in asymmetric relations relative to one another, though this is subject to manipulation and, in any case, we suspect that this is not a crucial matter. This paradigm may contain some fundamentally richer variants if only for the reason that it involves the manipulation of three-dimensional objects.

4.3.3.4 Conference Registration

We include this here because it is the domain that was chosen by the project of the Advanced Telephony Research Laboratory (ATR) in Japan, and because it has many of the properties of a good domain for our purposes. The idea is to decide what options would be open to a person registering for a particular conference (special sessions, hotel rooms, banquet, tours), to give to one person the job of registering people for the conference, presumably including filling out a form, and to assign to another person the role of someone wishing to attend.

This shares with many potential domains the difficulties of motivating participants in experimental situations and of determining when the tasks they have been given have been completed. We believe that it would not be satisfactory simply to tell one subject to imagine he or she wanted to register for a conference and another to fill out a registration form. There is no reason to suppose that what the subjects imagined would be at all realistic or that the data collected in various sessions would differ from one another in interesting or informative ways. In other words, it would be hard to invest this domain with the kind of manipulability we desire.

In this task, it would be important to be very specific about just what each participant was expected to achieve. For example, the person at the conference site would be expected to complete all the items on a form. The person intending to attend the conference would be given a sheet of information on the person whose role they were to play. For example, they would be given a sum of money to spend that would cover less than all the conference activities that they were interested in, and there might be constraints on when they could arrive and depart. But it remains difficult to see how the interactions could be manipulated in any but very gross ways.

We think that the design of experiments based on this domain could prove to be more subtle than in the others we have suggested and that it is also less manipulable. This is not to say that, with sufficient care and ingenuity, it could not be made to work.

4.3.3.5 Contract Negotiations

Like several of our other suggestions, this will be an example of a class of tasks from which we have chosen one just to simplify the discus-

sion. What is important here is that the object of the discussion is a document, not necessarily a contract, which is available to Verbmobil as well as to the participants in the discussion. We think of it as a contract because each participant will have an interest in changing, or maintaining, certain sections in the document and, in the interest of maximizing their interest, we will suppose that each will be scored by the number of his requirements he is able to achieve.

Suppose then that a document is displayed, either on a different screen before each participant, or on a single screen that both of them can see. If the screens were different, and each participant had a pointing device, such as a mouse, the possibilities for using this for reference resolution in the early stages are very attractive. Each participant has been given, in advance, a sheet showing a modified version of the document, with a number against each section that is different from the one on the screen. That number will be added to that participant's score if the corresponding section is in the modified form at the end of the discussion. Neither participant knows what is on the sheet that has been given to the other. In particular, they do not know how the other one will be scored for the amendments she or he is arguing for. The participants will be allowed a limited time to agree on the final form of the document. We will take it that the document is in a language known to both participants, but a version of the task could also be developed in which each had a version of the document in his own language. This would open up many interesting possibilities that we will not go into.

We find this scenario very appealing, for the following reasons:

1. It is the only scenario we have mentioned in which Verbmobil has just as easy access to the object of the conversation as the participants do.
2. Using a mouse, or similar device, Verbmobil can be allowed to see what part of that object the participants are pointing to at any given moment, although, in an experimental variation, it can also be denied access to this information.
3. The subject matter of the document can be varied to provide as wide a range of experimental situations as desired.
4. The kind of aim that the participants have in the dialog is known to Verbmobil in advance and, in the early stages, it is even possible to make the specific aims known.
5. The dialog is not driven by the imagination of the subjects, but by the interactive properties of the task.
6. The task is very manipulable, in our sense of the term, because

there is no end to the types and complexity of the amendments that subjects could be called upon to make.

7. Realistic situations that fit this paradigm are many and various.

4.3.3.6 Design Negotiations

The principal drawback of the scenario of contract negotiations is that the object of the conversation is itself a linguistic object. If its language is that of neither participant, then it is almost inevitable that words and phrases that it contains will figure very prominently in what the participants say so that the language of the conversation will be very mixed, providing special problems to both the speech analysis and translation parts of the interpretation system. On the other hand, the difficulties of designing a version of the task in which there were versions of the document in the languages of both participants could be severe. With this in mind, we suggest a modification in which the object of the conversation is something that can equally easily be made available to the system, but which is not primarily linguistic.

Suppose that what appears on the screen, or screens, that are before the participants is not a piece of text, but either a plan—say the layout of an exhibitor's booth at a forthcoming trade fair—or some data that can be presented in the form of a spreadsheet—say the budget for some proposed project. Once again, we suppose that each participant has been given a set of amendments, with associated scores, that he must urge on the other. We also suppose that there is a suitable interface to the display software that will enable changes to be made which will be not only displayed on the screen, but transmitted to Verbmobil.

This task has all the appeal of the previous one, while getting around the problems associated with using a third language.

4.4 Product Two

Product two of the Verbmobil program will provide translation on demand for participants who have a passive knowledge of a language of which neither is a fluent speaker, but in which most of their conversation will be conducted. In the course of an utterance, a participant can press a button to signal that he is now talking in his native language, and that what he says should be translated into the common language.

In the original conception, Verbmobil provides for interpretation of essentially everything that passes between the participants in a conversation, at least through the linguistic channel—it is assumed that whatever knowledge any of them may have of the languages of the others is not brought into play at all. However, if Verbmobil were being

used by people that had a common language, even if they knew it imperfectly, then the burden placed on the machine would be very much less and it would presumably be able to operate in something much closer to real time.

The simplest conception would be one in which all the participants had a good everyday knowledge of one language, say English, and used this as the main medium of communication. They would turn to Verbmobil only occasionally, for uncommon words, technical terms, complex constructions, or important points that they wished to verify. On these occasions, we may take it that the person needing help would press a button, causing Verbmobil to take as a fragment requiring translation whatever was said until the button was released. In the limit, this trivializes the translation job that Verbmobil has to do, reducing it to looking up terms in a dictionary. On the other hand, we can also imagine that, with a view to choosing correctly among competing translations for a term, the device takes some account of the context in which the term is used, for which purpose it may make use of some part of the spoken material before and after the moment when the button was pressed.

This version of Verbmobil is by no means as trivial as it might seem at first, though it is clearly a great deal simpler than product one. The machine carries a much smaller burden in this version; evidence of its imperfections would therefore be less frequently brought to the attention of its users and the overall impression should therefore be of a more reliable tool. However, while keeping users happy in this way, the system could also be taking note of the intervening conversation and using what it learnt in this way as context for the queries that were specifically directed to it. It could thus serve as a valuable source of data for researchers working on the principal product.

It would be important to limit the subject matter for which product two was used, just as it was for product one. Initially, the limitations should be no less severe than for product one, and for the same reasons. However, after the development of an initial prototype, the domain should be extended considerably. The reason for this is to avoid any tendency for product two to become, albeit covertly, a somewhat extended system for isolated speech recognition coupled with an electronic phrase book. In this version of Verbmobil, unlike product one, it is important to face the problems in perplexity that come with a larger vocabulary and open-ended possibilities for the construction of phrases and sentences. Naturally, there should continue to be a reasonable limit on the length of the segments that the system can be called upon to translate, and how often such segments occur. Without

such limits, product two would not be distinguishable from Verbmobil as originally conceived.

We favor using the same initial domain as for product one, say one of the negotiation tasks in which the object of the negotiation is available to the interpretation system. In the early stages, we see much benefit from sharing grammars, lexica, knowledge bases, and the like between the two products. After the initial prototype, the products will diverge to the extent that the coverage of product two will increase much more substantially. This should not involve a change in subject matter, but simply an enlargement of the original domain. This will have the advantage of making available to product one the material that would allow its coverage to be extended as soon as that seemed reasonable. Another way of effectively extending at least some parts of the domain of product two, if one of the negotiation tasks were used, would be to progressively deny Verbmobil access to the document, design, or whatever, that was the object of the conversation.

A variant of product two would be a device in which it would be the hearer, rather than the speaker that pressed the button. When a participant in the conversation heard a word or phrase that he did not understand, he could press a button that would enable him to get a translation of it. One possibility would be that he would hear an interpretation of the last so many seconds of speech. Another would be that he would be shown a transcript of the last so many seconds on a screen and would be invited to select the specific sequence for which interpretation was required. The latter suggestion, while doubtless easier to implement, has the disadvantage of requiring the hearer to be able to identify a specific sequence of words when the trouble might be precisely that he was unable to distinguish the sounds.

This latter version, if it could be made to work reliably and with little cognitive load on the user, could be of inestimable utility to a professional interpreter in providing technical terminology.

4.5 The Experimental Paradigm

In the early stages of the program, product one should be carefully simulated using professional interpreters and a set of data collected that will then be usable by all groups involved in the research and development. The results should then be processed so as to form a data base containing the expected inputs and outputs at each of the major interfaces in the system. A preliminary version of the eventual system, called the "breadboard" should be built with dummy programs in place of the major modules. The dummy programs will

produce "correct" outputs for examples in the experimental data by simply delivering what the data base says they should.

Verbmobil, and particularly product one, is an extremely adventuresome project. To compare it to the project of thirty years ago to put a man on the moon may seem pretentious, but this also required the coordination of a number of smaller projects each responsible for a component whose exact properties could not be fixed at the outset. Nevertheless, each component had to interact closely with other components in complex ways when the final product was assembled. We believe it is reasonable to expect a happy outcome from such an enterprise only when all the work is carried on within a carefully designed experimental paradigm.

It seems to us that the telephone interpretation project at ATR suffers, more than anything else, from that lack of a sufficiently well articulated experimental paradigm. A considerable investment was made early in the life of the project in collecting data. However, insufficient consideration was given to the way in which the data was collected and to the way in which it might be used in organizing the project as a whole. We think it likely that this is responsible for the impression of fragmentation and disjointedness that one gets at ATR. Excellent work on speech recognition, machine translation, and speech synthesis was conducted, apparently in the hope that a way will be found of assembling the results into a coherent whole in the end. This is not to criticize ATR, who were the pioneers in this field, nor is it to say that these matters are straightforward, but it is to say that they are important and should be given serious consideration from the very beginning.

Of the suggestions we are about to make, we say what we have said of others before: We are more concerned to exemplify the considerations that seem to us important than to lay down a particular set of rules on which we think the success of the project crucially depends. In other words, when in doubt about how much detail to include in our recommendations, we have erred in the direction of greater specificity. Our emphasis, in this section, will be on the version of Verbmobil we have been calling *product one*. Needless to say, similar considerations apply to experimental paradigms constructed for subsidiary goals, though these should clearly be folded together with the primary one to the extent possible.

In our view, the experimental paradigm that gives structure to the whole Verbmobil program needs to have two components. One consists of a body of data generated by human subjects, but showing how the

complete device, correctly operating, would perform in various situations. The second component is what we call the "breadboard." This is a complete system, built early in the history of the project and which, drawing on the data in appropriate ways, can simulate the performance of the final Verbmobil. The next two sections treat these two aspects of the experimental paradigm in turn.

4.5.1 Data Collection

The main reason for collecting a body of data will be to facilitate simulations of some components of Verbmobil so as to provide a realistic environment for experimentations with others. It is a fairly straightforward matter to simulate Verbmobil using professional interpreters. This is not to say that it is trivial. It is quite likely, for example, that professional interpreters will convey more in their intonation than can be expected of the young Verbmobil. This is simply one among several factors for which it will be important to control. Records of simulated Verbmobil sessions will constitute the primary data we will require.

There will be a strong, easily understandable, urge to vary the circumstances under which sets of these simulations are conducted, but the principle need will be for large amounts of data collected under circumstances that are as nearly identical as possible. Given the high cost usually associated with collecting data, the urge to collect many variants should therefore be held in check. The principal dimension along which variation is desirable will be dictated by the domain model that has been chosen. As we said earlier, a good domain for our purposes will be manipulable in the sense that it will be possible to specify within it tasks that put more or less strain on the communication channel.

The principal reason for wanting to collect fairly large numbers of conversations under essentially identical circumstances is the following: We want to be able to identify in the data base at least one, and preferably several translations that can be treated as "correct" results that should be among the outputs the machine system will be expected to produce. In a recent study, Henry Thompson (personal communication) of Edinburgh had fifty different translations into French made of a small number of utterances collected from experiments with the map task described above. The translations were made by professionals, amateurs, machines, bilinguals and people with almost no knowledge of French. There was no sentence that occurred more than once in the total corpus of translations collected. We do not find this result particularly surprising. However, it does underline the fact that the translation relation is very poorly defined and that it would be foolish to try to designate one translation as correct, and the one that

Verbmobil would ideally produce. However, we take it that Verbmobil will be at least potentially nondeterministic and we think it reasonable therefore that it be expected to produce at least one of a set of translations that have been designated correct. The importance of this will become clear shortly.

The data should be collected in a setting in which the interpreters are in a separate room from the participants in the conversation, unable to see them or anything that they can see. This is to reflect as closely as possible that fact that, within the time frame envisaged here, the only communication channel connecting the participants in the conversation will be the primary speech channel. High-quality audio recording, placing the contributions of each of the primary participants and of the interpreters on separate tracks should be used. Doing things in this way may preclude the use of cassette recorders, but we believe that the expense and added inconvenience will be amply repaid through the life of the project. We also see value in collecting a video record of the main conversation, though probably not of the interpreters.

In the Verbmobil program, the primary interest is in face-to-face conversations, and these should clearly be the primary target of the data-gathering effort. However, we think it very likely that some data on comparable conversations in which the participants cannot see one another may be of value in helping to determine the importance of information that passes otherwise than through the primary speech channel. Work in Edinburgh suggests that the efficiency with which people accomplish verbal tasks together is strongly influenced by how well they know one another, and this is a parameter for which it would be important to control.

4.5.2 Processing the Data

The data just described will serve as a set of samples of the input-output relation that Verbmobil should define. Much of the work in the Verbmobil program, especially in the early stages, might well be based on a version of the data base that has been cleaned up in certain ways, such as by removing false starts and eliminating flagrantly incorrect translations. More importantly, the initial data base needs to be enriched so as to become a sample not only of Verbmobil's overall input-output relation, but also of the input-output relation of each of its major components. For the moment, we assume that there will be a need for representations of each utterance of at least the following kinds:

1. For the original:
   - Tagged phonemic transcription,
   - Tagged orthographic transcription,
   - One or more syntactic analyses,
   - One or more interlingual representations, correlated with the syntactic analyses.
2. For each translation
   - Interlingual representation,
   - Syntactic representation,
   - Orthographic transcription,
   - Phonemic transcription.

These will be prepared "by hand" and, needless to say, if the formalism used in one of the major components is changed, it will generally mean that corresponding parts of the data base will have to be changed. Once the results of these analyses have been added to the data base, it will be possible to exercise any of the major components of the system against any part of the data, on the assumption that the other components are doing their jobs correctly.

Clearly, a specific step in the analysis should be undertaken only when the component that it simulates has reached a stage of design that fixes the form that the analysis should take. Equally clearly, the data bases should not be the only source of data against which components of the systems are tested. The main role of the data base is to make sure that there is a reasonable body of data, from as early a stage as possible, on which the system as a whole can operate.

4.5.3 The Breadboard

The term "breadboard" comes from electronic design and experimentation. It refers to an insulating board on which components could be laid out and wired together, usually without solder, so that they could easily be replaced in the course of experimentation. The breadboard we have in mind will be an analog of that, built in software. It can be thought of as the main program of an early prototype of Verbmobil, with a slot to accommodate each of the components that do the real work. In fact, we countenance the notion of several breadboards being built throughout the course of the program, corresponding to a number of different prototypes.

We recommend that the first breadboard be built very early, certainly by the end of the third year and, furthermore, that the place it provides for each of the major components of the eventual system should actually be occupied by a program from the start. The pro-

grams that initially fill all these slots will be simple dummies. What they will do is to map inputs found in the data base into the corresponding outputs, and nothing more. The experimental paradigm will never call upon them to handle any input not found in the data base. The interfaces through which these dummies communicate with the rest of the system will be just the same as the interfaces through which the more substantive modules that replace them in the eventual system communicate so that it will be a simple matter, in principle at least, to replace any of them with more substantive modules when they reach a suitable stage of development.

