Nine Issues in Speech Translation

Mark Seligman
Université Joseph Fourier, GETA, CLIPS, IMAG-campus, BP 53
385, rue de la Bibliothèque, 38041 Grenoble Cedex 9, France
and
Spoken Translation, Inc.
1100 West View Drive, Berkeley, CA 94705
mark.seligman@spokentranslation.com

Abstract

This paper sketches the author's research in nine areas related to speech translation: interactive disambiguation (two demonstrations of highly-interactive, broad-coverage speech translation are reported); system architecture; datastructures; the interface between speech recognition and analysis; the use of natural pauses for segmenting utterances; example-based machine translation; dialogue acts; the tracking of lexical co-occurrences; and the resolution of translation mismatches.

Introduction

This paper reviews some aspects of the author's research in speech translation (ST) since 1992. Since the purpose is to prompt discussion, the treatment is informal, programmatic, and speculative. There is frequent reference to work in progress—in other words, work for which evaluation is incomplete.

The paper sketches work in nine areas: interactive disambiguation; system architecture; datastructures; the interface between speech recognition and analysis; the use of natural pauses for segmenting utterances; example-based machine translation; dialogue acts; the tracking of lexical co-occurrences; and the resolution of translation mismatches. There is no attempt to provide a balanced survey of the speech translation scene. Instead, the hope is to provide a provocative and somewhat personal look at the field by spotlighting it from nine directions—in some respects, to offer an editorial rather than purely a report.

One of the most significant and difficult aspects of the speech translation problem is the need to effectively integrate many different sorts of knowledge: phonological, prosodic, morphological, syntactic, semantic, discourse, and domain knowledge should ideally work together to produce the most accurate and helpful translation. Thus a trend toward greater integration of knowledge sources is visible in much current speech translation research (e.g. in the VERBMOBIL project (Wahlster 1993), and most of the work described below is in this integrative direction. Many of the issues to be discussed here could in fact be addressed by dedicated pieces of software playing parts in an integrated ST system. The paper's conclusion will review the issues by sketching an idealized system of this sort—a kind of personal dream team in which the components are team members.

However, the first topic to be discussed is a renegade, headed in exactly the opposite direction. This is because, while continuing my concern with integration of speech translation system components, I have become interested in an alternative system design which, in sharp contrast, stresses a clean separation between speech recognition and translation. The thrust of this alternative "low road" or "quick and dirty" approach is to temporarily substitute intensive user interaction for system integration, thereby attempting a radical design simplification in hopes of fielding practical, broad-coverage systems as soon as possible.

To accommodate this renegade on the one hand and the team players on the other, the paper will be not only somewhat personal, but also two-faced. I will begin by advocating a "low road", non-integrated approach for the near term throughout Section 1. Two demonstrations of highly-interactive, broad-coverage speech translation will be reported and discussed. Then, performing an about-face, I will go on in the remaining sections to consider elements of a

Seligman, Mark. 2000. "Nine Issues in Speech Translation." Machine Translation, Vol. 15, Issue 1/2, June 2000, pages 149-186. Special Issue on Spoken Language Translation.

more satisfying integrated approach for the longer term.

1 Interactive Disambiguation

At the present state of the art, several stages of speech translation leave ambiguities which current techniques cannot yet resolve correctly and automatically. Such residual ambiguity plagues speech recognition, analysis, transfer, and generation alike.

Since users generally can resolve these ambiguities, it seems reasonable to incorporate facilities for interactive disambiguation into speech translation systems, especially those aiming for broad coverage. A good idea of the range of work in this area can be gained by browsing (Boitet 1996a).

In fact, (Seligman 1997) suggests that, by stressing such interactive disambiguation—for instance, by using highly-interactive commercial dictation systems for input, and by adapting existing techniques for interactive disambiguation of text translation (Boitet 1996b; Blanchon 1996)—practically usable speech translation systems may be constructable in the near term. In such "quick and dirty" or "low road" speech translation systems, user interaction is substituted for system integration. For example, the interface between speech recognition and analysis can be supplied entirely by the user, who can correct speech recognition (SR) results before passing them to translation components, thus bypassing any attempt at effective communication or feedback between SR and MT.

The argument, however, is not that the "high road" toward integrated and maximally automatic systems should be abandoned. Rather, it is that the "low road" of forgoing integration and embracing interaction may offer the quickest route to widespread usability, and that experience with real use is vital for progress. Clearly, the "high road" is the most desirable for the longer term: integration of knowledge sources is a fundamental issue for both cognitive and computer science, and maximally automatic use is intrinsically desirable. The suggestion, then, is that the low and high roads be traveled in tandem; and that even systems aiming for full automaticity recognize the need for interactive resolution when automatic resolution is insufficient. As progress is made along the "high road" and increasing knowledge can be applied to automatic ambiguity resolution, interactive resolution should be necessary less often. When it is necessary, its quality should be improved: questions put to the user should become more sensible and more tightly focused.

1.1 Two Interactive Demos

These design concepts have been informally and partly tested in two demos, first at the Machine Translation Summit in San Diego in October, 1997, and a second time at the meeting of C-STAR II (Consortium for Speech Translation Advanced Research) in Grenoble, France, in January, 1998. Both demos were organized and conducted under the supervision of Mary Flanagan, and both demo systems were based upon a text-based chat translation system previously built by Flanagan's team at CompuServe, Inc. The company's proprietary online chat technology was used, as distinct from Internet Relay Chat, or IRC (Pyra 1995).[1]

In an online chat session, users most often converse as a group, though one-on-one conversations are also easy to arrange. Each conversant has a small window used for typing input. Once the input text is finished, the user sends it to the chat server by pressing Return. The text comes back to the sender after an imperceptible interval, and appears in a larger window, prefaced by a header indicating the author. Since this larger window receives input from all parties to the chat conversation, it soon comes to resemble the transcript of a cocktail party, often with several conversations interleaved.

Each party normally sees the "same" transcript window. However, prior to the speech translation demos, CompuServe had arranged to place at the chat server a commercial translation system of the direct variety, enabling several translation directions. Once the user of this experimental chat system had selected a direction (say English-French), all lines in the transcript window would appear in the source language (in this case, English), even if some of the contributions originated in the target language (here, French). Bilingual text conversations were thus enabled between English typists and writers of French, German, Spanish, or Italian.

At the time of the demos, total delay from the pressing of Return until the arrival of translated text in the interlocutor's transcript window averaged about six seconds, well within tolerable limits for conversa-

1. My thanks to CompuServe, Inc., Mary Flanagan, and her staff are expressed in the Acknowledgements section. The opinions offered throughout this paper, however, are my own. CompuServe's chat translation project was discontinued in early 1998.

tion. (At the time of writing, another commercial chat translation service (*www.uni-verse.com*)—the only such service, now that the CompuServe project has been discontinued—typically gives a comparable throughput in 2-3 seconds.)

At the author's suggestion and with his consultation, highly-interactive speech translation demos were created by adding speech recognition front ends and speech synthesis back ends to CompuServe's text-based chat translation system. Two laptops were used, one running English input and output software (in addition to the CompuServe client, modified as explained below), and one running the comparable French programs.

Commercial dictation software was employed for speech recognition. For the first demo, both sides used discrete dictation, in which short pauses are required between words; for the second demo, English dictated continuously—that is, without required pauses—while French continued to dictate discreetly. (Continuous French was released just before the second demo, but because little testing time was available, a decision was made to forego its use.)

At the time of the demos, the discrete products allowed dictation directly into the chat input buffer, but the continuous products required dictation into their own dedicated window. Thus for continuous English input it became necessary to employ third-party software[2] to create a macro which (1) transferred dictated text to the chat input buffer and (2) inserted a Return as a signal to send the chat. (By March 1998, upgrades of the continuous software had already made this macro less necessary. Direct dictation to the chat window would then have been possible without it, with some sacrifice of advanced features for voice-driven interactive correction of errors.)

Commercial speech synthesis programs packaged with the discrete dictation products were used for voice synthesis. Using development software sold separately by the dictation vendor, CompuServe's chat client software was customized so that, as each text string returning from the chat server was written to the transcript window, it was simultaneously sent to the speech synthesis engine to be pronounced in the appropriate language. The text read aloud in this way was either the user's own, transmitted without changes, or the translation of an interlocutor's input.

2. SpeechLinks software from SpeechOne, Inc.

The first demo took place in an auditorium before a quiet audience of perhaps one hundred, while the second was presented to numerous small groups in a booth in a noisy room of medium size. Each demo began with ten scripted and pre-tested utterances, and then continued with improvised utterances, sometimes solicited from the audience—perhaps six in the first demo, and fifty or more in the second. Some examples of improvised sentences:

FRENCH: Qu'est-ce que vous étudiez? (What do you study?)

ENGLISH: Computer science. (L'informatique.)

FRENCH: Qu'est-ce que vous faites plus tard? (What are you doing later?)

ENGLISH: I'm going skiing. (Je vais faire du ski.)

FRENCH: Vous n'avez pas besoin de travailler? (You don't need to work?)

ENGLISH: I'll take my computer with me. (Je prendrai mon ordinateur avec moi.)

FRENCH: Où est-ce que vous mettrez l'ordinateur pendant que vous skiez? (Where will you put the computer while you ski?)

ENGLISH: In my pocket. (Dans ma poche.)

As these examples suggest, the level of language remained basic, and sentences were purposely kept short, with standard grammar and punctuation.

1.2 Discussion of Demos

A primary purpose of the chat speech translation demos was to show that speech translation is both feasible and suitable for online chat users, at least at the proof-of-concept level.

In my own view, the demos were successful in this respect. The basic feasibility of the approach appears in the fact that most demo utterances were translated comprehensibly and within tolerable time limits. It is true that the language, while mostly spontaneous, was consciously kept quite basic and standard. It is also true that there were occasional translation errors (discussed below). Nevertheless, the demos can plausibly be claimed to show that chatters making a reasonable effort could successfully socialize in this way. As preliminary evidence that many users could adjust to the system's limitations, we can remark that the dozen or so utterances suggested by the audience, once repeated verbatim by the demonstrators, were successfully recognized, translated, and pronounced in every case.

In addition to the general demo goals just mentioned, the author also had his own, more specific axes to grind from the viewpoint of speech translation research. I hoped the demos would be the first to show broad-coverage speech translation of usable quality; and I hoped they would highlight the potential usefulness of interactive disambiguation in moving toward practical broad-coverage systems.[3]

I believe that these goals, too, were reached. Coverage was indeed broad by contemporary standards. There was no restriction on conversational topic—no need, for instance, to remain within the area of airline reservations, appointment scheduling, or street directions. As long as the speakers stayed within the dictation and translation lexica (each in the tens of thousands of words), they were free to chat and banter as they liked.

The usefulness of interaction in achieving this breadth was also clear: verbal corrections of dictation results were indeed necessary for perhaps 5-10% of the input words. To give only the most annoying example, "Hello" was once initially transcribed as "Hollow." (Here we see with painful clarity the limitations of an approach which substitutes interactive disambiguation for automatic knowledge-based disambiguation: even the most rudimentary discourse knowledge should have allowed the program to judge which word was more likely as a dialog opener. On the other hand, the approach's capacity to compensate for lack of such knowledge was also clear: a verbal correction was quickly made, using facilities supplied by the dictation vendor.)

