Evaluating Commercial Spoken Language Translation Software

Harold SOMERS and Yuri SUGITA[1]

Centre for Computational Linguistics
UMIST, PO Box 88
Manchester M60 1QD, England
Harold.Somers@umist.ac.uk, sugita_yuri@yahoo.co.jp

Abstract

While spoken language translation remains a research goal, a crude form of it is widely available commercially for Japanese–English as a pipeline concatenation of speech-to-text recognition (SR), text-to-text translation (MT) and text-to-speech synthesis (SS). This paper proposes and illustrates an evaluation methodology for this noisy channel which tries to quantify the relative amount of degradation in translation quality due to each of the contributing modules. A small pilot experiment involving word-accuracy rate for the SR, and a fidelity evaluation for the MT and SS modules is proposed in which subjects are asked to paraphrase translated and/or synthesised sentences from a tourist's phrasebook. Results show (as expected) that MT is the "noisiest" channel, with SS contributing least noise. The concatenation of the three channels is worse than could be predicted from the performance of each as individual tasks.

1. Introduction

Evaluation is without doubt a major aspect of language engineering, including Machine Translation (MT). Although it is still true that no consensus exists regarding the best way to evaluate software, there is general agreement about some of the factors that must be taken into account when deciding what form an evaluation should take. MT evaluation has been much studied in recent years, so much so that it has been light-heartedly claimed that MT evaluation "is a better founded subject than machine translation" (Wilks, 1994:1). If this is no longer strictly true, it is because MT is arguably in pretty good shape, at least text-to-text MT of restricted texts or for restricted purposes. This paper however concerns a much less mature application, namely spoken language translation (SLT).

Until recently thought to be simply too difficult a task (cf. Krauwer, 2000:1), SLT has now established itself as a growing area for research and development. The apparently attractive option of hooking up a text MT system to speech recognition (SR) at one end and speech synthesis (SS) at the other – so-called linear pipeline architecture (cf. Seligman, 2000:156) – is rejected in all experimental SLT systems in recognition of the fact that spoken language is fundamentally different from written language. In the commercial world, however, things are different. In Japan especially, where text MT systems are widely used and have become a familiar application included as a free add-on with most computers, and where both speech recognition and speech synthesis have reached high levels of quality, the SR–MT–SS chain has proven irresistible.

The present paper proposes a methodology for evaluating SLT of this type: considering that concatenating three potentially error-prone processes is bound to provide a triple noisy channel, the aim is to establish how noisy each channel is, or, to put it another way, the relative negative impact of each of the processes on the overall SLT task. The methodology is illustrated via a small-scale evaluation of a commercial Japanese–English SLT system.

2. Background

As stated in the Introduction, many Japanese computers now come with SR, MT and SS already installed, or else cheaply available; in many cases, users are encouraged to attempt SLT in that the MT system offers the possibility of both input and output in either spoken or written form. Language pairs are almost inevitably Japanese and English, and one or other of the speech elements may be restricted to Japanese, but it is not unusual to find speech input and output as options in the MT window.

[1] Currently in Tsushi, Mieken, Japan.

One such system is Sourcenext's *Honkaku Hon'yaku*, currently one of the best-selling Japanese MT systems,[2] at the very modest price of ¥9800 (less than $100). The MT system itself was developed by NEC, and is described in the accompanying literature as a combination of rule-based and example-based methods. Its dictionary contains 526,000 words for basic use, with the option to purchase additional technical dictionaries available in 31 domains, increasing the vocabulary by 1.2m words. The Japanese SR function is SmartVoice, also developed by NEC. This system performs best when it has been trained to recognize an individual user, which is achieved with a 150-sentence training set. The English SS engine is the TTS system developed by Lernout & Hauspie.

The three functions can be combined to give four translation modes, namely

1. SR–MT–SS speech-to-speech
2. SR–MT speech-to-text
3. MT–SS text-to-speech
4. MT text-to-text

In the following sections, we shall briefly discuss the three functions, and some issues in separately evaluating them.

2.1 Speech Recognition

In SR an important distinction is made between speaker-independent and speaker-dependent systems, with the latter generally performing considerably better. Training permits the system to accustom itself to idiosyncrasies of both voice quality and allophonic realisation. However, even if individual phonemes are recognized with a high level of accuracy, there is still be the problem of homophone disambiguation. For Japanese, this problem is particularly acute. The simple phonological system (five vowels, 15 consonants, mostly open syllables with almost no consonant clusters) makes phoneme recognition fairly robust, but the high incidence of homophones makes the word-selection task extremely difficult. The same problem arises in text input (in word-processing for example), and Japanese speakers are accustomed to having to select the correct written form from a pop-up menu often showing 10 or more alternatives (see Fig. 1). SmartVoice is typical in requiring the user to confirm the word

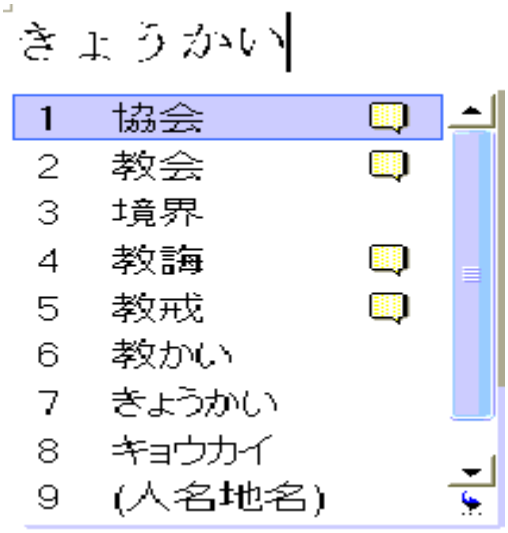

Figure 1. Pop-up menu offering the user a choice amongst competing orthographic representations of the given input, in this case *kyōkai*.

selection at the time of input. As we shall see, this feature has an important effect on our evaluation methodology.