The breadboard, then, will provide, from the earliest days of the program, an instantiation of the overall architecture of Verbmobil. This is the most important property that it can have. A researcher who is responsible for a particular component of the system will be able to conduct experiments on it in the context of the complete system, on the assumption that all the other modules work perfectly. It goes without saying that this is an idealization and that the conception of what a perfectly operating system would be like will change as the program progresses. Updating the data base so that it reflects the current view of how the final system would operate will require a substantial investment of effort. We believe that it will be repaid.

As well as providing an early embodiment of the Verbmobil architecture, it would be useful if the breadboard provided what has come to be known in computer science as a "shell," that is, a processor for a simple formal language by means of which an experimenter working with the system can exercise it, trace the workings of certain of its parts in greater or less detail, repeat specific parts of the process with minor variations, and so forth. By giving commands in the language of the shell, he or she would be able to insert particular versions of given components into the system, select examples from the data base to apply the system to, modify grammar rules, and call for detailed reports on the operation of any module.

4.6 Variants of Verbmobil

As well as product one, a number of variants of Verbmobil should be made for the strength that they will give to the main effort, for their intrinsic value, and for their value as insurance against unforeseen flaws in the original conception. The variation known as product two provides for interpretation among people who are able to carry out much of their conversation in a common language, but need help from time to time. Others involve providing the system with

information about the conversation that would be hard to derive just by listening.

Up until now, we have been concerned with what we have called products one and two. These are the goals to which the bulk of the resources available to the program should be devoted. However, we believe that it would be in the interests of the program that some of the resources should be made available for related efforts, because they will help the development of the primary products, because they are targets of opportunity immediately adjacent to the path towards these products, or because they provide some measure of insurance against unforeseen flaws in the conception of Verbmobil. Some of the suggestions that we have to make along these lines are intended to provid opportunities to test some of the important ideas and components of Verbmobil sooner than will be possible in the main system. Verbmobil is a long term program that can only hope to reach its product one after a very considerable investment in time and energy. It will be valuable to those close to the research, as well as to observers who might be inclined to lose faith in the attainability of product one, if the work can be shown to have practical value before that goal is reached.

The judgments we make about proposals for intermediate goals may become quite subtle because what we say about a particular proposal will depend on how we see it fitting into an overall plan. To be clear about this, consider a simple example. Suppose that it is proposed to build an interim version of Verbmobil whose interesting property is that it is capable of recognizing isolated words only, and not continuous speech. Now, the extent to which isolated speech recognition should be viewed as a step on the way to continuous speech recognition is a matter on which we have made our position clear. However, it may be argued that the use of isolated word recognition is justifiable on the grounds that (1) it provides a complete environment in which to develop other components, (2) it will not adversely affect the design of other components, and (3) its use will not stand in the way of developing the continuous recognition system that will eventually be needed. In fact, we could be convinced by these arguments provided that the part of the system that would do the isolated-word recognition already existed or could be realized at essentially no cost.

In what follows, we shall consider proposals that differ from the original in two ways: Either they will have reduced capabilities or what they can do automatically will be enhanced by arranging for human help. Particular systems may have different mixtures of these deficits and prosthetic devices. A third matter that we shall consider is ways

in which Verbmobil's performance may be enhanced not by changes in its design but by modifications of the environment in which it is used.

4.6.1 The Electronic Blackboard

Under the heading of the electronic blackboard, we collect a number of suggestions aimed at allowing speakers to provide supplementary information to Verbmobil through an ancillary channel. There are doubtless many more uses for the appendage we are about to suggest for Verbmobil than those we shall suggest.

Suppose that each participant in the conversation has before him a screen that is somehow sensitive to what part of it is being pointed at. Suppose, furthermore, that, whenever someone refers to something that has not been mentioned in the conversation before, he points to an unoccupied place on the screen, thereby causing an icon of some sort to appear at that position. The icon appears at the same time on the screen of all the other participants. When somebody refers to something that was introduced into the conversation earlier, they point to the corresponding icon which responds, say by changing color briefly, so that it is clear from the other screens, which icon is being pointed to.

We do not expect the participants to derive great benefit from watching icons being introduced and pointed to in this way, though it may be not entirely without value. The point is to provide Verbmobil with information about the referents of expressions that would be extremely difficult for it derive from the context of the discussion. Knowing what is being referred to, say by a pronoun, is often crucial for translating it correctly into another language, yet keeping track of that referent, and knowing when it is being referred to again is a task that lies almost entirely beyond our current technology. For the near term, this simple device could serve as a prosthesis for the referential capability that could substantially increase Verbmobil's capabilities.

There are several fine points relating to this version of the electronic blackboard that would need to be worked out. For example, if each icon looked exactly like every other, it would be easy to forget which referred to what. It would be altogether better if the icons were labeled or even if their shapes could somehow be made reminiscent of what they stood for. A stock of suitable icons could be developed and each word that might be used to refer to something could have a suitable icon listed against it in the dictionary. It remains to decide how the icons will be retrieved from the dictionary. One possibility is to make it a responsibility of Verbmobil. When a person points to an unoccupied place on the screen, Verbmobil identifies the current, or

the most recent, referring word, and places the corresponding icon at that position. If the wrong word is identified, the speaker signals the fact, say be pressing a button, and the next most likely interpretation is made. But possibly, this would be beyond the early capabilities of Verbmobil, and the responsibility would fall to the person making the reference. Perhaps the person would have to type a short label to go with the icon.

Another problem with the scheme is that the screen would rapidly become cluttered with icons, most of which would never be referred to again. One possibility for dealing with this would be to arrange that new icons were always introduced, say, at that top of the screen and that, as the discussion progressed, they migrated slowly towards the bottom where they would eventually disappear. When they were referred to again, they would be promoted to the top of the screen again, thus getting a new lease on life. Needless to say, this scheme could be quite difficult to manage and might be unnecessarily distracting for users. Perhaps it would be enough to just delete an icon from its original position when it no longer justified the space it was occupying, or maybe they could slowly fade away.

It is perhaps worth pointing out that, as well as enabling Verbmobil to keep track of what was being referred to, this version of the electronic blackboard could provide researchers with useful information of discourse structure, and the trajectories that referents trace through a conversation. These are data that would be a great deal more tedious to collect in other ways.

As we have remarked, the electronic blackboard might be used in a variety of other ways. Suppose, again, that Verbmobil could be made to pick icons from the dictionary in the manner just suggested. Suppose further that a word with several meanings had correspondingly many icons listed against it. The one it displayed on the screen would represent its best guess as to what the word meant in the given context. If that were wrong, the person who made the utterance could indicate the fact, and the machine could replace it with its second choice, and so on. Armed with a device that behaved in this way, Verbmobil would have invaluable additional information on both the meaning and the reference of lexical items. If the device were carefully designed, it seems to us that it need not put an intolerable burden on the early users of Verbmobil.

If the screen before each user had an associated keyboard, it might be used in a more pedestrian way as a communication channel between people engaged in the conversation and Verbmobil. It would be possible, for example, for the machine to put questions to a user through

this medium to clarify any aspect of what he had said. In principle, there is no limit to the disruption that this could cause, but it is possible that, in the early stages, the contribution that it could make to understanding as well as to the data that researchers could collect from it would warrant some measure of disruption.

4.6.2 Assistants

The later suggestions that we made for the use of an electronic blackboard were based on the idea of providing Verbmobil with human assistance with some of its most difficult problems. The problem with the suggestion is that it depends on getting that assistance from someone whose attention is presumably fully engaged in a discussion. But there are other ways of getting this kind of assistance, notably from a person, or persons, who have this as their only responsibility. Suppose that, in an adjacent room, there was some number of people whose job it was to answer questions put to them by Verbmobil. In the degenerate case, they would usurp all the functions of the machine but, in more interesting cases, they could fill limited roles, with potentially interesting effects.

Let us start from the assumption that one assistant is associated with each of the main participants, and that that assistant knows no relevant language other than the one spoken by the person he is responsible for. Beyond that, we will feel free to assign him a variety of different skills, from time to time. A rare skill, but one that would be of inestimable value in the present context, would be that of being able to operate a stenotype machine. Through this medium, Verbmobil could be provided with a transcript in real time of what each person was saying so that its responsibilities would be reduced to those of translating the transcript and generating the result in the form of speech. In other words, the speech-recognition component would be short-circuited altogether. The experimental value of a version of Verbmobil that operated in this manner would be hard to overestimate. The behavior of the device would be essentially indistinguishable from that of product one except that it might be capable of handling a richer domain and a greater variety of interactions. In other words, it might be well in a position to provide the kind of operational experience that will be of the essence in the Verbmobil program.

A considerably lesser skill that the assistants might exercise would be that of simply repeating every utterance made by the person they were assigned to shadow. It would be as though they were simultaneous interpreters from, and into, the same language. The value of this would be simply to supply Verbmobil with a "second opinion" on everything

that was said. The system would take it as its job to find an utterance in the intersection of information derived from the streams of sounds that it received from the two sources, and to make this the basis of its translation.

So far, the two roles in which we have cast the assistants have been based on the notion that the place in which Verbmobil will stand most in need of help is in the interpretation of the speech signal. In our view, however, aids to the translation component would be in as much, if not greater demand, and these would be even easier to fill. We imagine that the assistant assigned to a given conversant, who knows only the language of that conversant, is charged with answering any questions put to him by Verbmobil about what the conversant had said. This ability would, of course be useful in understanding the speech signal, but could also go a long way towards resolving the ambiguities that are such the bane of any translation system. As we have done before, we strongly urge those who will decide the fate of Verbmobil not to regard such suggestions as these as trivializations of their enterprise. They are ways of supporting the early deficiencies of a weak component and thus making it possible to free the enterprise as a whole of the need to proceed at the pace of its least well developed component.

4.7 System design

The design, programming, and documentation of Verbmobil should be uncompromising, never sacrificing elegance or theoretical motivation to performance or convenience. In particular, a regime such as the one based on tasks that are scheduled on an agenda should be used uniformly to control the operation of the main processes, almost all of which should be nondeterministic. An agenda-based system also allows modules to be separated from one another as dictated by the underlying theory while at the same time allowing freedom in the way the space of solutions to the overall problem is searched. The linguistic formalisms used in Verbmobil are important, but care should be taken not to confuse the notions of formalism and notation.

Up until now, we have been looking at Verbmobil from the outside. We have been concerned with the kinds of behavior we expect it to evince, with adjustments that might be made in the environment to make the expected behavior closer to what we want, and with variants of Verbmobil that, though more limited in one way or another, might be able to behave in the desired way more quickly. We now turn to the internals of the system. In this and the following sections, we shall

be interested in the principles of research methodology and system design that we think would be most likely to add value to the basic Verbmobil idea. The balance of the present section will be given over to some general exhortations. We will take up more specific points in later sections.

The following quotations come from Richard A. O'Keefe (1990).

> Elegance is not optional (1)
> What do I mean by that? I mean that in Prolog, as in most halfway decent programming languages, there is no tension between writing a beautiful program and writing an efficient program. If your Prolog code is ugly, the chances are that you either don't understand your problem or you don't understand your programming language, and in neither case does your code stand much chance of being efficient. In order to ensure that your program is efficient, you need to know what it is doing, and if your code is ugly, you will find it hard to analyze.
> Elegance is not optional (2)
> What do I mean by that? I mean that there is no program so trivial that it will not one day need maintaining. A clear straightforward program is going to be easier to maintain than an ugly one.

There is nothing about what O'Keefe is saying here that is specific to Prolog. We believe very strongly that if these exhortations had been heeded by even a small proportion of those who have invested so much of their lives in speech recognition and machine translation, the history of those fields might have been more interesting and an altogether greater source of inspiration to those who came afterwards. We are convinced that absolutely the only hope for managing the complexity of a project on the scale of Verbmobil, while maintaining the flexibility necessary for productive research, is to refuse all compromise on questions of elegance.

4.7.1 Programming

Since Verbmobil will be a collaborative venture, with work going on in parallel in a number of centers, it will be clearly necessary to adopt, not only common programming systems, but also programming conventions, interface specifications, standards of documentation, and the like. These are matters that are most often overlooked in educational institutions where there is relatively little experience with building large and robust systems.

4.7.1.1 Nondeterminism

On the matter of efficiency, the preeminent importance of human over machine efficiency cannot be too often stressed. One place where this is likely to become an issue is on decisions relating to nondeterminism. Almost every part of Verbmobil should be thought of as nondeterministic in the sense that, in order to proceed at all, it must take decisions on the basis of insufficient information; if more information were available, another decision would often have been seen to be better. Verbmobil must therefore be seen as a device that must seek solutions to its problems by searching an enormous space of possibilities. It would be possible to deny this and to force Verbmobil into the deterministic mold, say by introducing heuristics at every turn that purport to supply the information needed to make decisions when the information is actually not there. This has been the common practice in the design of systems in this field, and we are convinced that it would be a disastrous course for Verbmobil to follow.

We do not mean to argue that there is no place in Verbmobil for heuristics. Clearly the space of possibilities that it will have to search is far too great for exhaustive consideration, and even the final goal is not so readily distinguishable as to make it easy to recognize as soon as it is encountered. In fact, the eventual goal of the Verbmobil process is simply a state that is more desirable, on some poorly defined scale, than any others that can be found. So, heuristics are part even of the definition of success within the system.

The point we want to make, and to emphasize very strongly, is that the architecture of the program should be such as to keep the heuristics and the management of the search strategies separate from the parts of the program that actually define those strategies. A standard way of achieving this separation is through a regime according to which tasks are scheduled on an agenda from which they are retrieved for execution in accordance with a system of priorities. The point is that the priorities that tasks receive, the way these priorities are treated by the routines that maintain the agenda, and the crucial decision about when to abandon the search because it is thought that the best solution has probably already been found, are made by parts of the program that are independent of the tasks themselves and the routines that carry them out. With a regime of this sort, questions of whether a depth-first, a breadth-first or, as is more likely, some kind of beam search, is instituted can be determined quite independently of the main algorithms. It is our firm belief that a regime of this kind will amply repay any apparent cost in efficiency or program complexity. We shall

have more to say on the subject of tasks and agendas under the general heading of "modularity" in the next section.

4.7.2 Modularity

Modularity, along with commented code and structured programming, are the closest the programming profession comes to moral imperatives, and we certainly do not wish to be seen as opposing them in any way. Any system in which elegance is not optional must clearly adhere to the principles of modular design. But it is important to understand just how this notion interacts with that of an architecture that is based on tasks and an agenda.

The notion of a module, like many others of undoubted importance, is not well defined. A module can be likened to a room in that the wider the doorway that connects it to other parts of the building, the less convincing its claim to be a room. Modules are connected to the other parts of the programs by *interfaces*, and the most convincing modules are those that are connected by the smallest interfaces. To claim that part of a program is a module is tantamount to claiming that any variables that are mentioned there and also in other parts of the program belong to the interface. For these purposes, we treat as the same the arguments outside the module and the corresponding local variables in a routine inside. The more there are of these variables that enable the parts of the program on each side of the module boundary to influence one another, the less modular that module is.

A program that decomposes naturally into small modules connected by small interfaces will generally be highly regarded by good programmers, mainly because it will be easier to understand. Because it is easier to understand, a modular program will be more robust and easier to maintain. The reason for the relative perspicuity of a modular program is also fairly easy to understand, namely that, in order to understand a given part of it, one needs to understand only the module to which that part belongs and just as much of the remainder of the program as is reflected through the interface to that module.

Although modularity is conducive to the construction of perspicuous, and therefore robust programs, it is possible even for modular programs to be confusing and counterintuitive. This is what happens when, for example, the modules do not correspond to any natural way of breaking down the problem that the program is designed to solve. In the machine translation community and among computational linguists, this is naturally often taken to mean that the dividing lines between modules should follow the dividing lines between parts of linguistic theory. Accordingly, there should be a morphological module,

which should be separate from the syntactic and the phonological modules. Within the syntax module, there should be a grammar module as distinct from a parsing or generation strategy module. Certainly, a program that draws modular boundaries in this way will make more sense to a linguist.

The classic design of a program whose modular design reflects the structure of linguistic theory and which is intended, say, to analyze texts, is therefore one that passes the text first to an orthographic module. When this has finished its work, it provides a version of the text with normalized spelling to the morphological module. This continues working until it has produced an annotated segmentation of the text into morphemes, which constitute the input to the syntactic module, and so on. While this is indeed the classical design, there is nothing about the notion of modularity, or even of theoretically motivated modules, that favors this design especially. The theory says that the modules should be essentially those that we mentioned and, to be modules, they must each share as few variables as possible with the remainder of the program, but nothing says that the work done in one module should be complete before work in another begins. Modularity does indeed often parallel control structure in this way, but this is by no means required by the notion of modularity.