It should be stressed that the speech translation system of the CompuServe demos was by no means the first or only system to permit interactive monitoring of speech recognition output before translation. As far back as the C-STAR I international speech translation demonstrations of 1993 (www.itl.atr.co.jp/matrix/c-star/index .en.html), selection among SR candidates was essential for most participating systems. Similarly, selection among, or typed correction of, candidates is possible in most of the systems shown in the recent C-STAR II demos of July 22, 1999 (*www.c-star.org*).

The CompuServe experiments were, however, the first to demonstrate that a broad-coverage speech translation system of usable quality—a system capable of extending coverage beyond specialized domains toward unrestricted discourse—could be constructed by enabling users to ergonomically correct the output of a broad-coverage speech recognition component before passing the results to a broad-coverage machine translation component.

Ergonomic operation was an important element in the system's success. The SR correction facilities used in the experiments—the set of verbal revision commands supplied by the dictation product, including "scratch that", "correct < word > ", etc.—were designed for general use in a competitive market, and thus of necessity show considerable attention to ergonomic issues. (By contrast, the SR components of other research systems continue to rely on typed correction or menu selection.) Of course, a smooth human interface between SR and MT cannot by itself yield broad coverage; what it can do is to permit the unexpected combination of SR and MT components developed separately, with broad coverage rather than speech translation in mind.

This reliance on interactive correction raises obvious questions: Is the current amount and type of dictation correction tolerable for practical use? Would additional interaction for guiding or correcting translation be useful? Even if potentially useful, would it be tolerated, or would it break the camel's back?

Correction of dictation The interaction required in the current demos for correcting dictation is just that currently required for correcting text dictation in general. All current dictation products require interactive correction. The question is, do the advantages of dictation over typing nevertheless justify the cost of these products, plus the trouble of acquiring them, training them, and learning to use them? Their steadily increasing user base indicates that many users think so. (For the record, portions of this paper were produced using continuous dictation software.) My own impression is that, during the demos, continuous dictation with spoken corrections supplied correct text at least twice as fast as my own reasonably skilled typing would have done.

3. (Kowalski et al 1995) arranged the only previous demonstration known to the author of speech translation using dictated input. Since users (spectators at twin exposition displays in Boston, Massachusetts and Lyons, France) were untrained, little interactive correction of dictation was possible. For this and other reasons, translation quality was generally low (Burton Rosenberg, personal communication); but as the main purpose of the demo was to make an artistic and social statement concerning future hi-tech possibilities for cross-cultural communication, this was no great cause for concern. Text was transmitted via FTP, rather than via chat as in the experiments reported here. See (Seligman 1997) for a fuller account.

For readers who have never tried dictating, a description of the dictation correction process available in (Seligman et al 1998b) may help to realistically estimate the correction burden.

While a strict hands-off policy was adopted for the demos, it is worth noting that typed text and commands can be freely interspersed with spoken text and commands. It is sometimes handy, for instance, to select an error using the mouse, and then to verbally apply any of the above-mentioned correction commands. Similarly, when spelling becomes necessary, typing often turns out to be faster than spoken spelling. Thus verbal input becomes one option among several, to be chosen when—as often happens—it offers the easiest or fastest path to the desired text. The question, then, is no longer whether to type or dictate the discourse as a whole, but which mode is most convenient for the input task immediately at hand. As broad-coverage speech translation systems in the near term are likely to remain multimodal rather than exclusively telephonic, they can and should take advantage of this flexibility.

Correction of translation The current demos were not intended to demonstrate the full range of interactive possibilities. In particular, while *dictation* results were corrected online as just discussed, there was no comparable attempt at interactive disambiguation of *translation*. Thus, when ambiguities occurred, the speaker had no way to control or check the translation results.

For example, when the English partner concluded one dialog by saying, *It was a pleasure working with you*, the French partner saw and heard *C'était un plaisir fonctionner avec vous*—literally, "It has been a pleasure functioning with you." *Work*, in other words, had been translated as *fonctionner*, as would be appropriate for an input like *This program is not working*.

Such translation errors were not disruptive during the demos: they were infrequent, and many of the errors which did appear might be more amusing than bothersome in the sort of informal socializing seen in most online chat today.

However, errors arising from lexical and structural ambiguities might well be more numerous and more disruptive in future, more sensitive chat translation applications. Further, it seems doubtful that they can be eliminated in near-term systems aiming for both broad coverage and high-quality, even assuming effective use of multiple knowledge sources like those described below. Thus my own guess is that interactive resolution of ambiguities during chat translation would in fact prove valuable. Feedback concerning the translation, via some form of back-translation, would probably prove useful as well. Again, for discussion of possible techniques, see (Boitet 1996a and Blanchon 1996).

But even granting that interactive correction could raise the quality of speech translation, would users be willing to supply it? There is some indication that the degree of interaction now required in the demos to correct dictation may already be near the tolerable limit for chat as it is presently used (Flanagan 1997). A healthy skepticism concerning the practicality of real-time translation correction is thus warranted. I suspect, however, that users' toleration for interactive correction will turn out to depend on the application and the value of correct translation: to the extent that real-time machine translation can move beyond socializing into business, emergency, military, or other relatively crucial and sensitive applications, user tolerance for interaction can be expected to increase.

Ultimately, though, questions about the tradeoff between the burden of interaction and its worth should be treated as topics for research: using a specified system, what level of quality is required for given applications (specified in terms of tasks to be accomplished within specified time limits), and what types and amounts of interaction are required on average to achieve that quality level? Clearly, until speech translation systems with translation correction capabilities are built, no such experiments will be possible.

Having discussed the role of interactive disambiguation in ST, and having described two experiments with highly-interactive ST, we now turn on our heels as forecast toward more integration-oriented studies. We begin with considerations of ST system architecture.

2 System Architecture

An ideal architecture for "high road", or highlyintegrated, speech translation systems would allow global coordination of, cooperation between, and feedback among, components (speech recognition, analysis, transfer, etc.), thus moving away from linear or pipeline arrangements. For instance, speech recognition, as it moves through an utterance, should be able to benefit from preliminary analysis results for segments earlier in the utterance. The architecture should also be modular, so that a variety of con-

figurations can be tried: it should be possible, for instance, to exchange competing speech recognition components; and it should be possible to combine components not explicitly intended for work together, even if these are written in different languages or running on different machines.

Blackboard architectures have been proposed (Erman and Lesser 1980) to permit cooperation among components. In such systems, all participating components read from and write to a central set of datastructures—the blackboard. To share this common area, however, the components must all "speak a common (software) language." Modularity thus suffers, since it is difficult to assemble a system from components developed separately. Further, blackboard systems are widely seen as difficult to debug, since control is typically distributed, with each component determining independently when to act and what actions to take.

In order to maintain the cooperative benefits of a blackboard system while enhancing modularity and facilitating central coordination or control of components, (Seligman and Boitet 1994 and Boitet and Seligman 1994) proposed and demonstrated a "whiteboard" architecture for speech translation. As in the blackboard architecture, a central datastructure is maintained which contains selected results of all components. However, the components do not access this "whiteboard" directly. Instead, only a privileged program called the Coordinator can read from it and write to it. Each component communicates with the Coordinator and the whiteboard via a go-between program called a *manager*, which handles messages to and from the Coordinator in a set of mailbox files. Because files are used as data holding areas in this way, components (and their managers) can be freely distributed across many machines. (Mailbox files were extensively and successfully used in the French entry in the C-STAR II speech translation demo of July 22, 1999 (*www.c-star.org*).)

Managers are not only mailmen, but interpreters: they translate between the reserved language of the whiteboard and the native languages of the components, which are thus free to differ. In our demo, the whiteboard was maintained in a commercial Lisp-based object-oriented language, while components included independently-developed speech recognition, analysis, and word-lookup components written in C. Overall, the whiteboard architecture can be seen as an adaptation of blackboard architectures for client-server operations: the Coordinator becomes the main client for several components behaving as servers.

Since the Coordinator surveys the whiteboard, in which are assembled the selected results of all components, all represented in a single software interlingua, it is indeed well situated to provide central or global coordination. However, any degree of distributed control can also be achieved by providing appropriate programs alongside the Coordinator which represent the components from the whiteboard side. That is, to dilute the Coordinator's omnipotence, a number of demi-gods can be created. In one possible partly-distributed control structure, the Coordinator would oversee a set of agendas, one or more for each component.

A closely-related effort to create a modular "agent-based" (client-server-style) architecture with a central datastructure, usable for many sorts of systems including speech translation, is described in (Julia et al 1997). Lacking a central board but still aiming in a similar spirit for modularity in various sorts of translation applications is the project described in (Zajac and Casper 1997). Further discussion of speech translation architecture from the alternative viewpoint of the VERBMOBIL system appears in (Görz et al 1996). For discussion of a recent DARPA initiative stressing modular switching of components for experimentation, see (Aberdeen et al 1996).

3 Datastructures

We have argued the desirability for system coordination of a central datastructure where selected results of various components are assembled. The question remains how that datastructure should be arranged. The ideal structure should clarify all of the relevant relationships, in particular clearing up the matter of representational "levels"—a confusing term with several competing interpretations.

(Boitet and Seligman 1994) presented several arguments for the use of interrelated lattices for maintaining components' results. Here I present one possible elaboration, suggesting a multi-dimensional set of structures in which three meanings of "level" are kept distinct (Figure 1).

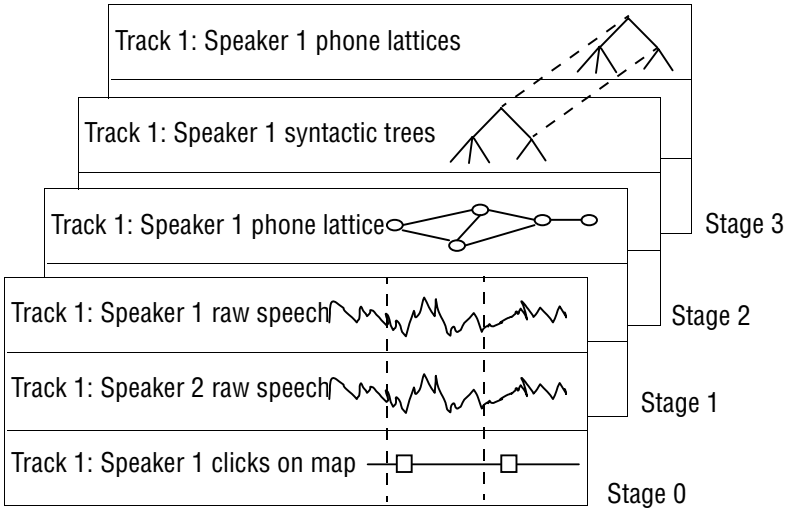

FIGURE 1: Multi-dimensional datastructures for speech translation

We first distinguish an arbitrary number of *Stages of Translation*, with each Stage viewable as a long scroll of paper extending across our view from left to right. Left-right is the time dimension, with earlier elements on the left. The Stage 0 scroll represents the raw input to the speech translation system, including for example the unprocessed speech input from both speakers and the record of one speaker's mouse clicks on an on-screen map, such as might be used for a direction-finding task. In its full extent from left to right, Stage 0 would thus include the raw input for a translation session once complete, e.g. for a dialogue to be translated.