There is much to suggest that human language processing does not follow any such modular structure in its control structure. Consider the following pair of sentences:

(2) The sheep found at the bottom of the ravine by the boy scouts early yesterday morning were near starvation

(3) The sheep found at the bottom of the ravine by the boy scouts early yesterday morning was near starvation

The difference is only that 3 has "was" where 2 has "were." This is possible only because the word "sheep" can function in English as either a singular or a plural noun. A person hearing either of these sentences clearly cannot decide in which of these capacities the word is functioning before reaching the main verb, three words before the end. By this time, however, he has understood that something is being said about one or more sheep that were at the bottom of a ravine where it, or they, was or were, found by some boy scouts early yesterday morning. In other words, by the time the morphological analysis of the verb takes place, the syntactic and semantic analysis of the early part of the sentence is already complete. If the recommendations we are making here are followed, this will also be a possible sequence of

events in Verbmobil. It is in no way at variance with the modular design of the system, motivated by linguistic theory.

In the architecture we propose, there is an agenda which is a collection of items, where each item is a record which, when it is removed from the agenda, causes a process to be carried out, generally in one of the linguistically motivated modules of the system. We use the term "task" to refer indifferently to the record that appears on the agenda and for the process that it initiates; occasionally, we use the term "task record" to distinguish the data structures that go on the agenda. Essentially all the variables in the modules that also appear in other parts of the system are shared by virtue of their appearance in task records. In other words, the task record essentially is the interface. Since these records are expected to be quite modest in size, the scheme makes for a naturally modular design.

However, the control structure of the system as a whole will generally jump back and forth between modules according to the order in which they are taken from the agenda. Since it is intended that the amount of computation done in each task be only very limited, this can be expected to happen very frequently.

4.7.3 Formalism

Since the appearance of "Aspects of the Theory of Syntax" (Chomsky 1965), many linguists have drawn a distinction between "competence," or the knowledge by virtue of which a person is a speaker of a particular language, and "performance," or the capacity that a person has to put such knowledge to work in speech and understanding. There is a close parallel between this distinction and one that computational linguists draw between the data structures that they use to represent grammars and dictionaries, and the programs that they write to exploit those data. Put shortly and oversimply, this is the distinction between data and algorithm.

Early linguistic computing systems, and notably early machine translation systems, did not maintain the difference between data and algorithm so that, when a new grammatical fact was discovered, for example, some part of the program itself had to be modified to accommodate it. Nowadays, it is generally accepted as good programming practice, as well as being an appropriate reflection of linguistic theory, to keep the two kinds of information separate. The program should thus be capable of working with any language provided only that it is furnished with a grammar and a dictionary in the appropriate form, or "formalism."

The term "formalism" plays an important role in linguistics; this is

also true of computational linguistics, where it is also responsible for a certain amount of confusion and misunderstanding. In linguistics, a formalism is associated with the linguistic theory and is a specification of the form the grammars and lexicons should have according to that theory. Often this is done by defining a formal language with the understanding that any expression in the language characterizes a rule or lexical entry. The conventions for interpreting the language specify how the rule is to be applied. The mistake is to assume that this is the only way, or even the best way, to characterize a formalism. In particular, to claim that one is using a particular formalism does not mean that one is using rules or lexical entries that conform to a certain syntax. Furthermore, if two people write rules that conform to the same syntax, this does not imply that they are both using the same formalism, for the interpretation conventions might well be different. We suggest that the notion of a *formalism* is closely related to that of an *interface*, and that a grammar or a lexicon in that formalism is like a module to which that interface gives access.

Consider the following three expressions:

```
a. S --> NP VP
b. S NP VP
c. NP VP S
```

Appropriately read, (a) is a context-free phrase-structure rule written according to the most widely used conventions. According to these conventions, every rule must contain the symbol "→" and there must be just one symbol, consisting of one or more other characters, to its left. Since the position of the "→" symbol is entirely predictable, it can be left out without any loss of information, giving the form in (b). The point of the "→" symbol is to make sure that two kinds of information that the rule contains are distinguished from one another. What is important is the distinction, and not the order in which they are written. Therefore, with a suitable modification of the conventions for interpreting the rules, (c) is also equivalent. We would want to say that what we have just done is to suggest three different *notations* for the same *formalism* and not three different formalisms.

There are probably few people who would disagree with what we have just said, but many would probably disagree with any claim to have a formalism without any particular associated notation, or to have a grammar in a particular formalism without having a particular set of rules set out in some notation or other. But consider the following:

```
rule(Phrase, Members)
```

where *Phrase* is the description of a phrase, and *Members* is a list of phrase descriptions. Either *Phrase* or *Members* must be instantiated.

The first line contains a Prolog term containing two arguments which are further characterized in the words that follow. This would be a natural way of describing the conventions for using a predicate in Prolog and since the original spirit of Prolog requires that predicates be modules within our understanding of the word, what we have just given is the specification of an interface. Now we can go on to say that any pair of items of which this predicate is true will be interpreted, say by a parsing program, as a rule in a context-free phrase-structure grammar. So we can define a grammar by providing definitions of the form

```
rule('S', ['NP', 'VP']).
```

In other words, we have provided a way of defining context-free grammars in a notation that is a subset of the Prolog programming language.

Our argument proceeds just one short step further, to the observation that nothing turns on whether the clauses that define the *rule* predicate are simple unit clauses, like the one above, or whether they involve inference and computation. For example, with appropriate definitions for the *bar* and *comp* predicates, the following trivial schema could generate an indefinite number of rules in accordance with X-bar theory:

```
rule(X1, [X2, Comp]) :-
  bar(X1, X2),
  comp(X2, Comp).
```

This rule schema, and others that one can easily imagine involving more computation, in no way impugn the claim that a particular, narrowly defined formalism is not in fact being used.

Notice that our characterization of the formalism says nothing about whether the total set of rules must be finite in size. We therefore take the view that much of the time that system designers spend on constructing interpreters for particular grammatical and lexical notations is, in fact, wasted. Many very suitable notations come quite readily to hand, and Prolog is an excellent example. Furthermore, the ability to compute sets of rules from underlying properties that the theory says they should have not only makes for a concise grammar, but also provides for the capture of significant generalizations. If formalisms are thought of as modules, and defined through their interfaces rather than through a description of an associated notation, it will be a great deal easier to modify them in the light of experience.

We hope that our remarks on the desirability of an agenda-based architecture and of facing the challenge of nondeterminism squarely will forestall any consideration of the practice, widespread in machine translation, of compiling grammars so as to combine control information with them, thus producing a classical deterministic program. There are many interesting kinds of transformations that can by applied to grammars by processes akin to compilation. This is not one of them.

4.8 Translation

The Verbmobil program should undertake some empirical investigations of translation and interpreting as done by humans. Translation necessarily involves a considerable amount of compromise. The idea of "negotiated translation" is one approach to this that should be investigated.

If the right decisions are made in the establishment of the Verbmobil program, it will establish standards for research in machine translation that the rest of the world will follow. The reasons are clear. There has never been a program of comparable scope, with such ambitious aims, and drawing on such a capable pool of researchers. The project of ATR in Japan failed to achieve its potential because, in our view, the lack of a sufficiently well articulated initial plan led to unnecessary fragmentation of the effort and too timid an approach to the difficult problems. The Verbmobil program should be undertaken in full cognizance of the exact status of the standard approaches to machine translation that have been developed since the 1950's, namely that they have failed. If the most widely used system in actual use today is Systran, it is because it is still the case that a primitive system with a large dictionary is more capable than the best system that can be built with a smaller one. We believe that genuine progress can be made in machine translation, but it will only come by firmly rejecting conventional wisdom and striking out in genuinely new directions. Needless to say, this policy entails risks and there can be no guarantee of ultimate success. The premium for such insurance as can be had will the diversity of the approaches pursued.

4.8.1 Translating and Interpreting

One of the things that stands out as remarkable, even against the remarkable history of machine translation, is the fact that almost no professional translators have played a noticeable part in it. Perhaps this is as it should be, but *a priori* it is difficult to see how the case would be made. The linguists and computer scientists that have de-

signed machine translation systems have taken it for granted that they knew what it meant for one text to be a translation of another—either that, or they were defeatist enough to think that the question would not arise in their work. It is not clear what, if anything, they imagined translators did during the years they spent studying their profession. Perhaps they imagined that they were engaged in matters of such subtlety that they would not concern a machine in the foreseeable future. And perhaps they were right in this to some extent.

We believe that it would be a grave mistake to continue this policy in the Verbmobil program. It was an arrogant and ill-conceived policy in the first place, but this program, as we have said before, is more adventurous than any that have preceded it and it is concerned with interpretation of the spoken word rather than with translating text. Interpretation is not concerned with the fine points of literary style, but it does require very special skills, especially if it is to be done in real time. Interpreting must squarely face the problems of ill-formed input with false starts, filled and unfilled pauses, and a host of other phenomena that are quite foreign to ordinary text.

There is probably some value in asking professional interpreters what they do and how they do it. What we are suggesting, however, is more radical than that. We propose that serious study of interpretation, as it bears on the interests of Verbmobil, be undertaken within the framework of the program. We have already expressed the view that data should be collected for use in the day-to-day development of Verbmobil, an enterprise that will involve simulation of the eventual device by professional interpreters. This data could also be useful for the present purpose, but we do not expect it to be all that is required.

Before discussing this proposal further, it may be important to lay one possible objection to rest. It is often expressed somewhat as follows: "It does not follow from the fact that humans perform this task that there is any virtue in attempting to do it the same way as humans do." We have two responses to this argument. The first is not intended to be as flippant as it may sound at first. It is simply that decades of determined work have failed to make more than minimal progress on the subject of how translation is done. Humans provide us, not only with the only notion we have of what a translation might be, but also with the only existence proof we have of a device that can do it. To persist in looking elsewhere for the solution is nothing short of perverse.

The second reason for using humans as models is this: Language, as we have been at pains to point out in our discussion of what makes translation difficult, is situated. The message is not encoded in the text;

it is at best reflected in the text. To get the whole message, the reader or the hearer must second guess the writer or the speaker constantly. He must read between the lines, and hear between the words, things that were communicated earlier, things that he knows about the writer or the speaker, things about the topic being discussed, and so on. To second guess a human being, one has to think like a human being, an enterprise for which it helps to actually be a human being. Failing that, it is surely advisable to behave as much like a human being as one can.

This is not the place to lay out a complete program of experimentation on interpreters and interpreting. We must be content to point out a few of the more obvious questions that would repay investigation. The first question is surely this: what is the time lag between the original and the interpretation and what factors cause it to change? Like the earliest and most primitive machine translation systems, professional interpreters start producing their rendering of a sentence before it is complete. How do they cope with situations when the word order of the two languages would make them wait until the end of a long sentence to get information that has to come early in the language they are interpreting into? Do they guess? Do they get around it using "place holders" of some sort? How do interpreters handle false starts and—perhaps more interesting—do they make their own? Translators, as opposed to interpreters, seem to produce a wide range of different renderings of the same text. Is this the same, or does the pressure they are under reduce the class of outputs that one gets for a given input? What are the effects of varying the speed of the input? In particular, does this have an effect on the range of outputs one gets? If so, we might have the beginnings of an objective measure of the difficulty of a particular interpretation task. We are convinced that finding the answers to questions like these will greatly influence the course of research in the Verbmobil program.

4.8.2 Translation as Compromise

We recommend the Verbmobil face squarely that fact that translation is inescapably a matter of compromise and adopt an approach in the same spirit as "translation by negotiation."

A tenet of the traditional view of machine translation that we believe has outlived its usefulness is that the possible lines of attack fall into two classes, those that use an interlingua and those that have a transfer component. The terms do not even seem to name notions

that should stand in opposition to one another, interlingua being a representation, and transfer being a process.

The principal objection to the so-called interlingual approach to machine translation has been what we can call the *partial match* argument. Briefly, translation can only exceptionally preserve meaning exactly, without adding or deleting something. Therefore, either the original and the translation must have different interlingual representations, or information must be systematically thrown away in arriving at an interlingual representation so as to increase the likelihood of finding a common form. On the level of meaning, there can be at best a partial match between original and the translation in all but a vanishingly small proportion of the cases.

The principal objection to the so-called transfer approach has been that it fails to recognize that translation is possible only to the extent that there is something that remains invariant under translation and that that is the most important thing that any system can attempt to capture.

The first objection is dealt with if the requirement that a translation must have the same interlingual representation as the original is abandoned in favor of something more flexible. The second goes away when it is shown that what remains invariant under translation is what a reasonable person would expect, namely the common part of the interlingual representations of the original and the translation.

We take it that translation is, in its very essence, a matter of compromise so that the core operation must be that of seeking a string in the target language that is as close in meaning to the original as possible. We therefore contemplate a process that derives a series of interlingual representations for the source, preferably in order of decreasing plausibility. For each of these, it attempts to find target strings with interlingual representations that differ as little as possible from these. At some point, one of these is chosen as being close enough to be allowed to stand as the translation. To the extent that this process is based on a representation of the text that is partial to neither of the languages, it is interlingual. To the extent that the representation of the target string is computed from that of the source, it is a transfer process.

The kind of process we have just sketched is the basis of an approach that we call "negotiated translation" which takes as primary the notion of translation as compromise. We assume that the interlingual representation can, and probably should, reflect information coming from a variety of different levels of abstraction in the analysis of the source and target strings. Certainly, it should contain the re-

sults of the finest lexical and compositional semantic analyses we know how to do. It should also reflect the "information structure" of the strings as well as possible, that is, it should say what information purports to be given, and what new. In the best of all possible worlds, it would probably not contain strictly grammatical information, such as the distinction between active and passive on the grounds that voice is presumably significant only to the extent that it reflects some more meaningful distinction. However, pending a better understanding of just what this distinction is and how it maps onto surface structure, we are prepared to countenance representing such things as active and passive directly.

With all of this information available in a single representation, a number of policies are open to the system. A natural one would be to attempt first to find a translation at as low a level as possible. In other words, if more than one translation is available that preserves the meaning of the original, then choose those that also preserve the information structure and, if there is still more than one candidate, choose the one that preserves the greatest number of grammatical properties. On the other hand, if there is no translation that preserves the meaning, then it becomes necessary to modify the interlingual representation until one is found that is close enough to the original and capable of being translated. This is where the idea of a negotiated translation (see 2.5.2.4) comes in.

4.9 Analysis and Generation

There should be no difference in the grammars and lexica that are used in the analysis and generation parts of the system. Translation by example is a promising notion that should be explored, but approaches based purely on statistics or connectionist schemes are not recommended.

Within the field of machine translation, it is usual to view the process of analyzing the source text as the obverse of the process of generating the target text. This is clear from Vauquois' triangle. In some cases, it has even proved possible to use the same programs for both purposes, though we know of no practical systems in which this is done. However, the part of morphological analysis and generation that is concerned with spelling rules or phonology, using a model based on finite-state transducers is so obviously similar that it is hard to see how one could justify writing two programs. In other places, notably in syntax, a single program might also be used.

Whether corresponding processes in analysis and generation use the

same programs is not a matter of great concern to us. However, it does seem to us that the grammars, or linguistic descriptions, from which these programs work, should be usable indifferently in either process. In other words, the grammar that is used to analyze English when it is the source language should be usable for generation when English is the target language. In the past, separate grammars have often been used. We think the reasons for this have not been good. They are essentially as follows: It is extremely difficult to write a grammar that covers all, and only, the sentences, or allowable utterances, in either language that the system must treat. Indeed, the enterprise may not even be well defined. It seems that the compromises we are therefore forced to make should go in different directions. For the purposes of analysis, we want to be able to analyze anything that may come along even if, in order to do so, we put ourselves in a position to analyze some things that we never expect to see. What we want, in other words is a so-called *covering* grammar. On the other hand, we want to generate only what is clearly acceptable, even at the expense of not being able to guarantee that we could generate every grammatical sentence. For this purpose, rather than a grammar that covers that actual language, we want one that is covered by it.