Stage 1 contains the results of the first stage of processing, whatever processes might be involved. This scroll, viewed as unrolling behind Stage 0, might for instance include twin sets of lattices representing the results of phoneme spotting within both speakers' raw input. Stages 2, 3, ... N unroll in turn behind Stage 1, receding in depth. Stage 2 might include source-language syntactic trees; Stage 3 might include semantic structures derived from these trees; and so on, through e.g. MT transfer and generation to the final Stage, a scroll behind all other scrolls, which might contain translated text annotated for speech synthesis. Pointers (diagonal light lines) would indicate relationships between elements in subsequent Stages.

Each Stage can be subdivided both vertically and horizontally. Vertical boundaries (vertical dashed lines) represent appropriate time or segment divisions, probably including utterances. Horizontal divisions represent *Tracks*, since at each Stage several separable signal sources may be under consideration. As already pointed out, Stage 1 in the figure includes two raw speech tracks and a track indicating mouse clicks on a map. Stage 2 might contain, in addition to tracks for the phoneme lattices already mentioned, other tracks (hidden from view) containing F0 curves extracted from the respective speech signals. Different Stages may have different numbers of Tracks, depending on the processes which define them.

Finally, within each Track at a given Stage, we can distinguish varying levels of *Height* on the page—that is, various values on the Y-axis corresponding to given time values along the X-axis. These can be given various interpretations as appropriate for the type of Track in question. When the Track contains syntactic trees, Height corresponds to syntactic rank, i.e. dominance, with dominant nodes usually covering longer time spans than dominated ones.

Confusion regarding the meaning of "level" bedevils many discussions of MT: it sometimes means a stage of processing, sometimes a mode or type of information, and sometimes a gradation of dominance or span. The hope is that, by clearly distinguishing these meanings as Stages, Tracks, or Height within tracks, we can help both programmers and programs keep their bearings amid a welter of information.

The multi-dimensional structures just described bear some resemblance to the *three-dimensional charts* of (Barnett et al 1990), used to track relationships between syntactic and semantic structures during analysis of queries to CYC knowledge bases (Lenat and Guha 1990). They were developed independently, however. 3-D charts were restricted to

two depths or stages (syntactic and semantic), lacked tracks, and made no explicit reference to height or rank.

The whiteboard demo reported in (Seligman and Boitet 1994) likewise made only partial use of the multidimensional structure: Stages and Height were explicitly represented and shown in the graphical user interface, with explicit representation of relations between structures in subsequent Stages; but Tracks were not yet included.

4 Interface between Speech Recognition and MT Analysis

In a certain sense, speech recognition and analysis for MT are comparable problems. Both require the recognition of the most probable sequences of elements. In speech recognition, sequences of short speech segments must be recognized as phones, and sequences of phones must be recognized as words. In analysis, sequences of words must be recognized as phrases, sentences, and utterances.

Despite this similarity, current speech translation systems use quite different techniques for phone, word, and syntactic recognition. Phone recognition is generally handled using hidden Markov models (HMMs); word recognition is often handled using Viterbi-style search for the best paths in phone lattices; and sentence recognition is handled through a variety of parsing techniques.

It can be argued that these differences are justified by differences of scale, perplexity, and meaningfulness. On the other hand, they introduce the need for interfaces between processing levels. The processors may thus become black boxes to each other, when seamless connection and easy communication might well be preferable. In particular, word recognition and syntactic analysis (of phrases, sentences, and utterances) should have a lot to say to each other: the probability of a word should depend on its place in the top-down context of surrounding words, just as the probability of a phrase or larger syntactic unit should depend on the bottom-up information of the words which it contains.

To integrate speech recognition and analysis more tightly, it is possible to employ a single grammar for both processes, one whose terminals are phones and whose non-terminals are words, phrases, sentences, etc.[4] This phone-grounded strategy was used to good effect e.g. in the HMM-LR speech recognition component of the ASURA speech translation system (Morimoto et al 1993), in which an LR parser extended a parse phone by phone and left to right while building a full syntactic tree.[5] The technique worked well for scripted examples. For spontaneous examples, however, performance was unsatisfactory, because of the gaps, repairs, and other noise common in spontaneous speech. To deal with such structural problems, an island-driven parsing style might well be preferable. An island-based chart parser, like that of (Stock et al 1989), would be a good candidate.

However, chart initialization presents some technical problems. There is no difficulty in computing a lattice from spotted phones, given information regarding the maximum gap and overlap of phones. But it is not trivial to convert that lattice into a "chart" (i.e. multipath finite state automaton) without introducing spurious extra paths. The author has implemented a Common Lisp program which does so correctly, based on an algorithm by C. Boitet (Seligman et al 1998a). The algorithm tracks, for each node of an automaton under construction, the lattice arcs which it reflects and the lattice nodes at their origins and extremities. An extension of the procedure permits the inclusion of null, or epsilon, arcs in the output automaton. The method has been successfully applied to lattices derived from dictionaries, i.e. very large corpora of strings. (Full source code and pseudocode are available from the authors.) Experiments with bottom-up island-driven chart parsing from charts initialized with phones are anticipated.

5 Use of Pauses for Segmentation

It is widely believed that prosody can prove crucial for speech recognition and analysis of spontaneous speech. (For one example of extensive related work in the framework of the VERBMOBIL system, see (Kompe et al 1997).) Several aspects of prosody might be exploited: pitch contours, rhythm, volume modulation, etc. However, (Seligman et al 1996) propose focusing on natural pauses as an aspect of prosody which is both important and relatively easy to detect automatically.[6]

4. Inclusion of other levels is also possible. At the lower limit, assuming the grammar were stochastic, one could even use sub-phone speech segments as grammar terminals, thus subsuming even HMM-based phone recognition in the parsing regime. At an intermediate level between phones and words, syllables could be used.

5. The parse tree was not used for analysis, however. Instead, it was discarded, and a unification-based parser began a new parse for MT purposes on a text string passed from speech recognition.

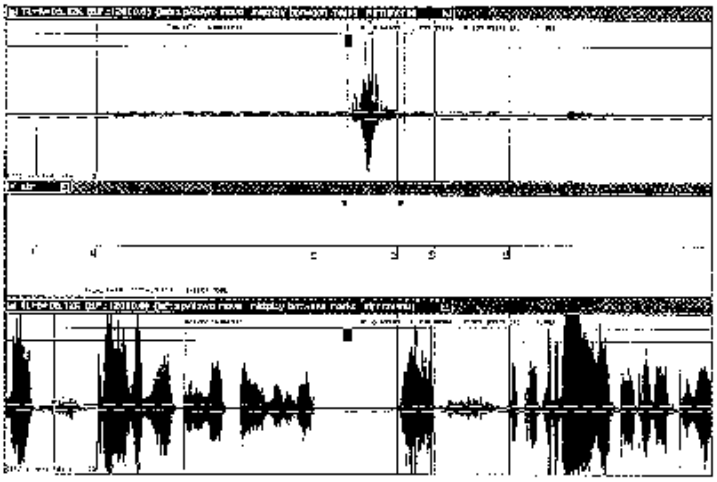

FIGURE 2: Interface used by the pause tagger

Given the frequency of utterances in spontaneous speech which are not fully well-formed—which contain repairs, hesitations, and fragments—strategies for dividing and conquering utterances would be quite useful. The suggestion is that natural pauses can play a part in such a strategy: that *pause units*, or segments within utterances bounded by natural pauses, can provide chunks which (1) are reliably shorter and less variable in length than entire utterances and (2) are relatively well-behaved internally from the syntactic viewpoint, though analysis of the relationships among them appears more problematic.

Our investigation began with transcriptions of four spontaneous Japanese dialogues concerning a simulated direction-finding task. The dialogues were carried out in the EMMI-ATR Environment for Multimodal Interaction (Loken-Kim et al 1993; Furukawa et al 1993), two using telephone connections only, and two employing onscreen graphics and video as well. In each 3 to 7 minute dialogue, a caller pretending to be at Kyoto station received from a pre-trained "agent'' directions to a conference center and/or hotel. In the multimedia setup, both the caller and agent could draw on onscreen maps and exchange typed information.

Morphologically tagged transcripts of the conversations were divided into turns by the transcriber, and included hesitation expressions and other natural speech features. We then added to the transcripts information concerning the placement and length of significant pauses. For our purposes, a significant pause was either a juncture of any length where breathing was clearly indicated (sometimes a bit less than 300 milliseconds) or a silence lasting approximately 400 milliseconds or more.

To facilitate pause tagging, we prepared a customized configuration of the Xwaves speech display program (Xwaves 1993) so that it showed synchronized but separate speech tracks of both parties on screen (Figure 2). The pause tagger, referring to the transcript, could use the mouse to draw labeled lines through the tracks indicating the starts and ends of turns; the starts and ends of segments within turns; and the starts and ends of response syllables which occur during the other speaker's turn. Visual placement of labels was quite clear in most cases. As a secondary job, the tagger inserted a special character into a copy of the transcript text wherever pauses occurred within turns.

After tagging, labels, bearing exact timing information, were downloaded to separate files. Because there should be a one-to-one mapping between labeled pauses within turns and marked pause locations in the transcript, it was then possible to create augmented transcripts by substituting accurate pause length information into the transcripts at marked pause points.

In studying the augmented transcripts, four specific questions were addressed: (1) Are pause units reliably shorter than whole utterances? If they were not, they could hardly be useful in simplifying analysis. It was found however, that, in the corpus investigated, pause units are in fact about 60% the length of entire

6. A related but distinct proposal appears in (Hosaka et al 1994).

utterances, on the average, when measured in Japanese morphemes. The average length of pause units was 5.89 morphemes, as compared with 9.39 for whole utterances. Further, pause units are less variable in length than entire utterances: the standard deviation is 5.79 as compared with 12.97. (2) Would hesitations give even shorter, and thus perhaps even more manageable, segments if used as alternate or additional boundaries? The answer seems to be that because hesitations so often coincide with pause boundaries, the segments they mark out are nearly the same as the segments marked by pauses alone. No combination of expressions was found which gave segments as much as one morpheme shorter than pause units on average. (3) Is the syntax within pause units relatively manageable? A manual survey showed that, once hesitation expressions are filtered from them, some 90% of the pause units studied can be parsed using standard Japanese grammars; a variety of special problems appear in the remaining 10%. (4) Is translation of isolated pause units a possibility? We found that a majority of the pause units in four dialogues gave understandable translations into English when translated by hand.