We believe that these requirements can be better met by a single grammar, possibly with appropriate "annotations," within a system based on tasks and an agenda. An assessment of the emerging value of an analysis or generation is a canonical example of something that should be reflected in the priorities assigned to tasks on the agenda. Suppose that the following is a pair of sentences according to some grammar, and that they both have the same interlingual representation:

(4) a. I explained to him that he should drive very carefully.
b. I explained that he should drive very carefully to him.

The sentence (4b) is allowed only because it is thought that it might appear in source texts, and not because it should ever be used in translations. Notice that, if the priority assigned to (4a) is greater than that of (4b), by however small an amount, then (4a) will *always* be preferred. Just how this will be achieved depends on the details of the grammar; it might be done by assigning different subcategorization frames to the verb "explain" in the lexicon, annotated with different priorities. On the other hand, (4b) might receive a lower priority on the more general principle that it involves more center embedding, a property that is susceptible to measurement.

4.9.1 Learning from Experience

We strongly advocate pursuing policies that would tend to increase the performance of Verbmobil in the light of experience. These include techniques like *translation by example*, where the device maintains and uses its own long-term memory, and statistical information supplied to the system from the outside.

In the limit, translation by example reduces to a technique that uses nothing but the examples in its memory; this is the IBM approach. A number of quite diverse techniques fall under this heading, most of them still at too early a stage in their development to assess with any confidence. We should like to see experimentation on a number of these approaches within the context of Verbmobil. However, we do not think that an approach based entirely on statistical methods, like the one currently under way at IBM is likely to succeed. We base this view mainly on the observation that the relations among items in a text that appear to be important for its translation tend to be spread rather widely through it, unlike those that are most important for speech recognition. The statistical approaches that have been proposed up to now, as well as those based on neural networks, have suffered from a lack of any ability to cope with recursive structure and thus to take cognizance of the structural hypothesis in language. However, as an adjunct to techniques based on linguistics, knowledge representation, and inference, we believe that statistics have a vital role to play. We believe that translation by example could be especially useful in providing a better treatment of collocations and selectional restrictions as well as in helping to resolve ambiguities at all levels in the system.

It is well known that Markov models can be used for what is called *tagging*, that is, the assignment of words in a text to parts of speech. Indeed with an inventory of around a dozen parts of speech, a hidden Markov model based on trigrams makes correct assignments to well over ninety per cent of the words in an average English text. The same technique could clearly also be used to assign weights to the various structures that a syntactic analyzer assigned to a sentence; all other things being equal, the structure based on the most highly valued part-of-speech assignment would be preferred. There is much to be learnt about statistical models of this kind. For example, it may be possible to improve on the results given by the trigram model in various ways. Here is a possible approach.

Suppose that a phrase-structure grammar is being used, involving feature structures of some kind, from which it is possible to extract a context-free core in some way. From this core, it is a straightforward

matter to construct a state-transition network of the kind that underlies LR(k) parsing, but without the assumption of determinism. There are several parsing schemes that would need such a transition network in any case. Now, in a manner substantially like that suggested by Pereira (1991) this can be made the basis of the Markov model and there is reason to hope that it would be more effective in constraining the operation of the parser in the desired way.

4.10 Speech

While always deriving maximum benefit from information directly available from the speech signal, including rhythm and prosody, speech recognition should also profit to the greatest possible extent from constraints imposed by other modules. No new work should be devoted to isolated speech recognition. For speech synthesis, an already existing system should be used if at all possible.

It is the general consensus that speech recognition has been just about as resistant to the onslaughts of science as machine translation. Statistical methods are credited with some recent advances, particularly in the United States in recent years but, as we have noted elsewhere, statistical methods usually cannot be expected to support a steady increase in capability. At first sight, Verbmobil would therefore seem to be saddled with the product of the weaknesses of two already very weak technologies. We believe that this could very easily be the case unless strong measures are adopted to counter the effect from the earliest stages. In particular, we believe that, while aiming for the best independent speech recognition techniques possible, it will be crucial to conduct the research on speech input from the start, and throughout the program, within the total Verbmobil setting, so that it will always be in a position to profit from constraints coming from the machine translation part of the system and from the domain model.

Work on speech recognition has concentrated very directly on the problem that a speech operated typewriter would have to solve, namely that of identifying a string of phonemes. This is not hard to understand—clearly there can be nothing worthy of being called "recognition" without this step. However, constraints on possible segmentations are welcome, wherever they come from, and we therefore think that somewhat more stress should be placed on other parts of the problem, notably rhythm and prosody. Just how important these aspects are for the total recognition problem is not known and would be a worthy goal of experimental research. We suspect that some surprising results might come from experiments in which human subjects were

called upon to recognize connected speech in which rhythmic distinctions had somehow been regularized or in which there was no variation in the F0 contour.

Despite the great amount of work that has been done on speech recognition, especially over the past decade, we believe that it is still far from clear what the size of the segment should be that a system should aim to recognize. Researchers have attempted to work with segments that are somewhat smaller, and also somewhat larger than a phoneme. It seems that there is virtue in a segment type that will call for boundaries to be recognized at points in the signal where the parameters are changing as slowly as possible. On these grounds, we should therefore have to disqualify both phonemes and complete words. As far as we know, no one has worked with segments that begin and end in regions of silence or in the middle of vowels. It might be argued that the number of such segments would be very large and that the consequent increase in perplexity of the system would, in all likelihood, not be repaid. Nevertheless, this still seems to us like an area worthy of investigation.

We have suggested that some work in the Verbmobil program be devoted to some systems other than product one. In particular, it might be possible to implement product two using technology that already exists for recognizing words in isolation and, if this can be done, it would be an excellent avenue to pursue. However, we are strongly of the opinion that no resources should be spent on new work on isolated-word representation. It seems to us that this is a line of attack that long ago reached the point of diminishing returns and that the best ways of attacking it are, in fact, not steps on the way to continuous speech recognition.

Speech synthesis is in a very different stage of development from either speech recognition or machine translation. Though current systems are by no means perfect, they can be made to perform adequately for most practical purposes and we therefore take the view that they should not be a major area of concentration for the Verbmobil program. For synthesis in English, quite adequate systems are available commercially. If this is not the case for German, we suspect that it soon will be. In any case, the Verbmobil program should commit no resources to developing speech synthesis technology unless the intellectual results of the proposed work promise to be just as interesting and important as the practical ones.

4.11 Dialog

A program of empirical research on dialog should be undertaken stressing face-to-face interactions.

We have stressed again and again in this report that the success of Verbmobil rests on its ability to profit from a variety of interlocking constraints on the conversation that it mediates coming from the domain model, intonation, syntax, and so on. Another source of constraints, whose importance has been recognized by researchers for a long time, is the overall structure of the dialog itself. Unfortunately, while linguists have shown considerable interest in the problems of dialog, they have so far been unable to develop formal methods for treating them comparable to those they apply to the grammar of sentences. Although many of those working in the field think of themselves as computational linguists, very little of their work has resulted in algorithms that could be applied in practical situations. We believe that Verbmobil will provide a framework for research in this field that will be close to ideal and that the aims of the program will depend very much on new results being obtained.

The problems of dialog, as we see it, need to be investigated in a wide framework. In particular, for present purposes, it will be important to take into account not just what people say to one another, and with what intonation patterns, but also what passes between them on channels other than speech. In other words, we think that it will be important to study the overall question of face-to-face interactions. We need, for example, to have an assessment of how much of the burden of communication is carried by each of the channels. It has already been shown, by the Edinburgh group and possibly others, that the amount of speech required to accomplish, say, the map task, is less when the participants can see one another's faces. It is possible that a close analysis of the data would give some clue as to what is passing by the nonverbal channel in these cases, and how important it is. In the same way, the functional load carried by intonation could be assessed if participants in a conversation were placed in a setting where variations in the F0 contour of what they said were eliminated.

Some of the most important advances in work on discourse seem to us to have concerned *speech acts* and we think it likely that this work will prove to be especially important in the context of face-to-face interactions where indirect speech acts appear to be particularly prevalent.

Bibliography

Abe, M., K. Shikano, and H. Kuwabara. 1990. Cross-Language Voice Conversion. In *ICASSP 90 Proceedings*, 345–348.

Akiyama, R. 1989. Our Experience in Using Systran. In *Machine Translation Summit*, ed. M. Nagao, 150–151. Tokyo. Omsha Ltd.

Allen, J., M.S. Hunnicutt, and D. Klatt. 1987. *From Text to Speech: the MITalk System.* Cambridge, England: Cambridge University Press.

Alleva, F., R. Bisiani, S. Forin, and R. Lerner. 1986. A Distributed System Architecture for Speech Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1573–1576. New York: IEEE. volume 3.

Alshawi, H., D. J. Arnold, D. M. Carter, J. Lindrop, K. Netter, S. G. Pulman, J-I. Tsujii, and H. Uszkoreit. 1991a. Studies in Machine Translation and Natural Language Processing: Assessment of Computational Linguistics and Machine Translation Formalisms. unpublished.

Alshawi, H., D. Carter, M. Rayner, and B. Gambäck. 1991b. Translation by Quasi Logical Form Transfer. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, held at Berkeley, Calif.*, 161–168. Morristown, N.J. Association for Computational Linguistics.

Alshawi, H., D. M. Carter, J. van Eijk, R. C. Moore, D. B. Moran, F. C. N. Pereira, A. G. Smith, and S. G. Pulman. 1988. 2nd Annual Report, CLE Project. Technical report. Cambridge, England: SRI International.

Alshawi, H., and J. van Eijck. 1989. Logical Forms in the Core Language Engine. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 25–32. Association for Computational Linguistics.

Amano, S., H. Hirakawa, and Y. Tsutsumi. 1989. AS-TRANSAC: The Toshiba Machine Translation System. In *Machine Translation Summit*, ed. M. Nagao, 107–112. Tokyo. Omsha Ltd.

Ananiadou, S. 1987. A Brief Survey of Some Current Operational Systems. In *Machine Translation Today: The State of the Art*, ed. M. King, 171–191.

Edinburgh, Scotland. Edinburgh University Press. Edinburgh Information Technology Series.

Arnold, D., and L. des Tombes. 1987. Basic Theory and Methodology in EUROTRA. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 7, 114–135. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Arnold, D., S. Krauwer, M. Rosen, L. des Tombes, and G. B. Varile. 1986. The ⟨C,A⟩,T Framework in Eurotra: A Theoretical Notation for Machine Translation. In *Proceedings of COLING-86*, 297–303. Bonn.

Arnold, D., and L. Sadler. 1990. The Theoretical Basis of MiMo. *Machine Translation* 5(3):195–222. Dordrecht, Holland: Kluwer.

Austin, A., D. Ayuso, M. Bates, R. Bobrow, R. Ingria, J. Makhoul, P. Placeway, and R. Schwartz. 1991. BBN HARC and DELPHI Results on the ATIS Benchmarks. In *Proceedings of the Speech and Natural Language Workshop*, 112–115. DARPA. Morgan Kaufmann, distributor.

Averbuch, A. 1987. Experiments with the Tangora 20,000 Word Speech Recogniser. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 701–704. New York. IEEE.

Bachenko, J., E. Fitzpatrick, and C. E. Wright. 1986. The Contribution of Parsing to Prosodic Phrasing in an Experimental Text-to-Speech System. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 145–155. Association for Computational Linguistics.

Bahl, L. R., P. F. Brown, P. V. De Souza, and R. L. Mercer. 1989. A Tree Based Stastical Language Model for Natural Language Speech Recognition. *Proceedings of the IEEE Transactions on Acoustics, Speech, and Signal Processing* 37(7):1001–1008. New York: IEEE.

Baker, J. K. 1975a. *Stochastic Modeling as a Means of Automatic Speech Research.* Doctoral dissertation, Carnegie Mellon University, Pittsburgh, Penn.

Baker, J. K. 1975b. The Dragon System—An Overview. *Proceedings of the IEEE Transactions on Acoustics, Speech, and Signal Processing* 1(23):24–29. ASSP-23.

Barnett, J., I. Mani, P. Martin, and E. Rich. 1991a. Reversible Machine Translation: What to Do When the Languages Don't Line Up. In *Reversible Grammar in Natural Language Processing*, ed. T. Strzalkowski, 61–70. Association for Computational Linguistics. ACL SIG Workshop.

Barnett, J., I. Mani, E. Rich, C. Aone, K. Knight, and J. C. Martinez. 1991b. Capturing Language-Specific Semantic Distinctions in Interlingua. In *Machine Translation Summit III: Proceedings*, 25–32. Pittsburgh, Penn.: Carnegie Mellon University.

Barwise, J., and J. Perry. 1983. *Situations and Attitudes.* Cambridge, Mass.: MIT Press.

Baum, L. E. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation of Probalistic Functions of Markov Processes. *Inequalities* 3:1–8.

Bear, J. 1986. A Morphological Recogniser with Syntactic and Phonological Rules. In *Proceedings of COLING-86*, 272–276. Bonn.

Bennett, W. S., and J. Slocum. 1985. The LRC Translation System. In *Machine Translation Systems*, ed. J. Slocum. Chap. 4, 111–140. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Biewer, A., C. Féneyrol, J. Ritzke, and E. Stegentritt. 1985. ASCOF: A Modular Multilevel System for French-German Translation. In *Machine Translation Systems*, ed. J. Slocum. Chap. 2, 49–84. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Bistritz, Y., H. Lev-Ari, and T. Kailath. 1987. Complexity Reduced Lattice Filters for Digital Speech Processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 0021–0024. New York. New York: IEEE. vol. 1.

Black, A., G. Ritchie, S. Pulman, and G. Russell. 1987. Formalism for Morphographemic Decription. In *EACL-3*, 11–17.

Blekman, M. S., and O. R. Polonskaya. 1987. Recent Developments in Commercial Machine Translation in the United States and Canada. *Automatic Documentation and Mathematical Linguistics* 21(5):43–51.

Block, H. U. 1991. Compiling Trace & Unification Grammar for Parsing and Generation. In *Reversible Grammar in Natural Language Processing*, ed. T. Strzalkowski, 100–108. Berkeley, Calif. Association for Computational Linguistics.

Boitet, C. 1987. Research and Development on Machine Translation and Related Techniques at Grenoble University (GETA). In *Machine Translation Today: The State of the Art*, ed. M. King, 133–153. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Boitet, C. 1988. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 93–108. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Boitet, C. 1989. GETA Project. In *Machine Translation Summit*, ed. M. Nagao, 54–65. Tokyo. Omsha Ltd.

Bouillon, P., and K. Boesefeldt. 1991. Applying an Experimental Machine Translation System to a Realistic Problem. In *Machine Translation Summit III: Proceedings*, 45–50. Pittsburgh, Penn.: Carnegie Mellon University.

Boves, L. 1991. The ESPRIT Project Polyglot. In *Speech and Natural Language Workshop*, 7–11. DARPA. Morgan Kaufmann, distributor.

Bresnan, J., and R. Kaplan. 1982. Lexical Functional Grammar: A Formal System for Grammatical Representation. In *The Mental Representation of Grammatical Relations*, ed. J. Bresnan. Cambridge, Mass.: MIT Press.

Brown, P. 1987. *The Acoustic-Modeling Problem in Automatic Speech Recognition.* Doctoral dissertation, Pittsburgh, Penn.: Carnegie Mellon University.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1989. A Statistical Approach to Machine Translation. Research Report RC 14773 (66226). Yorktown Heights, NY: IBM.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L. Mercer. 1992a. An Estimate of an Upper Bound for the Entropy of English. Computational Linguistics, March.

Brown, P. F., V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. 1992b. Class-based N-gram Models of Natural Language. Computational Linguistics, December.

Brown, P. F., J. C. Lai, and R. L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*

Brown, P.F., C-H. Lee, and J. C. Spohr. 1983. Bayesian Adaptation in Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 761–764. vol. 2.

Buchman, B. 1987. Early History of Machine Translation. In *Machine Translation Today: The State of the Art*, ed. M. King, 3–21. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Burk-Seligson, S. 1990. *The Bilingual Courtroom.* Chicago: Univ. of Chicago Press.

Carbonell, J. G., R. E. Cullingford, and A. V. Gershman. 1978. Towards Knowledge-based Machine Translation. In *Proceedings of the Seventh ICCL, COLING-78.*

Carbonell, J. G., R. E. Cullingford, and A. V. Gershman. 1981. Steps Toward Knowledge-based Machine Translation. In *IEEE Transactions on Pattern Analysis and Machine Translation*, 376–392. PAMI.

Carbonell, J. G., and M. Tomita. 1985. New Approaches to Machine Translation. Technical Report CMU-CS-85-143. Pittsburgh, Penn.: Carnegie Mellon University. Department of Computer Science, July.

Carbonell, J. G., and M. Tomita. 1987. Knowledge-Based Machine Translation, the CMU Approach. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 5, 68–89. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Chen, S.-C., J.-W. Wang, J.-S. Chang, and K.-Y. Su. 1991. Arch-Tran: A Corpus-based Statistics-oriented English-Chinese Machine Translation

System. In *Machine Translation Summit III: Proceedings*, 33–40. Pittsburgh, Penn.: Carnegie Mellon University.

Chevalier, M., J. Dansereau, and G. Poulin. 1978. TAUM-METEO: Description du Système. Technical report. Montréal, Canada: Université de Montréal.

Chomsky, N. 1957. *Syntactic Structures.* Janua linguarum. Series minor, No. 4. 's-Gravenhage: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Chomsky, N. 1970. Remarks on Nominalization. In *Readings in English Transformational Grammar*, ed. R. Jacobs and P. Rosenbaum. Waltham, Mass.: Ginn.

Chomsky, N. 1981. *Lectures on Goverment and Binding.* Dordrecht, Holland: Foris.

Choukri, K. 1990. SPRINT: Speech Processing and Recognition Using Integrated Neurocomputing Techniques. *International Journal of Neurocomputing* 2(2).

Chow, Y. L., M.O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, S. Roucos, and R. M. Schwartz. 1987. BYBLOS: The BBN Continuous Speech Recognition System. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 89–92. April. volume 1.

Cochard, J.-L. 1987. A Brief Look at a Typical Software Architecture. In *Machine Translation Today: The State of the Art*, ed. M. King, 117–123. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Cohen, M., H. Murveit, J. Bernstein, P. Price, and M. Weintraub. 1990. The DECIPHER Speech Recognition System. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.*

Cohen, M. H. 1989. *Phonological Structures for Speech Recognition.* Doctoral dissertation, Computer Science Div., University of California, Berkeley.

Cole, R. A. 1986. Phonetic Classification in New Generation Speech Recognition Systems. *Speech Technology* 43–46.

Cole, R. A., M. S. Phillips, B. Brennan, and B. Chigier. 1986a. The CMU Phonetic Classification System. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* volume 3.

Cole, R. A., R. M. Stern, and M. J. Lasry. 1986b. Performing Fine Phonetic Distinctions: Templates versus Features. In *Variability and Invariance in Speech Processes*, ed. J. S. Perkell and D. M. Klatt. Hillsdale, N.J.: L. Erlbaum Associates.

Cole, R. A., R. M. Stern, M. S. Phillips, S. M. Brill, P. Specker, and A. P. Pilant. 1983. Feature-Based Speaker Independent Recognition of English Letters. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Crookston, I. 1990. Machine Translation Formalism Issues and LFG Machine Translation. No idea how to get paper, July.

Cullingford, R. E., and B. A. Onyshkevych. 1987. An Experiment in Lexicon-Driven Machine Translation. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 16, 278–301. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Davis, K.H., R. Biddulph, and S. Balashek. 1952. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America* 24:637–642.

De Roeck, A. 1987. Linguistic Theory and Early Machine Translation. In *Machine Translation Today: The State of the Art*, ed. M. King, 38–57. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Denes, P.B., and M.V. Mathews. 1960. Spoken digit recognition using time-frequency pattern matching. *Journal of the Acoustical Society of America* 32:1450–1455.

Dudley, H., and S. Balashek. 1958. Automatic recognition of phonetic patterns in speech. *Journal of the Acoustical Society of America* 30:721–732.

Dymetman, M., and P. Isabelle. 1988. Reversible Logic Grammars for Machine Translation. In *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.* Pittsburgh, Penn. Carnegie Mellon University.

Dymetman, M., and P. Isabelle. 1990. Grammar Bidirectionality through Controlled Backward Deduction. In *Logic and Logic Grammars for Language Processing*, ed. P. Saint-Dizier and S. Szpakowicz. Chichester, England: Ellis Horwood.

Dymetman, M., P. Isabelle, and F. Perrault. 1990. A Symmetrical Approach to Parsing & Generation. In *Proceedings of COLING-90.* Helsinki.

Erman, L. D., and V. R. Lesser. 1980. The Hearsay-II Speech Understanding System: A Tutorial. In *Trends in Speech Recognition*, ed. W. A. Lea. Chap. 16, 361–381. Prentice-Hall Signal Processing Series. Englewood Cliffs, N.J.: Prentice Hall. Reprinted in Waibel and Lee 1990.

Estival, D. 1990. ELU User Manual. Technical Report 1. Geneva: ISSCO.

Estival, D., A. Ballim, G. Russell, and S. Warwick. 1990. A Syntax and Semantics for Feature-Structure Transfer. In *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language.* Austin, Texas.

Farwell, D., and Y. Wilks. 1990. Ultra: A Multi-lingual Machine Translator. Memoranda in Computer and Cognitive Science MCCS-90-202. New Mexico State University. Computer Research Laboratory.

Feynman, R. P. 1985. *Surely You're Joking, Mr. Feynman.* New York: W. W. Norton and Co.

Fillmore, C. J. 1968. The Case for Case. In *Universals in Linguistics Theory*, ed. E. Bach and R. T. Harms. New York: Holt, Rinehart and Winston.

Fissore, L., P. Laface, G. Micca, and R. Pieraccini. 1989. A Word Hypothesizer for a Large Vocabulary Continuous Speech Understanding System. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* New York: IEEE.

Furui, S. 1986. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 1:52–59. ASSP-34.

Gale, W. A., and K. W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* 177–184.

Garvin, P. L. 1967. The Georgetown-IBM Experiment of 1954: An Evaluation in Retrospect. In *Papers in Linguistics in Honor of Leon Dostert,* ed. W. M. Austin. 45–56. The Hague: Mouton.

Gazdar, G., E. Klein, G. Pullum, and I.Sag (ed.). 1985. *Generalized Phrase Structure Grammar.* Cambridge, Mass.: Harvard University Press.

Glass, J. R., and V. W. Zue. 1988. Multi-level Acoustic Segmentation of Continuous Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 429–432. New York: IEEE. volume 1.

Gold, B., and R. P. Lippmann. 1988. A Neural Network for Isolated-Word Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 44–47. New York: IEEE. volume 1.

Grice, H. P. 1975. Logic and Conversation. In *Syntax and Semantics,* ed. P. Cole and J. L. Morgan. 41–58. Seminar Press. Originally presented as the William James Lectures, Harvard University, 1967.

Guzmán de Rojas, I. 1988. ATAMIRI—Interlingual Machine Translation Using the Aymara Language. In *New Directions in Machine Translation,* ed. D. Maxwell, K. Schubert, and T. Witkam, 123–130. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Hanazawa, T., K. Kita, S. Nakamura, T. Kawabata, and K. Shikano. 1990. ATR HMM-LR Continuous Speech Recognition System. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* Reprinted in Waibel and Lee 1990.

Hasida, K. 1991. Common Heuristics for Parsing, Generation, and Whatever. In *Reversible Grammar in Natural Language Processing,* ed. T. Strzalkowski, 81–90. Association for Computational Linguistics. ACL SIG Workshop.

Hataoka, N., and A. H. Waibel. 1989. Speaker-Independent Phoneme Recognition on TIMIT Database Using Integrated Time-Delay Neural Networks (TDNNs). Technical Report CMU-CMT-89-115. Pittsburgh, Penn.: Carnegie Mellon University.

Haton, J.-P. 1984. *Knowledge-based and Expert Systems in Automatic Speech Recognition.* New Systems and Architectures for Automatic Speech

Recognition and Synthesis. Netherlands: Dordrecht.

Hirai, A., and A. Waibel. 1989. Phoneme-Based Word Recognition by Neural Network—A Step Toward Large Vocabulary Recognition. Technical Report CMU-CMT-89-114. Pittsburgh, Penn.: Carnegie Mellon University.

Hirakawa, H., H. Nogami, S.-Y. Amano, M. Kameyama, and J. R. Hobbs. 1991. EJ/JE Machine Translation System ASTRANSAC—Extensions toward Personalization. In *Machine Translation Summit III: Proceedings*, 73–80. Pittsburgh, Penn.: Carnegie Mellon University.

Hirschberg, J., and J. Pierrehumbert. 1986. The Intonational Structuring of Discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 136–144. Association for Computational Linguistics.

Hirschmann, L. 1986. Discovering Sublanguage Structures. In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, ed. R. Grishman and R. Kittredge. 211–234. Hillsdale, New Jersey: L. Erlbaum Associates.

Hobbs, J. R., W. Croft, T. Davies, D. Edwards, and K. Laws. 1986. Commonsense Metaphysics and Lexical Semantics. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 231–240. Association for Computational Linguistics.

Hobbs, J. R., and M. Kameyama. 1990. Translation by Abduction. Technical Report 484. Menlo Park, Calif.: SRI International.

Hobbs, J. R., M. Stickel, D. Appelt, and P. Martin. 1990. Interpretation as Abduction. Technical Report 499. Menlo Park, Calif.: SRI International.

Hon, H.-W., and K.-F. Lee. 1991. Recent Progress in Robust Vocabulary-Independent Speech Recognition. In *Proceedings of the Speech and Natural Language Workshop*. DARPA. Morgan Kaufmann, distributor.

Huang, W., R. Lippmann, and B. Gold. 1988. A Neural Net Approach to Speech Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 99–102. New York: IEEE. volume 1.

Huang, X. 1988. Semantic Analysis in XTRA, An English-Chinese Machine Translation System. *Computers and Translation* 3(2):101–120.

Huang, X. D. 1991. A Study of Speaker-Adaptive Speech Recognition. In *Proceedings of the Speech and Natural Language Workshop*. DARPA. Morgan Kaufmann, distributor.

Huang, X. D., and M. A. Jack. 1989. Semi-Continuous Hidden Markov Models for Speech Recognition. *Computer Speech and Language* 3(3):239–252. Reprinted in Waibel and Lee 1990.

Hutchins, W. J. 1988. Recent Developments in Machine Transaltion. In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 7–64. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Hutchins, W.J. 1986. *Machine Translation: Past, Present, Future.* Computers and Their Applications. Chichester, England: Ellis Horwood Limited.

IBM Speech Recognition Group. 1985. A Real-Time, Isolated-Word, Speech Recognition System for Dictational Transcription. In *IEEE Interational Conference on Acoustics, Speech, and Signal Processing.* volume 2.

Ikehara, S., S. Shirai, A. Yokoo, and H. Nakaiwa. 1991. Toward an Machine Translation System Without Pre-Editing: Effects of a New Method in ALT-J/E. In *Machine Translation Summit III: Proceedings*, 101–106. Pittsburgh, Penn.: Carnegie Mellon University.

Isabelle, P., and L. Bourbeau. 1985. TAUM-AVIATION: Its Technical Features and Some Experimental Results. In *Machine Translation Systems*, ed. J. Slocum. Chap. 7, 237–263. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Isabelle, P., M. Dymetman, and E. Mackiovitch. 1988. CRITTER: A Translation System for Agricultural Market Reports. In *Proceedings of the 12th International Conference on Computational Linguistics.* Budapest.

Isabelle, P., M. Dymetman, and E. Markovitch. 1986. Transfer and Modularity. In *Proceedings of COLING-86*, 115–117. Bonn.

Iso, K., and T. Watanabe. 1990. Speaker-Independent Word Recognition Using a Neural Prediction Model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* Reprinted in Waibel and Lee 1990.

Itakura, F. 1975. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoutics, Speech, and Signal Processing* 1:67–72. ASSP-23.

Jain, A. N. 1991. Parsing Complex Sentences with Structured Connectionist Networks. *Neural Computation* 3(1):110–120.

Jain, A. N., A. E. McNair, A. Waibel, H. Saito, A. G. Hauptmann, and J. Tebelskis. 1991. Connectionist and Symbolic Processing in Speech-to-Speech Translation: The Janus System. In *Machine Translation Summit III: Proceedings*, 113–118. Pittsburgh, Penn.: Carnegie Mellon University.

Japan Electronic Dictionary Research Institute LTD. 1988a. Concept Dictionary. Technical Report TR-009. EDR, November. Ver. 1.

Japan Electronic Dictionary Research Institute LTD. 1988b. Word Dictionary. Technical Report TR-008. EDR, November. Ver. 2.

Jäppinen, H., L. Kulikov, and A. Ylä-Rotiala. 1991. KIELKONE Machine Translation Workstation. In *Machine Translation Summit III: Proceedings*, 107–112. Pittsburgh, Penn.: Carnegie Mellon University.

Jelinek, F. 1976. Continuous Speech Recognition by Statistical Methods. In *Proceedings of the IEEE 64*, 532–556. volume 4.

Jelinek, F. 1985a. A Real-Time, Isolated-Word, Speech Recognition System for Dictation Transcription. In *Proceedings of the IEEE International*

Conference on Acoustics, Speech, and Signal Processing, 858–861. New York: IEEE.

Jelinek, F. 1985b. The Development of an Experimental Discrete Dictation Recognizer. *Proceedings of the IEEE* 73(11):1616–1624. Reprinted in Waibel and Lee 1990.

Jelinek, F., and R. L. Mercer. 1980. Pattern Recognition in Practice. In *Interpolated Estimation of Markov Source Parameters from Sparse Data.* 381–397. North Holland Publishing Company.

Jelinek, F., B. Merialdo, S. Roukas, and M. Strauss. 1991. A Dynamic Language Model for Speech Recognition. In *Proceedings of the Speech and Natural Language Workshop*, 293–295. DARPA. Morgan Kaufmann, distributor.

Jensen, K., and G. E. Heidorn. 1982. The Fitted Parse: 100 Percent Parsing Capability in a Syntactic Grammar of English. Research Report 42958. New York: IBM.

Johnson, R., and M. Rosner. 1987. Machine Translation and Software Tools. In *Machine Translation Today: The State of the Art*, ed. M. King, 154–167. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Johnson, R. L., and P. Whitelock. 1987. Machine Translation. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 8, 136–144. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Kaji, H. 1989. HICATS/JE: A Japanese to English Machine Translation System Based on Semantics. In *Machine Translation Summit*, ed. M. Nagao, 101–106. Tokyo. Omsha Ltd.

Kaji, H., A. Koizumi, and K. Yoshimura. 1989. A Semantics-Based Machine Translation System from Japanese into English. *Future Computing Systems* 2(3):247–259.

Kameyama, M., R. Ochitani, and S. Peters. 1991. Resolving Translation Mismatches With Information Flow. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 193–200. Berkeley, Calif. Association for Computational Linguistics.

Kamp, H. 1981. A Theory of Truth and Semantic Representation. In *Formal Methods in the Study of Language*, ed. J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof. 277–322. Mathematical Center Tracts. Amsterdam: Mathematisch Centrum.

Kaplan, R., K. Netter, J. Wedekind, and A. Zaenen. 1989. Translation by Structural Correspondences. In EACL-*4*. Manchester, England.

Kay, M. 1980. The Proper Place of Men and Machines in Language Translation. Technical Report CSL-80-11. Palo Alto, Calif.: Xerox Palo Alto Research Center, October.

Kay, M. 1984. Functional Unification Grammar: A Formalism for Machine Translation. In *Proceedings of COLING-84*, 75–78.

Kay, M. 1986. Machine Translation Will Not Work. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 268. Association for Computational Linguistics.

Kernighan, B. W., and D. M. Ritchie. 1978. *The C Programming Language.* Englewood, N.J.: Prentice-Hall. second edition.

Kimura, S. 1990. 100,000-Word Recognition Using Acoustic-Segment Networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 61–64.

King, M. 1986. Machine Translation Already Does Work. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 268–269. Association for Computational Linguistics.

King, M., and S. Perschke. 1987. EUROTRA. In *Machine Translation Today: The State of the Art*, ed. M. King, 373–391. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Kitano, H. 1990. Empirical Studies on the Utility of Genetic Algorithms for Training and Designing of Neural Networks. Technical Report CMU-CMT-90-120. Pittsburgh, Penn.: CMU, August.