The study provided encouragement for a "divide and conquer" analysis strategy, in which parsing and perhaps translation of pause units is carried out before, or even without, attempts to create coherent analyses of entire utterances.

As mentioned, parsability of spontaneous utterances might be enhanced by filtering hesitation expressions from them in preprocessing. Research on spotting techniques for such expressions would thus seem to be worthwhile. Researchers can exploit speakers' tendency to lengthen hesitations, and to use them just before or after natural pauses.

Use of pause information for "dividing utterances into meaningful chunks" during speech translation of Japanese is described in (Takezawa et al 1999). Pauses are used as segment boundaries in several commercial dictation products, but no descriptions are available.

6 Example-based ST

Example-based translation (Nagao 1984; Sato 1991) is translation by analogy. An example-based system translates source-language sentences by reference to an *example base*, or set of source-language utterances paired with their target-language equivalents. In developing such a system, the hope is to improve translation quality by reusing correct and idiomatic translations; to partly automate grammar development; and to gain insight into language learning.

Two EBMT systems are now being applied to speech translation: the TDMT (Transfer-driven MT) system developed at ATR (Furuse and Iida 1996; Iida et al 1996; Sumita and Iida 1992), used in the ATR-MATRIX speech translation system (Takezawa et al 1999); and the PanEBMT system (Brown 1996) of CMU, used along with transfer-based MT within the Multi-Engine MT architecture in the DIPLOMAT speech translation system (Frederking et al 1997).

Despite their common aims, the two systems differ substantially. The ATR system aims to supply a complete translation single-handed, and accordingly includes a full parser for utterances and a hand-built grammar (set of language patterns) to go with it. The CMU system, by contrast, operates as a component of a larger system: in general, its aim is to supply possible partial translations, or translation chunks, to be placed on a chart along with chunks supplied by other translation engines.[7] For this mission, the system requires neither parser nor grammar, relying instead on heuristics to align sub-elements of sentences in the example base at training time. Once it has put its chunks in place during translation, a separate process, belonging to the Multi-Engine MT architecture, will employ a statistical language model to select the best path through the pre-stocked chart in order to assemble the final output.

As the suggestions below relate to a tree-oriented and end-to-end view of example-based processing, the primary concern will be with systems of the ATR type. We begin with a sketch of this methodology.

Consider the Japanese noun phrase *kyouto no kaigi*. Its literal translation is "conference of Kyoto," but a more graceful translation would be "conference in Kyoto" or "Kyoto conference." We could hope to provide such improved translations if we had an example base showing for instance that *toukyou no kaigi* had been translated as "conference in Tokyo" or "Tokyo conference," and that *nyuu yooku no kaigi* had been rendered as "conference in New York" or "New York conference." The strategy would be to recognize a close similarity between the new input *kyouto no kaigi* and these previously translated noun phrases, based on the semantic similarity between *kyouto* on one hand and *toukyou* and *nyuu yooku* on the other. The same sort of pattern matching could be per-

7. PanEBMT operates solo only when the entire source expression can be rendered with a single memorized target expression.

formed against a noun phrase in the example base differing from the input at more one point: for example, *toukyou no mitingu* ("meeting in Tokyo"), where *mitingu* ("meeting") is semantically similar to *kaigi* ("conference"). At any number of such comparison points, semantic similarity of the relevant expressions can be assessed by reference to a semantic hierarchy—for example, a type hierarchy of semantic tags supplied by a thesaurus. A thesaurus associates a lexical item like *kaigi* with one or more semantic tags (e.g. CITY, SOCIAL-EVENT); and the similarity of two semantic tags can be defined as the distance one must rise in the relevant semantic hierarchy to reach a node which dominates both tags: the further, the more semantically distant. The four-level hierarchy of the *Kadokawa New World Category Dictionary* (Ohno and Hamanish 1981) has been used in this way in several studies.

Different translations of the Japanese *no* construction, for example as the English possessive (*tanaka-san no kuruma*, "Tanaka's car") would be distinguished by the distinct semantic types of their respective comparison points—in this case, e.g. PERSON and VEHICLE.

By replacing each comparison point in an expression like *kyouto no kaigi* with a variable, we can obtain a pattern like [?X no ?Y]. Such patterns can be embedded, giving [[?X no ?Y] no ?Z] or [?X no [?Y no ?Z]]. If we then receive an input like *kyouto no kaigi no ronbun* ("Kyoto conference paper," "paper at the conference in Kyoto"), we can determine which bracketing is most sensible—that is, we can parse the input—by extending the techniques already discussed for gauging semantic similarity. One possibility is to designate a *head* for each pattern, and to posit that a pattern's overall semantic type is the type of its head. Then semantic similarity scores can be calculated between an input like *kyouto no kaigi no ronbun* and an entire set of embedded patterns—that is, an entire pattern tree—by propagating similarity scores outward (upward). One can calculate similarity scores for several possible bracketings (trees), and choose the bracketing most semantically similar to the input. In this way, the calculation of semantic similarity guides structural disambiguation during analysis.

Having outlined the essentials of example-based processing in the tree-oriented style, we are now ready to discuss possible elaborations. The first involves the degree of separation between stages of example-based translation.

6.1 Separation of Example-based Analysis, Transfer, and Generation

Recall that semantic similarity calculation can be used to select an embedded set of patterns (a parse tree) from among several competitors. If each source language pattern (i.e. sub-tree) is associated with a unique target language pattern which provides its translation, then the selection of a complete source language tree will simultaneously and automatically provide a corresponding target language tree. In this way, an example-based analysis process can be made to automatically provide a *transfer* process as well—that is, a mapping of source language structures into target language structures. TDMT intentionally combines analysis and transfer in this way. The combination is seen as an advantage: the same mechanism which handles structural disambiguation simultaneously selects the right translation from among several candidates. However, the combination of phases does raise issues concerning the role of transfer in handling translation ambiguity and structural mismatches.

First, some translation applications may require an explicit account of translation ambiguity—that is, of the possibility of translating a given sub-tree or node in more than one way. For such applications, transfer might be treated as a separate phase of translation from source-language parsing. That is, since considerations of semantic similarity can guide the selection of target structure—just as they can guide the choice of analysis tree—we can recognize the possibility of *example-based transfer* as separate from example-based analysis. Furthermore, depending on the depth of analysis, even once a target language tree has been selected, ambiguity may arise in selecting target language surface forms to express it. Thus a separate *example-based generation* phase also becomes a possibility.

A second issue relates to structural mismatches between source and target. Should they be handled in the transfer phase of translation? Consider these translations, for example: *zou wa, hana ga nagai* (lit. "As for elephants, noses are long") > *elephants have long noses*; *Taeko wa, kami no ke wa nagai* (lit. "As for Taeko, hair is long") > *Taeko's hair is long or Taeko has long hair*; *watashi wa, taeko ga suki desu* (lit. "As for me, Taeko is beloved") > *I like/love Taeko*. In these cases, language-internal considerations dictate non-flat analyses on both the source and target sides. However, in each case, the source tree is differently configured from the target tree. Thus, to represent the correspondences completely, it is insufficient to

simply map one source node (one source pattern) into one target node (one target pattern); rather, we need to inter-map arbitrary sub-tree configurations (embedded pattern sets). In current implementations of tree-oriented EBMT, such general mappings between sub-trees are not supported during transfer; rather, they are handled by special-purpose post-processing routines. It might prove easier to arrange a more general treatment for such intermappings if transfer were treated as a separate translation phase.

An experiment reported in (Sobashima and Iida 1995) and (Sobashima and Seligman 1994) takes a first step toward clear separation of example-based translation phases: it presents an example-based treatment of analysis only. (Further information is given below.) A distinct example-based transfer phase including facilities for inter-mapping embedded patterns was envisaged, but has not yet been implemented.

6.2 Multiple Dimensions of Similarity

So far we have discussed the measurement of similarity along the semantic scale only. But utterances and structures can be compared along other dimensions as well. Thus for example, when assessing the similarity between a given pattern and the input pattern to be translated, we could ask not only how *semantically* similar its contained elements are to those of the input pattern, but how *syntactically* similar as well, or how *graphologically* or *phonologically* similar.

(Sobashima and Iida 1995) and (Sobashima and Seligman 1994) describe facilities for measuring and combining several sorts of similarity. Syntactic similarity, for instance, is measured with reference to a syntactic ontology, comparable to the thesaurus-based semantic hierarchy discussed above; and a score indicating overall similarity of respective variable elements in two patterns is calculated by combining syntactic and semantic similarity scores. The reported implementation also considered, as a factor in overall similarity, a score indicating graphological similarity: 1 for a complete match, and 0 in other cases. Future versions, however, might instead measure phonological similarity—for instance, by means of a phone type ontology indicating e.g. that /sh/ and /ch/ are similar sounds, while /sh/ and /k/are more different. Below, we briefly indicate how multiple similarity dimensions entered into the calculation of overall similarity.

Once we recognize the possibility of considering phonological similarity as a factor in overall similarity between patterns, we move example-based translation beyond text translation into the area of speech translation. We could, for instance, attempt to disambiguate the speech act of an utterance by comparing the prosodic contours of its elements with the contours of elements of labeled utterances in a database. Such prosodic comparisons might help, for example, to distinguish politely hesitant statements and yes-no questions in Japanese. These utterance types are syntactically marked by final particles *ga* and *ka*, which are phonologically quite difficult to distinguish; their prosodies, however, tend to be quite distinct.

In any case, use of similarity measurements along multiple dimensions as an aid to disambiguation would be very much in the spirit of the "high road", or integrative, approach to speech translation discussed throughout.[8]

6.3 Both Top-down and Bottom-up

In most current example-based systems, the applicability of a pattern is judged by the semantic match of its sub-elements against those of the input. These are *bottom-up* similarity judgments: the sub-elements provide evidence for the presence of the pattern as a whole. Usually absent, however, are corresponding *top-down* similarity judgments whereby the patterns give evidence for the sub-elements. (Sobashima and Iida 1995) and (Sobashima and Seligman 1994), however, do demonstrate application of both bottom-up and top-down similarity constraints. Further, similarity is measured in both directions along several dimensions (syntactic, semantic, and others), as suggested above. We now briefly describe the method.

First, some necessary background. Consider a *linguistic expression*, which may be either atomic or complex. Complex expressions are composed of variables and/or fixed lexical elements, as in [() no ()]. We calculate the *elemental similarity*, or **E-Sim**, of two expressions as a combined function of their syntactic, semantic, and phonological or graphological

8. The sort of generalization suggested here—from graded semantic similarity measurements to graded measurements of similarity along multiple dimensions—should not be confused with that of Generalized EBMT, the example-based technique proposed for CMU's PanEBMT engine. That engine utilizes no graded similarity measurements along any scale. Its generalization instead involves substitution of semantic tags for lexical items in examples and in input, so that e.g. "John Hancock was in Washington" becomes "<PERSON> was in <CITY>."

similarites. (In this calculation, fixed elements are treated differently from variable elements, and variable elements can be weighted to varying degrees: the heads of complex structures are differently weighted than non-heads.)

Now we are ready to consider top-down vs. bottom-up similarity measurement. We calculate the *structural similarity*, or *bottom-up* similarity, of two complex expressions by combining the elemental similarities of their respective elements. By contrast, the *top-down* factor in the similarity of two expressions **A** and **B** is a measure of the similarity of their respective contexts. We call this factor the *contextual similarity* of expressions **A** and **B**, and calculate it as the sum of the elemental similarities of their respective left and right neighbor expressions: C-Sim (A, B) = E-Sim(L-neighbor$_A$, L-neighbor$_B$) + E-Sim (R-neighbor$_A$, R-neighbor$_B$).

The final, or integrated, similarity score **Sim** for expressions $\mathbf{S}_1$ and $\mathbf{S}_2$, then, is the combination of their structural (bottom-up) similarity and their contextual (top-down) similarity: $Sim(S_1, S_2)$ = S-$Sim(S_1, S_2)$ * C-$Sim(S_1, S_2)$.

We have seen that **Sim** incorporates multi-dimensional similarity measurements applied both top-down and bottom-up (TD + BU). The next question is how to apply this score for example-based analysis. We now outline the method proposed in the cited papers.

6.4 Analysis with Multi-Dimensional, TD + BU Similarity Measurements

We can consider the training stage first. In this stage, an example base is prepared by bracketing and labeling the training corpus by hand. The labeling entry for a complex expression includes the number of elements it contains; the set of syntactic, semantic, and other classifying features of the complex structure as a whole; the classifying features of each sub-element; and the classifying features of the left and right contexts.

Now on to the analysis itself. After morphological processing, with access to a lexicon giving classifying feature sets (perhaps multiple sets) for each terminal, the main routine proceeds as follows: (1) Search the example base for the expression most similar to any contiguous subsequence in the input: find the longest similar matches from position 1 in the input, then from position 2, and so on, terminating if a perfect match is found. (2) Reduce, or rewrite, the covered subsequence, passing its similarity features to the rewritten structure. Go to (1). Continue the cycle until no further reduction is possible.

A preliminary experiment was conducted on 132 English and 129 Japanese sentences. This corpus was too small to permit meaningful statistical evaluation, but we can say that numerous sentences were successfully analyzed which might have yielded massive structural ambiguity. One example: "However, we do have single rooms with a shower for eight dollars and night and twin rooms with a bath for a hundred and forty dollars a night." Here, many spurious combinations, e.g. "shower for eighty dollars a night and twin rooms," were ignored in favor of the proper interpretations. Successful analysis of various uses of the article *a* was particularly notable. A full trace appears in the cited papers.

6.5 Similarity vs. Frequency

We have been discussing the uses of similarity calculations for the resolution of various sorts of ambiguity. We conclude this section by contrasting similarity-based disambiguation and *probability*-based disambiguation, an approach which is more widely studied at present. Several current parsers (e.g. Black et al 1993) are trained to resolve conflicts among competing analyses by using information about the relative frequencies, and thus probabilities, of the combinations of elements in question. At short range, n-gram statistics are used; at longer ranges, stochastic rules.

Several of the considerations raised above with respect to similarity-based disambiguation apply equally to probability-based disambiguation. For example, (Jurafsky 1993) stresses the need for multidimensional processing: in his parser—based upon the theory of grammatical constructions (Fillmore et al 1988; Kay 1990) and claimed to model several features of human parsing as observed in psycho-linguistic experiments—semantic as well as syntactic frequencies and probabilities are brought to bear in selecting the proper parse. Also stressed is the need for both top-down and bottom-up statistics in evaluating of parse tree as a whole.

Ideally, disambiguation approaches based upon similarity and approaches based upon occurrence probability should complement each other. However, I am aware of no attempts to combine the two.

7 Cue-based Speech Acts

Speech act analysis (Searle 1969)—analysis in terms of illocutionary acts like INFORM, WH-QUESTION,

REQUEST, etc.—can be useful for speech translation in numerous ways. Six uses, three related to translation and three to speech processing, will be mentioned here. Concerning translation, it is necessary to:

- *Identify the speech acts of the current utterance* Speech act analysis of the current utterance is necessary for translation. For instance, the English pattern "can you (VP, bare infinitive)?" may express either an ACTION-REQUEST or a YN-QUESTION (yes/no-question). Resolution of this ambiguity will be crucial for translation.

- *Identify related utterances* Utterances in dialogues are often closely related: for instance, one utterance may be a prompt and another utterance may be its response; and the proper translation of a response often depends on identification and analysis of its prompt. For example, Japanese *hai* can be translated as *yes* if it is the response to a YN-QUESTION, but as *all right* if it is the response to an ACTION-REQUEST. Further, the syntax of a prompt may become a factor in the final translation. Thus, in a responding utterance *hai, sou desu* (meaning literally "yes, that's right"), the segment *sou desu* may be most naturally translated as *he can*, *you will*, *she does*, etc., depending on the structure and content of the prompting question. The recognition of such prompt-response relationships will require analysis of typical speech act sequences.

- *Analyze relationships among segments and fragments* Early processing of utterances may yield fragments which must later be assembled to form the global interpretation for an utterance. Speech act sequence analysis should help fit fragments together, since we hope to learn about typical act groupings.

Concerning speech processing, it is necessary to:

- *Predict speech acts to aid speech recognition* If we can predict the coming speech acts, we can partly predict their surface patterns. This prediction can be used to constrain speech recognition. As already mentioned, for instance, Japanese utterances ending in *ka* and *ga*—respectively, YN-QUESTIONS and INFORMS—are difficult to distinguish phonologically. We earlier considered the use of prosodic information in resolving this uncertainty. Predictions as to the relative likelihood of these speech acts in a given context should further aid recognition.

- *Provide conventions for prosody recognition* Once spontaneous data is labeled, speech recognition researchers can try to recognize prosodic cues to aid in speech act recognition and disambiguation. For instance, they can try to distinguish segments expressing INFORMS and YN-QUESTIONS according to the F0 curves associated with them—a distinction which would be especially useful for recognizing YN-QUESTIONS with no morphological or syntactic markings.

- *Provide conventions for speech synthesis* Similarly, speech synthesis researchers can try to provide more natural prosody by exploiting speech act information. Once relations between prosody and speech acts have been extracted from corpora labeled with speech act information, researchers can attempt to supply natural prosody for synthesized utterances according to the specified speech acts. For instance, more natural pronunciations can be attempted for YN-QUESTIONS, or for CONFIRMATION-QUESTIONS (including tag questions in English, as in *The train goes east, doesn't it?*).

While a well-founded set of speech act labels would be useful, it has not been clear what the theoretical foundation should be. As a result, no speech act set has yet become standard, despite considerable recent effort. (See for example the website of the Discourse Resource Initiative (Duff 1999), with links to recent workshops, or browse (Walker 1999), especially regarding attempted standardization of Japanese discourse labeling (Ichikawa et al 1999)). Labels are still proposed intuitively, or by trial and error.

Speakers' goals can certainly be analyzed in many ways. However, (Seligman et al 1995) hypothesize that only a limited set of goals is conventionally expressed in a given language. For just these goals, relatively fixed expressive patterns are learned by speakers when they learn the language. In English, for instance, it is conventional to express certain suggestions or invitations using the patterns "Let's *" or "Shall we *?" In Japanese, one conventionally expresses similar goals via the patterns "(V, combining stem)*mashou*" or "(V, combining stem)*masen ka*?"

The proposal is to focus on discovery and exploitation of these conventionally-expressible speech acts, or *Cue-based Speech Acts* (CAs).[9] The relevant expressive patterns and the contexts within which they are found have the great virtue of being objectively observable; and assuming the use of these patterns

9. Called Communicative Acts in (Seligman et al 1995) and Situational Formulas in (Seligman 1991).

is common to all native speakers, it should be possible to reach a consensus classification of the patterns according to their contextualized meaning and use. This functional classification should yield a set of language-specific speech act labels which can help to put speech act analysis for speech translation on a firmer foundation.

The first reason to analyze speech acts in terms of observable linguistic patterns, then, is the measure of objectivity thus gained: the discovery process is to some degree empirical, data-driven, or corpus-based. A second reason is that automated cue-based analysis, being shallow or surface-bound, should be relatively quick as opposed to plan-based analysis. Plan-based analysis may well prove necessary for certain purposes, but it is quite expensive. For applications like speech translation which must be carried out in nearly real time, it seems wise to exploit shallow analysis as far as possible.

With these advantages of cue-based processing—empirical grounding and speed—come certain limitations. When analyzing in terms of CAs, we cannot expect to recognize all communicative goals. Instead, we restrict our attention to communicative goals which can be expressed using conventional linguistic cue patterns. Communicative goals which cannot be described as Cue-based Speech Acts include utterance goals which are expressed non-conventionally (compare the non-conventional warning *May I call your attention to a potentially dangerous dog* to the conventional WARNING *Look out for the dog!*); or goals which are expressed only implicitly (*It's cold outside* as an implicit request to shut the window); or goals which can only be defined in terms of relations between utterances. (While speakers often repeat an interlocutor's utterance to confirm it, we do not use a REPEAT-TO-CONFIRM CA, since it is apparently signaled by no cue patterns, and thus could only be recognized by noting inter-utterance repetition.)

Given that the aim is to classify expressive patterns according to their meaning and function, how should this be done? (Seligman 1991) and (Seligman et al 1995) describe a paraphrase-based approach: native speakers are polled as to the essential equivalence of expressive patterns in specified discourse contexts. If by consensus several patterns can yield paraphrases which are judged equivalent in context, and if the resulting pattern set is not identical to any competing pattern set, then it can be considered to define a Cue-based Speech Act. ((Knott and Dale 1992) and (Knott 1996) describe a similar *substitution*-based approach to the discovery of discourse relations, as opposed to speech acts.)

Cue-based Speech Acts are defined in terms of monolingual conventions for expressing certain communicative goals using certain cue patterns. For translation purposes, however, it will be necessary to compare the conventions in language A with those in language B. With this goal in mind, the discovery procedure was applied to twin corpora of Japanese-Japanese and English-English spontaneous dialogues concerning transportation directions and hotel accommodations (Loken-Kim et al 1993). CAs were first identified according to monolingual criteria. Then, by observing translation relations among the English and Japanese cue patterns, the resulting English and Japanese CAs were compared. Interestingly, it was found that most of the proposed CAs seem valid for both English and Japanese: only two out of 27 CAs seem to be monolingual for the corpus in question.

We have been outlining a cue-based approach to recognition of speech or discourse acts, with the assumption that some sort of parsing would be employed to recognize cue patterns. This methodology can be compared with statistical recognition approaches: speech or discourse act labels are posited in advance, and statistical models are subsequently built which attempt to identify the acts according to their sequence (Reithinger 1995; Nagata and Morimoto 1993) or according to the words they contain (Alexandersson et al 1997; Reithinger and Klesen 1997).