Kitano, H. 1991. Toward High Performance Machine Translation: Preliminary Results from Massively Parallel Memory-Based Translation on SNAP. In *Machine Translation Summit III: Proceedings*, 93–100. Pittsburgh, Penn.: Carnegie Mellon University.

Kittredge, R. I. 1987. The Significance of Sublanguage for Automatic Translation. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 4, 59–67. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Knowles, F., G. Jelinek, and Wood M. M. 1989. Alvey Project. In *Machine Translation Summit*, ed. M. Nagao, 45–49. Tokyo. Omsha Ltd.

Kobayashi, Y., and Y. Niimi. 1986. A New Architecture of Speech Understanding Systems—A Hybrid of a Hierarchical and a Network Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1585–1588. New York: IEEE.

Kosaka, M., V. Teller, and R. Grishman. 1988a. A Comparative Study of Japanese and English Sublanguage Patterns. Technical Report 15. New York University: Courant Institute of Mathematical Sciences, June.

Kosaka, M., V. Teller, and R. Grishman. 1988b. A Sublanguage Approach to Japanese-English Machine Translation. In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 109–122. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Koskenniemi, K. 1984. A General Computational Model for Word-form Recognition and Production. In *Proceedings of COLING-84*, 178–181.

Kowalski, R. 1979. *Logic for Problem Solving.* New York: Elsevier North Holland.

Kubala, F., S. Austin, C. Barry, J. Makhoul, P. Placeway, and R. Schwartz. 1991. BYBLOS Speech Recognition Benchmark Results. In *Proceedings*

of the Speech and Natural Language Workshop, 14–27. DARPA. Morgan Kaufmann, distributor.

Kubala, F., Y. Chow, A. Derr, M. Feng, O. Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandegrift. 1988. Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word REsource Management Database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 291–294. New York: IEEE. volume 1.

Kubala, F., and R. Schwartz. 1990. A New Paradigm for Speaker-Independent Training and Speaker Adaptation. *Proceedings of the Speech and Natural Language Workshop* 306–310. Morgan Kaufmann, distributor.

Kugler, M., G. Heyer, R. Kese, B. von Kleist-Retzow, and G. Winkelmann. 1991. The Translator's Workbench: An Environment for Multi-Lingual Text Processing and Translation. In *Machine Translation Summit III: Proceedings*, 81–84. Pittsburgh, Penn.: Carnegie Mellon University.

Landsbergen, J. 1987. Isomorphic Grammars and Their Use in the ROSETTA Translation System. In *Machine Translation Today: The State of the Art*, ed. M. King, 351–372. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Leavitt, J., E. Nyberg, S. Nirenburg, and C. Defrise. 1991. DIOGENES-90. Technical Report CMU-CMT-91-123. Pittsburgh, Penn.: Carnegie Mellon University, April.

Lee, K.-F. 1989. *Automatic Speech Recognition: The Development of the* SPHINX *System.* Kluwer International Series in Engineering and Computer Science, Vol. 62. Dordrecht: Holland: Kluwer.

Lee, K.-F., and H. W. Hon. 1988. Large-Vocabulary Speaker-Independent Continuous Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.*

Lehrberger, J., and L. Bourbeau. 1988. *Machine Translation: Linguistic Characteristics of Machine Translation Systems and General Methodology of Evaluation.* Linguisticae Investigationes: Supplementa, Studies in French and General Linguistics, Vol. 15. Amsterdam: John Benjamins Publishing Company. first edition.

Lennig, M., V. Gupta, P. Kenny, P. Mermelstein, and D. O'Shaughnessy. 1990. An 86,000-Word Recognizer Based on Phonemic Models. In *Proceedings of the Speech and Natural Language Workshop*, 391–396. DARPA. Morgan Kaufmann, distributor.

Lesser, V. R., R. D. Fennell, L. D. Erman, and R. D. Reddy. 1975. The Hearsay II Speech Understanding System. *Proceedings of the IEEE Transations on Acoustics, Speech, and Signal Processing* 1:11–24. ASSP-23.

Leung, H. C., and V. W. Zue. 1988. Some Phonetic Recognition Experiments Using Artificial Neural Nets. In *Proceedings of the IEEE International*

Conference on Acoustics, Speech, and Signal Processing, 422–425. New York: IEEE. volume 1.

Logan, J.S., B.G. Greene, and D.B. Pisoni. 1989. Segmental Intelligibility of Synthetic Speech Produced by Rule. *Journal of the American Statistical Association* 86(2):566–581.

Lowerre, B. T. 1976. *The Harpy Speech Recognition System.* Doctoral dissertation, Pittsburgh, Penn.: Carnegie Mellon University.

Lowerre, B. T. 1980. The Harpy Speech Understanding System.

Lubensky, David. 1988. Learning Spectral-Temporal Dependencies Using Connectionist Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 418–421. New York: IEEE. volume 1.

Lytinen, S. L. 1987. Integrating Syntax and Semantics. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 17, 302–316. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Maas, H.-D. 1987. The Machine Translation System SUSY. In *Machine Translation Today: The State of the Art*, ed. M. King, 209–246. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Mann, J. S. 1987. Get Smart! Industrial Strength Language Processing from Smart Communications. *Language and Technology* 3:12–15.

Marchuk, Y. N., and I. B. Vorozhtsova. 1985. Ariane-78 Machine Translation Sytem and Prospects for its Use. *Automatic Documentation and Mathematical Linguistics* 19(5):40–43.

Matsunaga, S., S. Sagayama, S. Homma, and S. Furui. 1990. A Continuous Speech Recognition System Based on a Two-Level Grammar Approach. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, 589–592.

Maxwell, D., K. Schubert, and T. Witkam (ed.). 1988. *New Directions in Machine Translation.* Distributed Language Translation 4, Dordrecht, Holland. Foris.

McCord, M. C. 1986. Design of LMT: A Prolog-based Machine Translation System. Research Report RC 11801 (53031). Yorktown Heights, N.Y.: IBM, March.

McCord, M. C. 1988. Design of a Prolog-based Machine Translation System. Research Report RC 13536 (60496). Yorktown Heights, N.Y.: IBM, February. Revision of 1986 paper of same title.

McCord, M. C. 1989. A New Version of the Machine Translation System LMT. Research Report RC 14710 (65948). Yorktown Heights, N.Y.: IBM, June.

McDonald, D. D. 1987. Natural Language Generation: Complexities and Techniques. In *Machine Translation: Theoretical and Methodological Is-*

sues, ed. S. Nirenburg. Chap. 12, 192–224. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Melby, A. 1987a. Creating an Environment for the Translator. In *Machine Translation Today: The State of the Art*, ed. M. King, 124–132. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Melby, A. 1987b. On Human-Machine Interaction in Translation. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 9, 145–154. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Miller, L. G., and S. E. Levinson. 1988. Syntatic Analysis for Large Vocabulary Speech Recognition Using a Context-free Covering Grammar. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 271–274. New York: IEEE. volume 1.

Mitamura, T. 1991. An Efficient Interlingua Translation System for Multilingual Document Production. In *Machine Translation Summit III: Proceedings*, 55–62. Washington D. C.

Montague, R. M. 1974. *Formal Philosophy.* New Haven, Conn.: Yale University Press.

Moore, R. C. 1989. Unification-Based Semantic Interpretation. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 33–41. Association for Computational Linguistics.

Morii, S., K. Niyada, S. Fujii, and M. Hoshimi. 1985. Large Vocabulary Speaker-Independent Japanese Speech Recognition System. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 866–869. New York: IEEE.

Muraki, K. 1989. PIVOT: Two-Phase Machine Translation System. In *Machine Translation Summit*, ed. M. Nagao, 113–115. Tokyo. Omsha Ltd.

Murveit, H., J. Butzberger, and M. Weintraub. 1991. Speech Recognition in SRI's Resource Management and ATIS Systems. In *Proceedings of the Speech and Natural Language Workshop*, 94–99. DARPA. Morgan Kaufmann, distributor.

Murveit, H., and M. Weintraub. 1988. 1000-Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 115–118. New York: IEEE. volume 1.

Nagao, M. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In *Artificial and Human Intelligence*, ed. A. Elithorn and R. Banerji. Amsterdam. North-Holland.

Nagao, M. 1987. Role of Structural Transformation in a Machine Translation System. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 15, 262–277. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Nagao, M. (ed.). 1989a. *A Japanese View of Machine Translation in Light of the Considerations and Recommendations Reported by* ALPAC, *U. S. A.* Japan Electronic Industry Development Association.

Nagao, M. 1989b. *Machine Translation: How Far Can It Go?* Oxford: Oxford University Press. translated by N. D. Cook.

Nagao, M. 1991. Current State and Problems of Machine Translation. Department of Electrical Engineering, Kyoto University.

Nagao, M., J.-I. Tsujii, and J.-I. Nakamura. 1985. The Japanese Government Project for Machine Translation. In *Machine Translation Systems*, ed. J. Slocum. Chap. 5, 141–186. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Nagata, M., and K. Kogure. 1990. HPSG-Based Lattice Parser for Spoken Japanese in a Spoken Language Translation System. In *Proceedings of the European Conference on AI.* London. Pitman.

National Research Council. 1966. Language and Machines: Computers in Translation and Linguistics. Technical report. Washington D.C.: Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences. publication 1416.

Nerbonne, J., and J. Laubsch. 1991. An Overview of NLL. Technical report. Hewlett Packard.

Newell, A. et al. 1973. *Speech understanding systems: Final report of a study group.* Amsterdam: North-Holland Pub. Co.

Ney, H. 1987. Dynamic Programming Speech Recognition Using a Context-Free Grammar. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 0069–0072. New York: IEEE. volume 1.

Niemann, H., A. Brietzmann, U. Ehrlich, and G. Sagerer. 1986. Representation of a Continuous Speech Understanding and Dialog System in a Homogeneous Semantic Net Architecture. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1581–1584. New York: IEEE. volume 3.

Nirenburg, S. 1987a. A Distributed Generation System for Machine Translation: Background, Design, Architecture and Knowledge Structures. Technical Report CMU-CMT-87-102. Pittsburgh, Penn.: CMU, June.

Nirenburg, S. 1987b. Knowledge and Choices in Machine Translation. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 1, 1–21. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Nirenburg, S., and K. Goodman (ed.). 1991. *The KBMT Project: A Case Study in Knowledge-Based Machine Translation.* San Mateo, Calif.: Morgan Kaufmann.

Nirenburg, S., and A. K. Joshi (ed.). 1987. *Machine Translation: Theoretical and Methodological Issues.* Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Nirenburg, S., R. McCardell, E. Nyberg, P. Werner, S. Huffman, E. Kenschaft, and I. Nirenburg. 1988a. DIOGENES-88. Technical Report CMU-CMT-88-107. Pittsburgh, Penn.: Carnegie Mellon University, June.

Nirenburg, S., R. McCardell, E. Nyberg, P. Werner, S. Huffman, E. Kenschaft, and I. Nirenburg. 1988b. Diogenes-88. Technical Report CMU-CMT-88-107. Pittsburgh, Penn.: Carnegie Mellon University, June.

Nirenburg, S., V. Raskin, and A. B. Tucker. 1987. The Structure of Interlingua in TRANSLATOR. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 6, 90–113. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Norvig, P., and R. Wilensky. 1990. A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation. In *Proceedings of COLING-90.* Helsinki.

Ochsman, R. B., and A. Chapanis. 1974. The Effects of 10 Communication Modes on the Behavior of Teams During Cooperative Problem Solving. *International Journal of Man-Machine Studies* 6:579–619.

O'Keefe, R. 1990. *The Craft of Prolog.* Cambridge, Mass.: MIT Press.

Okumura, A., K. Muraki, and S. Akamine. 1991. Multi-lingual Sentence Generation from the PIVOT Interlingua. In *Machine Translation Summit III: Proceedings*, 67–72. Pittsburgh, Penn.: Carnegie Mellon University.

Oshio, T. 1989. Applied Testing of HICATS/JE for Japanese Patent Abstracts. In *Machine Translation Summit*, ed. M. Nagao, 140–144. Tokyo. Omsha Ltd.

Oviatt, S. L., P. R. Cohen, and A. M. Podlozny. 1990. Spoken Language and Performance During Telephone Interpretation. In *Proceedings of the International Conference on Spoken Language Processing.*

Oviatt, S.L., and P.R. Cohen. 1992. Spoken Language in Interpreted Telephone Dialogues. *Computer Speech and Language* 6(3):277–302.

Paul, D. B. 1991. New Results with the Lincoln Tied-Mixture HMM CSR System. In *Proceedings of the Speech and Natural Language Workshop.* Morgan Kaufmann, distributor, DARPA.

Peckham, J. 1982. A Real Time Hardware Continuous Speech Recognition System. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Peckham, J. 1991. Speech Understanding and Dialogue over the Telephone: An Overview of the ESPRIT SUNDIAL Project. In *Proceedings of the Speech and Natural Language Workshop,* 14–27. DARPA. Morgan Kaufmann, distributor.

Peirce, C. S. 1932. Lectures on Pragmatism. In *Elements of Logic: Collected Papers of C. S. Peirce*, ed. C. Hartshorne and P. Weiss. Chap. 6 and 7. Cambridge, Mass.: Harvard University Press. volume 5.

Pereira, F. C. N., and D. H. D. Warren. 1983. Parsing as Deduction. In *Proceedings of the 21st Annual Meeting of the Association for Computational*

Linguistics, 137–144. Morristown, N.J. Association for Computational Linguistics.

Perlmutter, D. (ed.). 1983. *Studies in Relational Grammar.* Chicago: University of Chicago Press.

Perschke, S. 1989. EUROTRA Project. In *Machine Translation Summit*, ed. M. Nagao, 37–44. Tokyo. Omsha Ltd.

Petitpierre, P. 1987. Software Background for Machine Translation: a Glossary. In *Machine Translation Today: The State of the Art*, ed. M. King, 111–116. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Phillips, M. S., J. Glass, and V. Zue. 1991. Modelling Context Dependency in Acoustic-Phonetic and Lexical Representations. In *Proceedings of the Speech and Natural Language Workshop*, 71–76. DARPA. Morgan Kaufmann, distributor.

Piron, C. 1988. Learning from Translation Mistakes. In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 233–242. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Poesio, M. 1991. Relational Semantics and Scope Disambiguation. In *Situation Theory and Its Applications II*, ed. M. Gawron, J. Barwise, G. Plotkin, and S. Tutiya. Lecture Notes, no. 26. Stanford, Calif.: CSLI Publications.

Pollack, I., and J.M. Pickett. 1963. Intelligibility of Excerpts from Conversational Speech. *Language and Speech* 6:165–171.

Pollard, C., and I. Sag. 1987. *Information-Based Syntax and Semantics.* Lecture Notes, no. 13. Stanford, Calif.: CSLI Publications.

Pullum, G., and G. Gazdar. 1982. Natural Languages and Context Free Languages. *Linguistics and Philosophy* 4:471–504.

Pustejosky, J. 1987. An Integrated Theory of Discourse Analysis. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 11, 168–191. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Rabiner, L. R., and B. H. Juang. 1986. An Introduction to Hidden Markov Models. *IEEE* ASSP *Magazine* 1(3):4–16.

Rabiner, L. R., S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon. 1979. Speaker-Independent Recognition of Isolated Words Using Clustering Techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27(4):336–349. Reprinted in Waibel and Lee 1990.

Rabiner, L. R., J. G. Wilpon, and F. K. Soong. 1988. High Performance Connected Digit Recognition, Using Hidden Markov Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 119–122. New York: IEEE. volume 1.

Rabiner, L.R., S.E. Levinson, and M.M. Sondhi. 1984. On the Performance of Isolated Word Speech Recognizers Using Vector Quantization

and Temporal Energy Contours. *AT&T Bell Laboratories Technical Journal* 63(7):1245–1260.

Raskin, V. 1987. Linguistics and Natural Language Processing. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 3, 42–58. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Reddy, D. R. 1976. Speech Recognition by Machine: A Review. *IEEE Proceedings* 64(4):502–531. Reprinted in Waibel and Lee 1990.

Reddy, D. R., and V. Zue. 1983. Recognizing Continuous Speech Remains an Illusive Goal. *IEEE Spectrum* 84–87:

Reder, L. M. 1983. What Kind of Pitcher Can a Teacher Fill? Effects of Priming in Sentence Comprehension. *Journal of Verbal Learning and Verbal Behavior* 22(2):189–202.