Certain speech act sequences may indeed turn out to be typical; and certain words may indeed prove to be unusually common in, and thus symptomatic of, arbitrarily defined speech acts. Thus statistical techniques are indeed likely to be helpful for recognition of conventional speech acts when they are implied or expressed non-conventionally, or for recognition of speech acts which are not conventional but nevertheless appear useful for some applications. Further, even for conventional speech acts which are conventionally expressed, efficiency considerations may sometimes favor statistical recognition techniques over pattern recognition: once cue-based acts were identified using our methods and a sufficiently large training corpus had been hand-labeled, statistical models might certainly be built to permit efficient identification in context. However, statistical recognition approaches alone cannot provide a principled way to *discover* (that is, posit or hypothesize) the labels in the first place, and this is what we seek.

The current CA set has been applied in three studies: (Black and Campbell 1995) attempted to associate speech acts, including CAs, with intonation contours in hopes of improving speech synthesis; (Iwadera et al 1995) employed the CA set in attempts to parse discourse structure; and (Jokinen and Tanaka 1998) used CAs in topic tracking experiments.

8 Tracking Lexical Co-occurrences

In the processing of spontaneous language, the need for predictions at the morphological or lexical level is clear. For bottom-up parsing based on phones or syllables, the number of lexical candidates is explosive. It is crucial to predict which morphological or lexical items are likely so that candidates can be weighted appropriately. (Compare such lexical prediction with the predictions from Cue-based Speech Acts discussed above. In general, it is hoped that by predicting CAs we can in turn predict the structural elements of their cue patterns. We are now shifting the discussion to the prediction of open-class elements instead. The hope is that the two sorts of prediction will prove complementary.)

N-grams provide such predictions only at very short ranges. To support bottom-up parsing of noisy material containing gaps and fragments, longer-range predictions are needed as well. Some researchers have proposed investigation of associations beyond the n-gram range, but the proposed associations remain relatively short-range (about five words). While stochastic grammars can provide somewhat longer-range predictions than n-grams, they predict only within utterances. Our interest, however, extends to predictions on the scale of several utterances.

Thus (Seligman et al 1999; Seligman 1994a) propose to permit the definition of windows in a transcribed corpus within which co-occurrences of morphological or lexical elements can be examined. A flexible set of facilities (CO-OC) has been implemented in Common Lisp to aid collection of such discourse-range co-occurrence information and to provide quick access to the statistics for on-line use.

A window is defined as a sequence of minimal segments, where a segment is typically a turn, but can also be a block delimited by suitable markers in the transcript.

Sparse data is somewhat less problematic for long-range than for short-range predictions, since it is in general easier to predict what is coming "soon" than what is coming next. Even so, there is never quite enough data; so smoothing will remain important. CO-OC can support various statistical smoothing measures. However, since these techniques are likely to remain insufficient, a new technique for semantic smoothing is proposed and supported: researchers can track co-occurrences of semantic tokens associated with words or morphs in addition to co-occurrences of the words or morphs themselves. The semantic tokens are obtained from standard on-line thesauri. The benefits of such semantic smoothing appear especially in the possibility of retrieving reasonable semantically-mediated associations for morphs which are rare or absent in a training corpus.

Subsections 8.1 to 8.3 describe CO-OC's operations in somewhat greater detail. Subsection 8.4 sketches possible application for the co-occurrence information harvested by the program.

8.1 Windows and Conditional Probabilities

As mentioned, we first permit the investigator to define minimal segments within the corpus: these may be utterances, sections bounded by pauses or significant morphemes such as conjunctions, hesitations, postpositions, etc. Windows composed of several successive minimal segments can then be recognized: Let Si be the current segment and N be the number of additional segments in the window as it extends to the right. N = 2 would, for instance, give a window three segments long with Si as its first segment. Then if a given word or morpheme M1 occurs (at least once) in the initial segment, Si, we attempt to predict the other words or morphemes which will co-occur (at least once) anywhere in the window.

Specifically, a conditional probability Q can be defined as follows: $Q(M1, M2) = P(M2 \in Si \cup Si+1 \cup Si+2 \ldots Si+N \mid M1 \in Si)$, where M1, M2 ... are morphemes, S1, S2 ... are minimal segments, and N is the width of window in segments. Q is thus the conditional probability that M2 is an element of the union of segments Si, Si + 1, Si + 2, and so on up to Si + N, given that M1 is an element of Si. Both the segment definition and the number of segments in a window can be adjusted to vary the range over which co-occurrence predictions are attempted.

For initial experiments, we used a morphologically-tagged corpus of 16 spontaneous Japanese dialogues concerning direction-finding and hotel arrangements (Loken-Kim et al 1993). We collected common-noun/common-noun, common-noun/verb, verb/common-noun, and verb/verb conditional probabilities in a three-segment window (N = 2). Conditional probability Q was computed among all morph

pairs for these classes and stored to a database; pairs scoring below a threshold (0.1 for the initial experiments) were discarded. We also computed and stored the mutual information for each morph pair, using the standard definition as in (Fano 1961).

Fast queries of the database are then enabled. A central function is GET-MORPH-WINDOW-MATES, which provides all the window mates for a specified morph which belong to a specified class and have scores above a specified threshold for the specified co-occurrence measure (conditional probability or mutual information).

The intent is to use such queries in real time to support bottom-up, island-driven speech recognition and analysis. To support the establishment of island centers for such parsing, we also collect information on each corpus morph in isolation: its hit count and the segments it appears in, its unigram probability and probability of appearance in a given segment, etc. Once island hypotheses have been established based on this foundation, co-occurrence predictions will come into play for island extension.

8.2 Semantic Smoothing

As mentioned, CO-OC supports the use of standard statistical techniques (Nadas 1985) for smoothing both conditional probability and mutual information. In addition, however, we enable semantic smoothing in an innovative way. Thesaurus categories — *cats* for short — are sought for each corpus morph (and stored in a corpus-specific customized thesaurus for fast access). The common-noun *eki* (station), for instance, has among others the cat label "725a" (representing a semantic class of posts-or-stations) in the standard Kadokawa Japanese thesaurus (Ohno and Hamanish 1981).

Equipped with such information, we can study the co-occurrence within windows of cats as well as morphs. For example, using N = 2, GET-CAT-WINDOW-MATES finds 36 cats co-occurring with "725a", one of the cats associated with *eki* (station), with a conditional probability Q greater than 0.10, including "459a" (*sewa*, taking-care-of or looking-after), "216a" (*henkou*, transfer), and "315b" (*ori*, getting-off). Since we have prepared an indexed reverse thesaurus for our corpus, we can quickly find the corpus morphs which have these cat labels, respectively *miru*, "look", *mieru*, "can see, visible"; *magaru*, "turn"; and *oriru*, "get off". The resulting morphs are related to the input morph *eki* via semantic rather than morph-specific co-occurrence. They thus form a broader, smoothed group.

This semantic smoothing procedure — morph to related cats, cats to co-occurring category window-mates, cats to related morphs — has been encapsulated in the function GET-MORPH-WINDOW-MATES-VIA-CATS. It permits filtering, so that morphs are output only if they belong to a desired morphological class and are mediated by cats whose co-occurrence likelihood is above a specified threshold.

Thesaurus categories are normally arranged in a type hierarchy. In the Kadokawa thesaurus, there are four levels of specificity: "725a" (posts-or-stations), mentioned above, belongs to a more general category "725" (stations-and-harbors), which in turn belongs to "72" (institutions), which belongs to "7" (society). Accordingly, we need not restrict co-occurrence investigation to cats at the level given by the thesaurus. Instead, knowing that "725a" occurred in a segment Si, we can infer that all of its ancestor cats occurred there as well; and can seek and record semantic co-occurrences at every level of specificity. This has been done; and GET-MORPH-WINDOW-MATES-VIA-CATS has a parameter permitting specification of the desired level of semantic smoothing. The more abstract the level of smoothing, the broader the resulting group of semantically-mediated morpheme co-occurrences. The most desirable level for semantic smoothing is a matter for future experimentation.

8.3 Evaluation

We are presently reporting the implementation of facilities intended to enable many experiments concerning morphological and morpho-semantic co-occurrence; the experiments themselves remain for the future. Clearly, further testing is necessary to demonstrate the reliability and usefulness of the approach. (A principle aim would be to determine how large the corpus must be before consistent co-occurrence predictions are obtained.) Nevertheless, some indication of the basic usability of the data is in order.

Tools have been provided for comparing two corpora with respect to any of the fields in the records relating to morphs, morph co-occurrences, cats, or cat co-occurrences. Using these, we treated 15 of our dialogues as a training corpus, and the one remaining dialogue as a test corpus. We compared the two corpora in terms of conditional probabilities for morph co-occurrences. (In both cases, statistically unsmoothed scores were used for simplicity of interpretation.)

We found 5162 co-occurrence pairs above a conditional probability threshold of 0.10 in the training corpus and 1552 in the test. Since 509 pairs occurred in both corpora, the training corpus covered 509 out of 1552, or 33 percent, of the test corpus. That is, one third of the morph co-occurrences with conditional probabilities above 0.10 in the test corpus were anticipated by the training corpus.

This coverage seems respectable, considering that the training corpus was small and that neither statistical nor semantic smoothing was used. More important than coverage, however, is the presence of numerous pairs for which good co-occurrence predictions were obtained. Such predictions differ from those made using n-grams in that they need not be chained, and thus need not cover the input to be useful: if consistently good co-occurrence predictions can be recognized, they can be exploited selectively.

The figures obtained for cats and cat co-occurrences are comparable.

8.4 Possible Applications

A weighted co-occurrence between morphemes or lexemes can be viewed as an association between these items; so the set of co-occurrences which CO-OC discovers can be viewed as an associative or semantic network. Spreading activation within such networks is often proposed as a method of lexical disambiguation. (For example, if the concept MONEY has been observed, then the lexical item *bank* has the meaning closest to MONEY in the network: "savings institution" rather than "edge of river", etc.) Thus disambiguation becomes a second possible application of CO-OC's results, beyond the abovementioned primary use for constraining speech recognition. (See (Schütze 1998) or (Veling and van der Weerd 1999) concerning the use of co-occurrence networks for disambiguation, though without comparable segmentation or semantic smoothing.)

A third possible use is in the discovery of topic transitions: we can hypothesize that a span within a dialogue where few co-occurrence predictions are fulfilled is a topic boundary. (Compare e.g. (Morris and Hirst 1991), (Hearst 1994), (Nomoto and Nitta 1994), or (Kozima and Furugori 1994).) Once the new topic is determined, appropriate constraints can be exploited, e.g. by selecting a relevant sub-grammar.

9 Translation Mismatches

During translation, when the source and target expressions contain differing amounts of information, a *translation mismatch* is said to occur. For example, the English sentence *He ate* may be translated by Japanese *tabemashita*. In this case, because the explicit pronoun is suppressed, information concerning person and number is lost. Similarly, *He bought the books* may be translated as *hon wo kaimashita*. Here, the pronoun is once again suppressed, and information about the object of the verb is lost as well: Japanese does not express either its number or its determinateness.