Riley, M. D., and A. Ljolje. 1991. Lexical Access with a Statistically-derived Phonetic Network. In *Proceedings of the Speech and Natural Language Workshop*, 289–292. DARPA. Morgan Kaufmann, distributor.

Rimon, M., M. McCord, U. Schwall, and P. Martinez. 1991. Advances in Machine Translation Research in IBM. In *Machine Translation Summit III: Proceedings*, 11–18. Pittsburgh, Penn.: Carnegie Mellon University.

Ryan, J. P. 1989. Systran: A Machine Translation System to Meet User Needs. In *Machine Translation Summit*, ed. M. Nagao, 116–121. Tokyo. Omsha Ltd.

Sadler, L., I. Crookston, D. Arnold, and A. Way. 1990. LFG and Translation. In *Third International Conference on Theoretical and Methodological Issues in Machine Translation*. Linguistics Research Center, Austin, Texas.

Sadler, L., I. Crookston, and A. Way. 1989. Co-description, Projection, and 'Difficult' Translation. In *Working Papers in Language Processing*. University of Essex. Department of Language and Linguistics.

Sadler, V. 1989a. The Bilingual Knowledge Bank (BKB). Technical report. Utrecht, Holland: BSO Research.

Sadler, V. 1989b. Translating with a Simulated Bilingual Knowledge Bank (BKB). Technical report. Utrecht, Holland: BSO Research.

Saeki, N. 1989. Machine Translation System at Mazda. In *Machine Translation Summit*, ed. M. Nagao, 136–139. Tokyo. Omsha Ltd.

Saito, H., and M. Tomita. 1986. On Automatic Composition of Stereotypic Documents in Foreign Languages. In *Proceedings of the 1st International Conference on Applications of Artificial Intelligence to Engineering Problems*, 179–192. Berlin. Springer-Verlag.

Sakoe, H. 1979. Two-Level DP-Matching—A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27(6):588–595. Reprinted in Waibel and Lee 1990.

Sakoe, H., and S. Chiba. 1971. A Dynamic Programming Approach to Continuous Speech Recognition. In *Proceedings of the International Congress on Acoustics*. Budapest, Hungary. Akademiai Kiado. Paper 20C–13.

Sakoe, H., R. Isotani, K. Yoshida, K. Iso, and T. Watanabe. 1989. Speaker-Independent Word Recognition Using Dynamic Programming Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 29–32. Reprinted in Waibel and Lee 1990.

Sakurai, K., M. Ozeki, and Y. Nishihara. 1991. Machine Translation Application for the Translation Agency. In *Machine Translation Summit III: Proceedings*, 63–66. Pittsburgh, Penn.: Carnegie Mellon University.

Sampson, C. 1987. Machine Translation: A Nonconformist's View of the State of the Art. In *Machine Translation Today: The State of the Art*, ed. M. King, 91–108. Edinburgh Information Technology Series. Edingburgh, Scotland. Edinburgh University Press.

Sato, S., and M. Nagao. 1989. Memory-based Translation. *Information Processing Society of Japan. Working Group - NL 9(70).* (In Japanese).

Sato, S., and M. Nagao. 1990. Toward Memory-based Translation. In *Proceedings of COLING-90.* Helsinki.

Schneider, T. 1989. The Metal System, Status 1987. In *Machine Translation Summit*, ed. M. Nagao, 122–127. Tokyo. Omsha Ltd.

Schneider, T. 1991. The METAL System. In *Machine Translation Summit III*, 41–44. Pittsburgh, Penn.: Carnegie Mellon University.

Schroeder, M. R. 1977. Machine Processing of Acoustic Signals: What Machines Can Do Better than Organisms (and Vice Versa). In *Recognition of Complex Acoustic Signals*, ed. T. H. Bullock. Life Sciences Research Report. Berlin. Abakon Verlagsgesellschaft [in Komm.].

Schubert, K. 1987. *Metataxis: Contrastive Dependency Syntax for Machine Translation.* Distributed Language Translation 2. Dordrecht, Holland: Foris.

Schubert, K. 1988. The Architecture of DLT—Interlingual or Double Direct. In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 131–144. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Schütz, J., G. Thurmair, and R. Cencioni. 1991. An Architecture Sketch of Eurotra-II. In *Machine Translation Summit III: Proceedings*, 3–10. Pittsburgh, Penn.: Carnegie Mellon University.

Scott, B. E. 1989. The Logos System. Paper delivered at the Machine Translation Summit Conference. Munich. August, 1989.

Sells, P. 1985. *Lectures on Contemporary Syntactic Theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar.* Lecture Notes, no. 3. Stanford, Calif.: CSLI Publications.

Shah, R. 1989. Translation of Engineering Documentation with METAL. In *Machine Translation Summit*, ed. M. Nagao, 152–159. Tokyo. Omsha Ltd.

Shann, P. 1987. Machine Translation: A Problem of Linguistic Engineering or of Cognitive Modelling. In *Machine Translation Today: The State of the Art*, ed. M. King, 71–90. Edinburgh Information Technology 4. Edinburgh, Scotland. Edinburgh University Press.

Shieber, S. M. 1986. *An Introduction to Unification-Based Theories of Grammar*. Lecture Notes, no. 4. Stanford University: CSLI Publications.

Shieber, S. M. 1988. A Uniform Architecture for Parsing and Generation. In *Proceedings*, 614–619. Budapest, July-December. ACL. 12th International Conference.

Shieber, S. M., G. van Noord, R. Moore, and F. Pereira. 1989. A Semantic Head-Driven Generation Algorithm for Unification-Based Formalisms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 7–17. Association for Computational Linguistics.

Shikano, K. 1985. Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition. Technical report. Pittsburgh, Penn.: Carnegie Mellon University, May.

Shikano, K., K. Lee, and D. R. Reddy. 1986. Speaker Adaptation through Vector Quantization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Sigurd, B. 1988. Translating to and from Swedish by SWETRA—A Multilanguage Translation System. In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 205–218. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Silverman, K. E. A. 1987. *The Structure and Processing of Fundamental Frequency Contours.* Doctoral dissertation, University of Cambridge, Cambridge, England.

Sinaiko, H. W., and G. R. Klare. 1972. Further Experiments in Language Translation: Readability of Computer Translation. *Review for Applied Linguistics* 15:1–29.

Slocum, J. 1985a. A Machine(-Aided) Translation Bibliography. In *Machine Translation Systems*, ed. J. Slocum. Chap. 8, 265–341. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Slocum, J. 1985b. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. In *Machine Translation Systems*, ed. J. Slocum. Chap. 1, 1–47. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Slocum, J., W. S. Bennet, J. Bear, M. Morgan, and R. Root. 1987. METAL: The LRC Machine Translation System. In *Machine Translation Today: The State of the Art*, ed. S. Michaelson and Y. Wilks. Chap. 17, 319–350. Edinburgh University Press.

Snell, B. M. (ed.). 1979. *Translating and the Computer.* Amsterdam. North-Holland. Proceedings of a seminar organized by the Technical Translation and Informatics Groups of Aslib.

Stentiford, F., and M. G. Steer. 1987. A Speech Driven Language Translation System. Presented at the European Speech Technology Conference, September, 1987.

Stern, R. M., W. H. Ward, A. G. Hauptmann, and J. Leon. 1987. Sentence Parsing with Weak Grammatical Constraints. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 380–383. New York: IEEE.

Strzalkowski, T. (ed.). 1991. *Reversible Grammar in Natural Language Processing.* Berkeley, Calif. Association for Computational Linguistics.

Sumita, E., and H. Iida. 1991. Experiments and Prospects of Example-Based Machine Translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* 185–192. Berkeley, Calif. Association for Computational Linguistics.

Sumita, E., and Y. Tsutsumi. 1988. A Translation Aid System Using Flexible Text Retrieval Based on Snytax-Matching. TRL Research Report, IBMTR-87-1019. Tokyo Research Laboratory.

Tanaka, A., and S. Kamiya. 1986. A Speech Processing Based on a Syllable Identification by Using Phonological Patterns. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2231–2234. New York: IEEE.

Tanaka, H. 1989. Commercial Machine Translation. In *Machine Translation Summit,* ed. M. Nagao, 91–92. Tokyo. Omsha Ltd.

Tesnière, L. 1959. *Eléments de Syntaxe Structurale.* Paris: Klincksieck.

Texas Instruments Speech Recognition Group. 1987. TI Speech Recognition Technology Development, October. In *Proceedings of the DARPA Speech Recognition Workshop.*

Thagard, P. R. 1978. The Best Explanation: Criteria for Theory Choice. *The Journal of Philosophy* 75(2):76–92.

Thompson, H. S., and J. D. Laver. 1987. The Alvey Speech Demonstrator—Architecture, Methodology, and Progress to Date. In *Official Proceedings of Speech Tech '87,* ed. A. Donly. New York. Media Dimensions.

Tomabechi, H., H. Kitano, T. Mitamura, L. Levin, and M. Tomita. 1989. Direct Memory Access Speech-to-Speech Translation: A Theory of Simultaneous Interpretation. Technical Report CMU-CMT-89-111. Pittsburgh, Penn.: Carnegie Mellon University.

Tomita, M. 1984. The Design Philosophy of Personal Machine Translation System. Technical Report CMU-CS-84-142. Pittsburgh, Penn.: Carnegie Mellon University. Department of Computer Science.

Tomita, M. 1985. Feasibility Study of Personal/Interactive Machine Translation. Technical Report CMU-CS-85-140. Pittsburgh, Penn.: Carnegie Mellon University. Department of Computer Science.

Tomita, M. 1986. An Efficient Word Lattice Parsing Algorithm for Continuous Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1569–1572. New York: IEEE. volume 3.

Tomita, M., and J. G. Carbonell. 1987. The Universal Parser Architecture for Knowledge-Based Machine Translation. Technical Report CMU-CMT-87-101. Pittsburgh, Penn.: Carnegie Mellon University.

Tomita, M., and J. G. Carbonell. 1989. CMU Project. In *Machine Translation Summit*, ed. M. Nagao, 72–82. Tokyo. Omsha Ltd.

Tomita, M., and E. H. Nyberg. 1988. Generation Kit and Transformation Kit Version 3.2: User's Manual. Technical Memo CMU-CMT-88-MEMO. Pittsburgh, Penn.: Carnegie Mellon University.

Trost, H. 1990. The Application of Two Level Morphology to Nonconcatenative German Morphology. In *Proceedings of COLING-90*, 371–376.

Tsuji, Y. 1989. Machine Translation with Japan's Neighboring Countries. In *Machine Translation Summit*, ed. M. Nagao, 50–53. Tokyo. Omsha Ltd.

Tsujii, J. 1988. What is a Cross-Linguistically Valid Interpretation of Discourse? In *New Directions in Machine Translation*, ed. D. Maxwell, K. Schubert, and T. Witkam, 157–166. Distributed Language Translation 4. Dordrecht, Holland. Foris.

Tsujii, J. 1989. Mu Project. In *Machine Translation Summit*, ed. M. Nagao, 66–71. Tokyo. Omsha Ltd.

Tucker, A. B. 1987. Current Strategies in Machine Translation Research and Development. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 2, 22–41. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Turn, R. 1974. The Use of Speech for Man-Computer Communication. Technical Report 1386-ARPA. RAND Corporation.

Uchida, H. 1986. Fujitsu Machine Translation System: ATLAS. *Future Generations Computer Systems* 2(2):95–100.

Uchida, H. 1989. ATLAS II: A Machine Translation System Using Conceptual Structure as an Interlingua. In *Machine Translation Summit*, ed. M. Nagao, 93–100. Tokyo. Omsha Ltd.

Uchida, H., and T. Kakizaki. 1989. Electronic Dictionary Project. In *Machine Translation Summit*, ed. M. Nagao, 83–87. Tokyo. Omsha Ltd.

Uszkoreit, H. 1986. Categorical Unification Grammars. In *Proceedings of COLING-86*, 187–194. Bonn. Institut für angewandte Kommunikations- und Sprachforschung e.V.

Uszkoreit, H. 1991. Strategies for Adding Control Information to Declarative Grammars. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 237–245. Berkeley, Calif. Association for Computational Linguistics.

Van Noord, G. 1990a. Reversible Unification-Based Machine Translation. In *Proceedings of COLING-90*. Helsinki.

Van Noord, G. 1990b. *User Manual,* PATRA *18, Software Environment of the MiMo2 Translation System.* Utrecht, Holland: OTS RUU.

Vasconcellos, M. 1985. Management of the Machine Translation Environment: Interaction of Functions at the Pan American Health Organization. In *Tools for the Trade, Translating and the Computer 5: Proceedings of a Conference...,* ed. V. Lawson. 115–129. London: Aslib.

Vasconcellos, M., and M. León. 1985a. SPANAM and ENGSAM: Machine Translation at the Pan American Health Organization. In *Machine Translation Systems,* ed. J. Slocum. Chap. 6, 187–236. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Vasconcellos, M., and M. León. 1985b. SPANAM and ENGSAM: Machine Translation at the Pan American Health Organization. *Computational Linguistics* 11(2–3):122–136.

Vauquois, B. 1975. *La Traduction Automatique à Grenoble.* Documents de Linguistique Quantitative, Vol. 24. Paris: Dunod.

Vauquois, B., and C. Boitet. 1985. Automated Translation at Grenoble University. In *Machine Translation Systems,* ed. J. Slocum. Chap. 3, 85–110. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Viterbi, A. J. 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory* IT-13(2):260–269.

Waibel, A. 1986. Recognition of Lexical Stress in a Continuous Speech Understanding System—A Pattern Recognition Approach. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2287–2290. New York: IEEE.

Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. 1988. Phoneme Recognition: Neural Networks vs. Hidden Markov Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 107–110. New York: IEEE. volume 1.

Waibel, A., and K.-F. Lee (ed.). 1990. *Readings in Speech Recognition.* San Mateo, Calif.: Morgan Kaufmann.

Waibel, A., H. Sawai, and K. Shikano. 1989. Consonant and Phoneme Recognition by Modular Constructions of Large Phonemic Time-Delay Neural Networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.* Reprinted in Waibel and Lee 1990.

Waibel, A. H. 1986. *Prosody and Speech Recognition.* Doctoral dissertation, Pittsburgh, Penn.: Carnegie Mellon University.

Walker, D. D. 1987. Knowledge Resource Tools for Accessing Large Text Files. In *Machine Translation: Theoretical and Methodological Issues,* ed. S. Nirenburg. Chap. 14, 247–261. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Ward, N. 1991. A Flexible, Parallel Model of Natural Language Generation. Technical Report UCB/CSD 91/629. Berkeley, Calif.: University of California, Berkeley, April.

Ward, W. H., A. G. Hauptmann, R. M. Stern, and T. Chanak. 1988. Parsing Spoken Phrases Despite Missing Words. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 275–278. New York: IEEE. volume 1.

Warwick, S. 1987. An Overview of Post-ALPAC Developments. In *Machine Translation Today: The State of the Art*, ed. M. King, 22–37. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Weaver, W. 1955. Translation. In *Machine Translation of Languages*, ed. W. N. Locke and A. D. Booth. 15–23. Cambridge, Mass.: MIT Press.

Wehrli, E. 1987. Recent Developments in Theoretical Linguistics and Implications for Machine Translation. In *Machine Translation Today: The State of the Art*, ed. M. King, 58–70. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

Weischedel, R. M. 1989. A Hybrid Approach to Representation in the Janus Natural Language Processor. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 193–202. Association for Computational Linguistics.

Weischedel, R. M., and L. A. Ramshaw. 1987. Reflections on the Knowledge Needed to Process Ill-Formed Language. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 10, 155–167. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Wheeler, P. 1987. Systran. In *Machine Translation Today: The State of the Art*, ed. M. King, 192–208. Edinburgh Information Technology Series. Edinburgh, Scotland. Edinburgh University Press.

White, J. S. 1986. What Should Machine Translation Be? In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 267. Association for Computational Linguistics.

White, J. S. 1987. The Research Environment in the METAL Project. In *Machine Translation: Theoretical and Methodological Issues*, ed. S. Nirenburg. Chap. 13, 225–246. Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

Wilks, Y. 1973. An Artificial Intelligence Approach to Machine Translation. In *Computer Models of Thought and Language*, ed. R. Schank and K. Colby. 114–151. San Francisco: W. H. Freeman.

Wilks, Y. 1975. A Preferential Pattern-Seeking Semantics for Natural Language Inference. *Artificial Intelligence* 6:53–74.

Wilks, Y. 1976a. Parsing English II. In *Computational Semantics*, ed. E. Charniak and Y. Wilks. 155–184. Amsterdam: North-Holland.

Wilks, Y. 1976b. Semantics and World Knowledge in Machine Translation. American Journal for Computational Linguistics 2, microfiche 48, 67-69.

Wilks, Y. 1979. Machine Translation and Artificial Intelligence. In *Translating and the Computer*, ed. B. M. Snell. 27–43. Amsterdam: North-Holland.

Wilks, Y., and D. Farwell. 1990. A White Paper on Research in Pragmatics-based Machine Translation. Memoranda in Computer and Cognitive Science MCCS-90-188. New Mexico State University. Computer Research Laboratory.

Wilpon, J. G., L. R. Rabiner, and A. Bergh. 1982. Speaker-Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary. *The Journal of the Acoustical Society of America* 2(72):390–396.

Wu, H. Y., P. Badwin, Y. M. Cheng, and B. Guerin. 1987. Vocal Tract Simulation: Implementation of Continuous Variations of the Length in a Kelly-Lochbaum Model, Effects of Area Function Spatial Sampling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 0009–0012. New York: IEEE. volume 1.

Yamaguchi, Y., and T. Matsumoto. 1990. A Neural Network Approach to Multi-Language Text-to-Speech System. In *Proceedings of the International Conference on Spoken Language Processing*, 325–328. Kobe, Japan.

Yannakoudakis, E. J., and P. J. Hutton. 1987. *Speech Synthesis and Recognition Systems.* Chichester, England: Ellis Horwood.

Young, S. J., N. H. Russell, and J. H. S. Thornton. 1988. Speech Recognition in VODIS II. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 441–444. New York: IEEE. volume 1.

Zaharin, Y. 1987. Towards an Analyzer (parser) in a Machine Translation System Based on Ideas from Expert Systems. *Computational Intelligence* 4(2):180–191.

Zajac, R. 1986. SCSL: A Linguistic Specification Language for Machine Translation. In *Proceedings of COLING-86.* Bonn.

Zajac, R. 1989. A Transfer Model Using a Typed Feature Structure Rewriting System with Inheritance. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 1–6. Association for Computational Linguistics.

Zajac, R. 1990. A Relational Approach to Translation. In *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language.* University of Texas as Austin.

Zajac, R. 1991. A Uniform Architecture for Parsing, Generation, and Transfer. In *Reversible Grammar in Natural Language Processing*, ed. T. Strzalkowski, 71–80. Association for Computational Linguistics. ACL SIG Workshop.

Zeevat, H., E. Klein, and J. Calder. 1987. An Introduction to Unification Categorial Grammar. In *Edinburgh Working Papers in Cognitive Science: Categorial Grammar, Unification Grammar and Parsing*, ed. J. N. Haddock, E. Klein, and G. Morill. Edinburgh: Edinburgh University Press. volume 1.

Zock, M., and G. Sabah (ed.). 1988. *Advances in Natural Language Generation: An Interdisciplinary Perspective.* Communication in Artificial Intelligence, Vol. 1. Norwood, New Jersey: Ablex Publishing.

Zue, V. W. 1985. The Use of Speech Knowledge in Automatic Speech Recogition. In *Proceedings of the* IEEE, 1602–1615. 11(73).

Zue, V. W., and R. Cole. 1979. Experiments on Spectrogram Reading. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* 116–119.

CSLI Publications

Lecture Notes

The titles in this series are distributed in the United States and Canada by the University of Chicago Press and may be purchased in academic or university bookstores. They may be ordered directly from the distributor at 11030 South Langley Avenue, Chicago, IL 60628 (USA) or by phone 1-800-621-2736, (312) 568-1550. You may also order by fax at 1-800-621-8471 or (312) 660-2235 or Telex, 28-0206 (answerback UCPRESS CGO).

A Manual of Intensional Logic. van Benthem, 2nd edition. No. 1. 0-937073-29-6 (paper), 0-937073-30-X (cloth)

Emotion and Focus. Nissenbaum. No. 2. 0-937073-20-2 (paper)

Lectures on Contemporary Syntactic Theories. Sells. No. 3. 0-937073-14-8 (paper), 0-937073-13-X (cloth)

An Introduction to Unification-Based Approaches to Grammar. Shieber. No. 4. 0-937073-00-8 (paper), 0-937073-01-6 (cloth)

The Semantics of Destructive Lisp. Mason. No. 5. 0-937073-06-7 (paper), 0-937073-05-9 (cloth)

An Essay on Facts. Olson. No. 6. 0-937073-08-3 (paper), 0-937073-05-9 (cloth)

Logics of Time and Computation. Goldblatt, 2nd edition. No. 7. 0-937073-94-6 (paper), 0-937073-93-8 (cloth)

Word Order and Constituent Structure in German. Uszkoreit. No. 8. 0-937073-10-5 (paper), 0-937073-09-1 (cloth)

Color and Color Perception: A Study in Anthropocentric Realism. Hilbert. No. 9. 0-937073-16-4 (paper), 0-937073-15-6 (cloth)

Prolog and Natural-Language Analysis. Pereira and Shieber. No. 10. 0-937073-18-0 (paper), 0-937073-17-2 (cloth)

Working Papers in Grammatical Theory and Discourse Structure: Interactions of Morphology, Syntax, and Discourse. Iida, Wechsler, and Zec (Eds.). No. 11. 0-937073-04-0 (paper), 0-937073-25-3 (cloth)

Natural Language Processing in the 1980s: A Bibliography. Gazdar, Franz, Osborne, and Evans. No. 12. 0-937073-28-8 (paper), 0-937073-26-1 (cloth)

Information-Based Syntax and Semantics. Pollard and Sag. No. 13. 0-937073-24-5 (paper), 0-937073-23-7 (cloth)

Non-Well-Founded Sets. Aczel. No. 14. 0-937073-22-9 (paper), 0-937073-21-0 (cloth)

Partiality, Truth and Persistence. Langholm. No. 15. 0-937073-34-2 (paper), 0-937073-35-0 (cloth)

Attribute-Value Logic and the Theory of Grammar. Johnson. No. 16. 0-937073-36-9 (paper), 0-937073-37-7 (cloth)

The Situation in Logic. Barwise. No. 17. 0-937073-32-6 (paper), 0-937073-33-4 (cloth)

The Linguistics of Punctuation. Nunberg. No. 18. 0-937073-46-6 (paper), 0-937073-47-4 (cloth)

Anaphora and Quantification in Situation Semantics. Gawron and Peters. No. 19. 0-937073-48-4 (paper), 0-937073-49-0 (cloth)

Propositional Attitudes: The Role of Content in Logic, Language, and Mind. Anderson and Owens. No. 20. 0-937073-50-4 (paper), 0-937073-51-2 (cloth)

Literature and Cognition. Hobbs. No. 21. 0-937073-52-0 (paper), 0-937073-53-9 (cloth)

Situation Theory and Its Applications, Vol. 1. Cooper, Mukai, and Perry (Eds.). No. 22. 0-937073-54-7 (paper), 0-937073-55-5 (cloth)

The Language of First-Order Logic (including the Macintosh program, Tarski's World 4.0). Barwise and Etchemendy, 3rd Edition. No. 23. 0-937073-99-7 (paper)

Lexical Matters. Sag and Szabolcsi (Eds.). No. 24. 0-937073-66-0 (paper), 0-937073-65-2 (cloth)

Tarski's World: Macintosh Version 4.0. Barwise and Etchemendy. No. 25. 1-881526-27-5 (paper)

Situation Theory and Its Applications, Vol. 2. Barwise, Gawron, Plotkin, and Tutiya (Eds.). No. 26. 0-937073-70-9 (paper), 0-937073-71-7 (cloth)

Literate Programming. Knuth. No. 27. 0-937073-80-6 (paper), 0-937073-81-4 (cloth)

Normalization, Cut-Elimination and the Theory of Proofs. Ungar. No. 28. 0-937073-82-2 (paper), 0-937073-83-0 (cloth)

Lectures on Linear Logic. Troelstra. No. 29. 0-937073-77-6 (paper), 0-937073-78-4 (cloth)

A Short Introduction to Modal Logic. Mints. No. 30. 0-937073-75-X (paper), 0-937073-76-8 (cloth)

Linguistic Individuals. Ojeda. No. 31. 0-937073-84-9 (paper), 0-937073-85-7 (cloth)

Computational Models of American Speech. Withgott and Chen. No. 32. 0-937073-98-9 (paper), 0-937073-97-0 (cloth)

Verbmobil: A Translation System for Face-to-Face Dialog. Kay, Gawron, and Norvig. No. 33. 0-937073-95-4 (paper), 0-937073-96-2 (cloth)

The Language of First-Order Logic (including the Windows program, Tarski's World 4.0). Barwise and Etchemendy, 3rd edition. No. 34. 0-937073-90-3 (paper)

Turing's World. Barwise and Etchemendy. No. 35. 1-881526-10-0 (paper)

The Syntax of Anaphoric Binding. Dalrymple. No. 36. 1-881526-06-2 (paper), 1-881526-07-0 (cloth)

Situation Theory and Its Applications, Vol. 3. Aczel, Israel, Katagiri, and Peters (Eds.). No. 37. 1-881526-08-9 (paper), 1-881526-09-7 (cloth)

Theoretical Aspects of Bantu Grammar. Mchombo (Ed.). No. 38. 0-937073-72-5 (paper), 0-937073-73-3 (cloth)

Logic and Representation. Moore. No. 39. 1-881526-15-1 (paper), 1-881526-16-X (cloth)

Meanings of Words and Contextual Determination of Interpretation. Kay. No. 40. 1-881526-17-8 (paper), 1-881526-18-6 (cloth)

Language and Learning for Robots. Crangle and Suppes. No. 41. 1-881526-19-4 (paper), 1-881526-20-8 (cloth)

Hyperproof. Barwise and Etchemendy. No. 42. 1-881526-11-9 (paper)

Mathematics of Modality. Goldblatt. No. 43. 1-881526-23-2 (paper), 1-881526-24-0 (cloth)

Feature Logics, Infinitary Descriptions, and Grammar. Keller. No. 44. 1-881526-25-9 (paper), 1-881526-26-7 (cloth)

Tarski's World: Windows Version 4.0. Barwise and Etchemendy. No. 45. 1-881526-28-3 (paper)

German in Head-Driven Phrase Structure Grammar. Pollard, Nerbonne, and Netter. No. 46. 1-881526-29-1 (paper), 1-881526-30-5 (cloth)

Other CSLI Titles Distributed by UCP

Agreement in Natural Language: Approaches, Theories, Descriptions. Barlow and Ferguson (Eds.). 0-937073-02-4 (cloth)

Papers from the Second International Workshop on Japanese Syntax. Poser (Ed.). 0-937073-38-5 (paper), 0-937073-39-3 (cloth)

The Proceedings of the Seventh West Coast Conference on Formal Linguistics (WCCFL 7). 0-937073-40-7 (paper)

The Proceedings of the Eighth West Coast Conference on Formal Linguistics (WCCFL 8). 0-937073-45-8 (paper)

The Phonology-Syntax Connection. Inkelas and Zec. 0-226-38100-5 (paper), 0-226-38101-3 (cloth)

The Proceedings of the Ninth West Coast Conference on Formal Linguistics (WCCFL 9). 0-937073-64-4 (paper)

Japanese/Korean Linguistics. Hoji (Ed.). 0-937073-57-1 (paper), 0-937073-56-3 (cloth)

Experiencer Subjects in South Asian Languages. Verma and Mohanan (Eds.). 0-937073-60-1 (paper), 0-937073-61-X (cloth)

Grammatical Relations: A Cross-Theoretical Perspective. Dziwirek, Farrell, Bikandi (Eds.). 0-937073-63-6 (paper), 0-937073-62-8 (cloth)

The Proceedings of the Tenth West Coast Conference on Formal Linguistics (WCCFL 10). 0-937073-79-2 (paper)

On What We Know We Don't Know. Bromberger. 0-226-075400 (paper), (cloth)

The Proceedings of the Twenty-fourth Annual Child Language Research Forum. Clark (Ed.). 1-881526-05-4 (paper), 1-881526-04-6 (cloth)

Japanese/Korean Linguistics, Vol. 2. Clancy (Ed.). 1-881526-13-5 (paper), 1-881526-14-3 (cloth)

Arenas of Language Use. Clark. 0-226-10782-5 (paper), (cloth)

Japanese/Korean Linguistics, Vol. 3. Choi (Ed.). 1-881526-21-6 (paper), 1-881526-22-4 (cloth)

The Proceedings of the Eleventh West Coast Conference on Formal Linguistics (WCCFL 11). 1-881526-12-7 (paper)

Phrase Structure and Grammatical Relations in Tagalog. Kroeger. 0-937073-86-5 (paper), 0-937073-87-3 (cloth)

Theoretical Aspects of Kashaya Phonology and Morphology. Buckley. 1-881526-02-X (paper), 1-881526-03-8 (cloth)

Books Distributed by CSLI

The Proceedings of the Fourth West Coast Conference on Formal Linguistics (WCCFL 4). 0-937073-43-1 (paper)

The Proceedings of the Fifth West Coast Conference on Formal Linguistics (WCCFL 5). 0-937073-42-3 (paper)

The Proceedings of the Sixth West Coast Conference on Formal Linguistics (WCCFL 6). 0-937073-31-8 (paper)

Hausar Yau Da Kullum: Intermediate and Advanced Lessons in Hausa Language and Culture. Leben, Zaria, Maikafi, and Yalwa. 0-937073-68-7 (paper)

Hausar Yau Da Kullum Workbook. Leben, Zaria, Maikafi, and Yalwa. 0-93703-69-5 (paper)

Ordering Titles Distributed by CSLI

Titles distributed by CSLI may be ordered directly from CSLI Publications, Ventura Hall, Stanford, CA 94305-4115 or by phone (415)723-1712, (415)723-1839. Orders can also be placed by FAX (415)723-0758 or e-mail (pubs@csli.stanford.edu).

All orders must be prepaid by check or Visa or MasterCard (include card name, number, and expiration date). California residents add 8.25% sales tax. For shipping and handling, add $2.50 for first book and $0.75 for each additional book; $1.75 for first report and $0.25 for each additional report.

For overseas shipping, add $4.50 for first book and $2.25 for each additional book; $2.25 for first report and $0.75 for each additional report. All payments must be made in U.S. currency.

Overseas Orders

The University of Chicago Press has offices worldwide which serve the international community.

Mexico, Central America, South America, and the Caribbean (including Puerto Rico): EDIREP, 5500 Ridge Oak Drive, Austin, Texas 78731 U. S. A. Telephone: (512) 451-4464. Facsimile: (512) 451-4464.

United Kingdom and Europe: (VAT is added where applicable.) International Book Distributors, Ltd., Campus 400, Maylands Avenue, Hemel Hempstead HP2 7EZ, England. Telephone: 0442 881900/Telex: 82445. Facsimile: 0442 882099. Internet: 536-2875@MCIMAIL.COM

Australia, New Zealand, South Pacific, Africa, Middle East, China (PRC), Southeast Asia, and India: The University of Chicago Press, International Sales Manager, 5801 South Ellis Avenue, Chicago, Illinois 60637 U.S.A. Telephone: (312)702-7706. Facsimile: (312)702-9756. Internet: dblobaum@press.uchicago.edu

Japan: Libraries and individuals should place their orders with local booksellers. Booksellers should place orders with our agent: United Publishers Services, Ltd.,

Kenkyu-sha Building, 9 Kanda Surugadai 2-chome, Chiyoda-ku, Tokyo, Japan. Telephone: (03)3291-4541. Facsimile: (03)3293-8610. Telex: J33331 (answerback UPSTOKYO). Cable: UNITEDBOOKS TOKYO.

Korea, Hong Kong, and Taiwan, R.O.C.: The America University Press Group, 3-21-18-206 Higashi-Shinagawa, Shinagawa-ku, Tokyo 140, Japan. Telephone: (03)3450-2857. Facsimile: (03)3472-9706.

Internet Gopher Access: University of Chicago Press catalogs can be searched on-line by connecting to the University of Chicago Press gopher:

press-gopher.uchicago.edu