Suppressing such information during translation is less difficult than arranging for its addition when translating in the opposite direction. When translating from Japanese to English, for instance, how is a program to determine whether an entity is determinate, or plural, or third-person? Of course, such problems are not unique to speech translation—they are equally present in text translation. In spoken translation, though, there is the added difficulty of resolving them in real time.

The first observation we can make about mismatch resolution is that it is in some respects akin to ambiguity resolution. In both cases, information is missing which must be supplied somehow: in translation mismatches, missing information must be filled in; in ambiguity resolution, missing information must guide a choice. In light of this similarity, the interactive resolution techniques suggested above for ambiguity resolution can be suggested for mismatches as well. For example, it would be relatively straightforward to put up a menu offering a choice between singular and plural—or "one vs. many", etc. Granted, other sorts of information, for example concerning determinateness, would be trickier to elicit in non-technical terms. (One possible formulation: "Can the audience easily identify which one is meant?" See again (Boitet 1996a) for discussion.) Of course, handling many such requests would be tedious, so interface design would be crucial. And again, the hope is that the need for interaction will shrink as knowledge source integration advances.

A second observation about mismatch resolution is that, when missing information cannot be accurately computed and is excessively burdensome for users to supply, it can simply be left missing. For example, for translating *hon wo kaimashita* when the correct English would be *He bought the books*, the incomplete translation **bought book* could be produced. This broken English would at least allow the hearer to infer the correct meaning from context, more or less as a hearer of the original Japanese (or a hearer of "real" broken English) would have to do. Further,

supplying insufficient information is usually better than supplying incorrect information: in the same situation, **bought book* would be far less confusing than, say, *I bought a book*. (wrong on several counts). Thus far, however, I am aware of no speech translation programs which purposely abstain when in doubt.

Ideally, however, translation software will do its best to resolve mismatches before requesting help from the user or throwing in the towel. Researchers in this area have tended to create programs focusing on a specific sort of mismatch. For example, (Murata and Nagao 1993) propose an expert system for supplying number and definiteness information, and thus articles, during Japanese-English translation.

In a similar spirit, (Seligman 1994b) describes a program for resolving the references of zero pronouns in the ASURA speech translation system (Morimoto et al 1993), thus supplying the missing pronouns for translation. The program, based upon the theory of centering (Sidner 1979; Grosz et al 1983; Grosz et al 1986; Joshi and Weinstein 1981; Walker et al 1990; Takeda and Doi 1994), follows unpublished work by Masaaki Nagata. It is invoked from within specially-modified transfer rules for verbs, and can work alongside other pronoun resolution techniques, e.g. those making use of Japanese honorific information (Dohsaka 1990). No evaluations have yet been made.

Other mismatch problems to be addressed are surveyed in (Seligman et al 1993) in the context of Japanese-English or Japanese-German transfer. These include the determination of tense (Japanese, for example, does not have an explicit future tense); aspect (Japanese lacks explicit cues which would license a choice between *He is studying* and *He has been studying*); intimacy (as required for a choice between German *du* and *Sie*—Japanese does supply a great deal of information concerning politeness, formality, relative status, etc., but none of these map cleanly into the German distinction); choice of possessive determiners (Japanese often uses only namae, or "name", where English would employ *your name*); and several other sorts of mismatch.

Conclusions

The first section of the paper described a "low road" or "quick and dirty" approach to speech translation, in which interactive disambiguation of speech recognition and translation is temporarily substituted for system integration. This approach, I believe, is likely to yield broad-coverage systems with usable quality sooner than approaches which aim for maximally automatic operation based upon tight integration of knowledge sources and components.

Two demonstrations of "quick and dirty" speech translation over the Internet were reported. For the demos, an experimental chat translation system created by CompuServe, Inc. was provided with front and back ends, using commercial dictation products for speech input and commercial speech synthesis engines for speech output. The dictation products' standard interfaces were used to interactively debug dictation results. While evaluation of these experiments remained informal, coverage was much broader than in most ST experiments to date—in the tens of thousands of words. While interactive control of translation was lacking, output quality was probably sufficient for many social exchanges.

But while the "low road" may offer the fastest route to usable broad-coverage speech translation systems, automatic operation based upon knowledge source integration is certain to remain desirable in the longer run. Hence the balance of the paper has concentrated on aspects of integrated systems.

Taken together, the nine areas of research examined in the paper suggest a nine-item wish list for an experimental speech translation system. (1) The system would include facilities for interactive disambiguation of both speech and translation candidates. (2) Its architecture would allow modular reconfiguration and global coordination of components. (3) It would employ a perspicuous set of datastructures for tracking information from multiple processes: stages of translation, multiple tracks, and height, span, or dominance of nodes would be clearly distinguished. (4) The system would employ a grammar whose terminals were phones, recognizing both words and syntactic structures in a uniform and integrated manner, e.g. via island-driven chart parsing. (5) Natural pauses and other aspects of prosody would be used to segment utterances and otherwise aid analysis. (6) Similarity-based techniques for resolving ambiguities, comparable to those of example-based MT, would be effectively used. Stages of translation yielding potential ambiguities would be kept distinct; similarity would be measured along several dimensions (e.g. syntactic and phonological in addition to semantic); top-down as well as bottom-up constraints would be exercised; and disambiguation using both probability-based and similarity-based techniques would be used in complementary fashion. (7) Speech or dialogue acts would be defined in terms of their cue patterns, and analyses based upon

them would be exploited for speech recognition and analysis. (8) Semantically smoothed tracking of lexical co-occurrences would provide a network of associations useful for speech recognition, lexical disambiguation, and topic boundary recognition. And finally, (9) a suite of specialized programs would help to resolve translation mismatches, for instance to supply referents for zero pronouns.

Acknowledgements

Warmest appreciation to CompuServe, Inc. for making the chat-based speech translation demonstrations possible. In particular, thanks are due to Mary Flanagan, then Manager, Advanced Technologies, and to Sophie Toole, then Supervisor, Language Support. Ms. Flanagan authorized and oversaw both demos. Ms. Toole organized and conducted the Grenoble demo and played an active role in making the speech recognition and speech synthesis software operational. Thanks also to Phil Jensen and Doug Chinnock, translation system engineers. The demos made use of pre-existing proprietary software.

Work on all nine of the issues discussed here began at ATR Interpreting Telecommunications Laboratories in Kyoto, Japan. I am very grateful for the support and stimulation I received there.

Thanks also to numerous colleagues at GETA (Groupe d'Étude pour la Traduction Automatique) at the Université Joseph Fourier in Grenoble, France; and at DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) in Saarbrücken, Germany.

However, the opinions expressed throughout are mine alone.

References

Aberdeen, J., S. Bayer, C. Caskey, L. Damianos, A. Goldschen, L. Hirschman, D. Loehr, and H. Trappe. 1996. "Implementing Practical Dialogue Systems With the DARPA Communicator Architecture." In *Proceedings of IJCAI-99, Workshop NLP-2, Knowledge and Reasoning in Practical Dialogue Systems*, Stockholm, Sweden, July 31 - August 6, 1999.

Alexandersson, J., N. Reithinger, and E. Maier. 1997. "Insights into the Dialogue Processing of VERBMOBIL." In *Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLP '97*, pages 33-40, Washington, DC, 1997.

Barnett, J., K. Knight, I. Mani, and E. Rich. 1990. "Knowledge and Natural Language Processing." *Communications of the ACM*, Vol. 33, No. 8 (Aug. 1990), pages 50-71.

Blanchon, H. 1996. "A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France, August 11 - 15, 1996.

Black, A. W. and N. Campbell. 1995. "Predicting the Intonation of Discourse Segments from Examples in Dialogue Speech." In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems*, Vigsø, Denmark, 1995.

Black, E., R. Garside, and G. Leech. 1993. *Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach*. Language and Computers: Studies in Practical Linguistics No. 8. Ropodi. Amsterdam, Atlanta GA.

Boitet, C., ed. 1996a. *Proceedings of the International Seminar on Multimodal Interactive Disambiguation (MIDDIM-96)*, Col de Porte, France, August 11 - 15, 1996.

Boitet, C.. 1996b. "Dialogue-based Machine Translation for Monolinguals and Future Self-explaining Documents." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France, August 11 - 15, 1996.

Boitet, C. and M. Seligman. 1994. "The 'Whiteboard' Architecture: A Way to Integrate Heterogeneous Components of NLP Systems." In *Proceedings of COLING-94*, Kyoto, Aug. 5 - 9, 1994.

Brown, R. 1996. "Example-based Machine Translation in the Pangloss System." In *Proceedings of COLING-96*, pages 169-174, Copenhagen, August 1996.

Dohsaka, K. 1990. "Identifying the Referents of Zero-Pronouns in Japanese Based on Pragmatic Constraint Interpretation." In *Proceedings of ECAI90*, pages 240-245.

Duff, D. 1999. Discourse Resource Initiative home page. *http://www.georgetown.edu/luperfoy/ Discourse-Treebank/dri-home.html*

Erman, L.D. and V.R. Lesser. 1980. "The Hearsay-II Speech Understanding System: A Tutorial." In *Trends in Speech Recognition*, W.A. Lea, ed., Prentice-Hall, pages 361-381.

Fano, R. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA.

Fillmore, C., P. Kay, and M.C. O'Connor. 1988. "Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone." *Language* 64, pages 501-538.

Flanagan, M. 1997. "Machine Translation of Interactive Texts." In *Proceedings of Machine Translation Summit VI*, San Diego, CA, November, 1997.

Frederking, R., A. Rudnicky, and C. Hogan. 1997. "Interactive Speech Translation in the DIPLOMAT Project." In *Workshop on Spoken Language Translation, 35th Meeting of the Association for Computational Linguistics*, ACL-97, Madrid, Spain, July 7-12,1997.

Furukawa, R., F. Yato, and K. Loken-Kim. 1993. *Analysis of Telephone and Multimedia Dialogues*. Technical Report TR-IT-0020, ATR Interpreting Telecommunications Laboratories, Kyoto. (In Japanese)

Furuse, O. and H. Iida. 1996. "Incremental Translation Using Constituent Boundary Patterns." In *Proceedings of COLING-96*, pages 412-417, August 1996, Copenhagen.

Görz, G., M. Kesseler, J. Spilker, and H. Weber. 1996. "Research on Architectures for Integrated Speech/Language Systems in VERBMOBIL." In *Proceedings of COLING-96*, Copenhagen, August 1996.

Grosz, B., A. Joshi, and S. Weinstein. 1983. "Providing a Unified Account of Definite Noun Phrases in Discourse." In *Proceedings of the 21st Annual Meeting of the ACL*, pages 44-50, Cambridge, MA.

Grosz, B., A. Joshi, and S. Weinstein. 1986. *Towards a Computational Theory of Discourse Interpretation*. Unpublished Manuscript.

Hearst, M. 1994. "Multi-Paragraph Segmentation of Expository Text." In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, June 27-30, 1994.

Hosaka, J. and M. Seligman, H. Singer. 1994. "Pause as a Phrase Demarcator for Speech and Language Processing." In *Proceedings of COLING-94*, Kyoto, Aug. 5-9, 1994.

Ichikawa, A., M. Araki, Y. Horiuchi, M. Ishizaki, S. Itabashi, T. Itoh, H. Kashioka, K. Kato, H. Kikuchi, H. Koiso, T. Kumagai, A. Kurematsu, K. Maekawa, S. Nakazato, M. Tamoto, S. Tutiya, Y. Yamashita, and T. Yoshimura. "Evaluation of Annotation Schemes for Japanese Discourse." In *Proceedings of the Workshop: Towards Standards and Tools for Discourse Tagging*, ACL-99, page 26, College Park, MD, June 21, 1999.

Iida, H., E. Sumita, and O. Furuse. 1996. "Spoken Language Translation Method Using Examples." In *Proceedings of COLING-96*, pages 1074-1077, Copenhagen, August 1996.

Iwadera, T., M. Ishizaki, and T. Morimoto. 1995. "Recognizing an Interactional Structure and Topics of Task-oriented Dialogues." In *Proceedings of ESCA Research Workshop on Spoken Dialogue Systems*, Vigsø, Denmark, 1995.

Jokinen, K. and H. Tanaka. 1998. "Context Management with Topics for Spoken Dialogue Systems." In *Proceedings of COLING-ACL 1998*, pages 631-637, Montreal, Quebec, Canada, August 10-14, 1998.

Joshi, A. and S. Weinstein. 1981. "Control of Inference: Role of Some Aspects of Discourse Structure—Centering." In *Proceedings of IJCAI-81*, pages 385-387.

Julia, L., L. Neumeyer, M. Charafeddine, A. Cheyer, and J. Dowding. 1997. "HTTP:// WWW.SPEECH. SRI.COM/ DEMOS/ATIS.HTML." In *Working Notes, Natural Language Processing for the World Wide Web*. AAAI-97 Spring Symposium, Stanford University, March 24-26, 1997.

Jurafsky, D. 1993. *A Cognitive Model of Sentence Interpretation: the Construction Grammar Approach*. Technical Report TR-93-077. International Computer Science Institute and Department of Linguistics, University of California at Berkeley.

Kay, P. 1990. "Even." *Linguistics and Philosophy* 13, pages 59-216.

Knott, A. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.

Knott, A. and R. Dale. 1992. "Using Linguistic Phenomena To Motivate A Set Of Rhetorical Relations." Technical Report Rp-34, Human Communication Research Centre, University Of Edinburgh. Also in *Discourse Processes* 18(1) 1995, pages 35-62.

Kozima, H. and T. Furugori. 1994. "Segmenting Narrative Text into Coherent Scenes." *Literary and Linguistic Computing*, Volume 9, Number 1.

Kompe, R., A. Kiessling, H. Niemann, E. Noeth, A. Batliner, S. Schachtl, R. Ruland, and H.U. Block. 1997. "Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 811-814, Munich, April 1997, IEEE Computer Society Press.

Kowalski, P., B. Rosenberg, and J. Krause. 1995. *Information Transcript*. Biennale de Lyon d'Art Contemporain. December 20, 1995 to February 18, 1996. Lyon, France.

Lenat, D. and R.V. Guha. 1990. *Building Large Knowledge-based Systems*. Reading, MA: Addison-Wesley, 1990.

Loken-Kim, K., F. Yato, K. Kurihara, L. Fais, and R. Furukawa. 1993. *EMMI-ATR Environment for Multi-modal Interaction*. Technical Report TR-IT-0018, ATR Interpreting Telecommunications Laboratories, Kyoto. (In Japanese).

Morimoto, T., T. Takezawa, F. Yato, et al. 1993. "ATR's Speech Translation System: ASURA." In *Proceedings of Eurospeech-93*, Vol. 2, pages 1291-1294.

Morris, J. and G. Hirst. 1991. "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text." *Computational Linguistics*, 17, pages 21-48.

Murata, M. and M. Nagao. 1993. "Determination of Referential Property and Number of Nouns in Japanese Sentences for Machine Translation into English." In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-93)*, Kyoto, Japan, July 14-16, 1993, pages 218-225.

Nadas, A. 1985. "On Turings's Formula for Word Probabilities." *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-33, pages 1414-1416, December, 1985.

Nagao, M. 1984. "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle." In *Artificial and Human Intelligence*, North-Holland, pages 173-180.

Nagata, M. and T. Morimoto. 1993. "An Experimental Statistical Dialogue Model to Predict the Speech Act Type of the Next Utterance." In *Proceedings of the International Symposium on Spoken Dialogue (ISSD-93)*, pages 83-86, Waseda University, Tokyo, November 10-12, 1993.

Nomoto, T. and Y. Nitta. 1994. "A Grammatico-statistical Approach to Discourse Partitioning." In *Proceedings of COLING-94*, Aug. 5-9, 1994, Kyoto, Japan.

Ohno, S. and M. Hamanish. 1981. *Kadokawa Ruigo Shinjiten (Kadokawa New Word Category Dictionary)*. Kadokawa Shoten. January 30, 1981.

Pyra, M. 1995. *Using Internet Relay Chat*. Indianapolis, IN: Que Corporation, 1995.

Reithinger, N. 1995. "Some Experiments in Speech Act Prediction." In *Working Notes, AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*. Stanford University, March 27-29, 1995.

Reithinger, N. and M. Klesen. 1997. "Dialogue Act Classification Using Language Models." In *Proceedings of EuroSpeech-97*, pages 2235-2238, Rhodes, 1997.

Sato, S. 1991. *Example-based Machine Translation*. Doctoral thesis, Kyoto University.

Schütze, H. 1998. "Automatic Word Sense Discrimination." *Computational Linguistics*, volume 24, number 1, pages 97-124.

Searle, J. 1969. *Speech Acts*. Cambridge: Cambridge University Press, 1969.

Seligman, M. 1991. *Generating Discourses from Networks Using an Inheritance-Based Grammar*. Dissertation, Department of Linguistics, University of California, Berkeley.

Seligman, M. 1994a. *CO-OC: Semi-automatic Production of Resources for Tracking Morphological and Semantic Co-occurrences in Spontaneous Dialogues*. Technical Report TR-IT-0084, ATR Interpreting Telecommunications Laboratories, Kyoto.

Seligman, M. 1994b. *CNTR: Basic Functions for Centering Experiments with ASURA*. ATR Technical Report TR-I-0085.

Seligman, M. 1997. "Interactive Real-time Translation via the Internet." In *Working Notes, Natural Language Processing for the World Wide Web*. AAAI-97 Spring Symposium, Stanford University. March 24-26, 1997.

Seligman, M. and C. Boitet. 1994. "A 'Whiteboard' Architecture for Automatic Speech Translation." In *Proceedings of the International Symposium on Spoken Dialogue*, ISSD-93, Waseda University, Tokyo, Nov. 10 - 12, 1993.

Seligman, M., C. Boitet, and B. Meddeb-Hamrouni. 1998a. "Transforming Lattices into Non-deterministic Automata with Optional Null Arcs." In *Proceedings of COLING-ACL 98*, Montreal, Canada, August 10-14, 1998.

Seligman, M., J. Alexandersson, and K. Jokinen. 1999. "Tracking Morphological and Semantic Co-occurrences in Spontaneous Dialogues." In *Proceedings of IJCAI-99, Workshop NLP-2, Knowledge and Reasoning in Practical Dialogue Systems*, Stockholm, Sweden, July 31 - August 6, 1999.

Seligman, M., J. Hosaka, and H. Singer. 1996. "'Pause Units' and Analysis of Spontaneous Japanese Dialogues: Preliminary Studies." In *Notes of the ECAI-96 Workshop on Dialogue Processing in Spoken Language Systems*, August 12, 1996, Budapest, Hungary. Republished in *Lecture Notes in Artificial Intelligence* 1236, E. Maier, M. Mast, and S. LuperFoy, eds., Springer, 1997.

Seligman, M., L. Fais, and M. Tomokiyo. 1995. *A Bilingual Set of Communicative Act Labels for Spontaneous Dialogues*. Technical Report TR-IT-0081, ATR Interpreting Telecommunications Laboratories, Kyoto.

Seligman, M., M. Flanagan, and S. Toole. 1998b. "Dictated Input for Broad-coverage Speech Translation." In *Association for Machine Translation in the Americas (AMTA-98), Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component*. Langhorne, PA, October 28, 1998.

Seligman, M., M. Suzuki, and T. Morimoto. 1993. "Semantic-Level Transfer in Japanese-German Speech Translation: Some Experiences." Technical Report NLC93-13 of the Institute of Electronics, Information, and Communication Engineers (IEICE). May 21, 1993.

Sidner, C. 1979. *Toward a Computational Theory of Definite Anaphora Comprehension in English*. Technical Report AI-TR-537, MIT.

Sobashima, Y. and H. Iida. 1995. "A Multi-dimensional Analogy-based, Context-dependent, Bottom-up Parsing Method for Spoken Dialogues." In *Proceedings of NLPRS (Natural Language Processing Pacific Rim Symposium) 1995*, Vol. 2, pages 586-591.

Sobashima, Y. and M. Seligman. 1994. "Parsing Method for Example-based Analysis Integrating Multiple Knowl-

edge Sources." In *Proceedings of the 49th General Meeting of the Information Processing Society of Japan*.

Stock, O., R. Falcone and P. Insinnamo. 1989. "Bi-directional Charts: A Potential Technique For Parsing Spoken Natural Language Sentences." *Computer Science and Language* (1989) 3, 219-237.

Sumita, E. and H. Iida. 1992. "Example-based Transfer of Adnominal Particles into English." *IEICE Trans. Inf. Syst.*, Vol. E75-D, No. 4, pages 585-594.

Takeda, S. and N. Doi. 1994. "Centering in Japanese: a Step Towards Better Interpretation of Pronouns and Zero-Pronouns." In *Proceedings of COLING-94*, Kyoto, Japan, Aug. 5-9, 1994.

Takezawa, T., F. Sugaya, and A. Yokoo. 1999. "ATR-MATRIX: A spontaneous speech translation system between English and Japanese." In *ATR Journal*, 2:29-33, June 1999.

Veling, A. and P. van der Weerd. 1999. "Conceptual Grouping in Word Co-occurrence Networks." In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99)*, volume 2, pages 694-699, Stockholm, Sweden, July 31-August 6, 1999.

Wahlster, W. 1993. *VERBMOBIL: Translation of Face-to-Face Dialogs*. Research Report RR-93-34, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany, 1993.

Walker, M., ed. 1999. *Proceedings of the Workshop, Towards Standards and Tools for Discourse Tagging*, ACL-99, College Park, MD, June 21, 1999.

Walker, M., M. Iida, and S. Cote. 1990. "Centering in Japanese Discourse." In *Proceedings of COLING-90*, page 1, Helsinki.

Xwaves93. 1993. Entropic Research Laboratory, 1993.

Zajac, R. and M. Casper. 1997. "The Temple Web Translator." In *Working Notes, Natural Language Processing for the World Wide Web*. AAAI-97 Spring Symposium, Stanford University, March 24-26, 1997.