Where the output from SR is text, there are well-established evaluation measures, notably word-error rate (WER) and sentence-error rate (SER), in either case given as an average percentage over a set of samples. As its name suggests, WER is the number of incorrect, omitted or inserted words divided by the number of words in the target phrase. The SER is a simpler percentage reflecting right/wrong decisions and tends to be pessimistic (Tillmann et al., 2000:55), since a single error in a long sentence will cause the whole sentence to be rejected. Since for most SR tasks, overall comprehension can usually survive a few small errors, WER is to be preferred, sometimes with refinements to take account of the importance or otherwise of the misrecognised word(s).

One problem for the WER in our case however is that written Japanese does not indicate word boundaries. Instead, we split the text stream into morphemes, using the Japanese morphological analyser ChaSen (Matsumoto et al., 2000) available free online. Morpheme error rate (MER) is slightly more punitive than WER, but this is more than compensated by the fact that the SR task is massively aided by the human user choosing amongst alternatives. In particular, we wish to

[2] www.computernews.com/award/2003_s.htm (in Japanese).

assess the degree to which SR errors degrade the overall translation result.

2.2 Machine Translation

Many distinctions are made in discussions of MT evaluation. Our evaluation is a "declarative" evaluation (Arnold et al., 1993:7) aimed at end-users in that we aim to evaluate the output rather than the process. Such MT evaluations focus on grammaticality and style, intelligibility, fidelity and so on. In choosing the form of our MT evaluation, we took into account the most likely use for an SLT system. Almost without exception, cutting-edge SLT research systems focus on task-oriented cooperative dialogues, usually between monolingual participants. Accordingly, our MT evaluation aims at testing the fidelity in translation of short utterances taken from a tourist's phrasebook. This is of course similar to the "Traveller Task" defined in the EUTRANS project (Amengual et al., 1997). "Fidelity" is measured in terms of a subject's ability to infer correctly the intended meaning of the utterance. We are not interested in the grammar or style of the output. As above, we want to know to what extent the MT system is responsible for the overall quality of the SLT.

Taking phrases from a tourist's phrase book helps us avoid some of the well-documented difficulties in translating real-time dialogues, such as hesitations, false starts, repetitions, fragmental phrases, complex topicalization, metonymical phrases, inconsistent expressions, and so on. Other features, such as ellipses, anaphora, idiomatic expressions for etiquette, may still be present to some extent, as will problems common to both SLT and text MT – lexical and syntactic ambiguity above all else.

2.3 Speech synthesis

Although SS is considerably more advanced than either SR or MT, there are still pitfalls, especially in text-to-speech synthesis. For many languages the mapping of orthography to speech is relatively straightforward, inasmuch as a word's pronunciation can be given more or less unequivocally in a dictionary. Exceptions to this general rule are heteronyms such as *tear*, *read*, and so on. However, generally this process does not present a major obstacle, in particular for a language like English which has been studied substantially. Other aspects less well studied include prosodic features such as pitch, loudness and duration, all of which can affect the perceived meaning, as well as contributing to the naturalness of the synthetic speech. In some respects, however, SS is considered "a solved problem" (Kay et al., 1994:140). In fact, for a wide range of applications, for a number of languages imperfect but nonetheless acceptable synthesis is currently available (*idem*.).

Like MT, SS can be evaluated for its accuracy, intelligibility or style (cf. van Heuven and van Bezooijen, 1995). As with our MT evaluation, our aim is to assess the extent to which the synthesised speech permits the hearer to infer correctly the intended meaning of the utterance. Once again, our aim is to quantify the contribution of the SR component to the SLT.

3. Method

Our goal is to take a complex process consisting of three elements and to compare the contribution of each of the three to the overall process. It is therefore logical to try to evaluate each of the processes in isolation, as well as all (logical) combinations of them. We therefore conducted six different but closely related evaluations, as follows:

1. speech recognition (SR)
2. speech synthesis (SS)
3. text-to-text translation (MT)
4. speech-to-text translation (SR+MT)
5. text-to-speech translation (MT+SS)
6. speech-to-speech translation (SR+MT+SS)

Figure 2 shows more clearly the relationship between these six "modes", in particular the idea that mode 6 is in some sense a combination of 1+2+3, or 4+2, or 5+1. Another way to look at it is

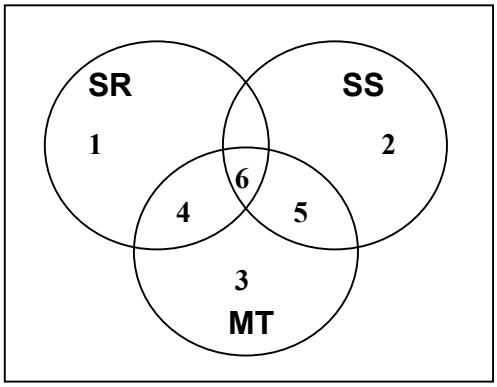

Figure 2. Three modes – six tests

that mode 4 assumes "perfect" SS and so on. It will be of interest to see to what extent the results bear out this notion of the whole as the sum of its parts.[3]

3.1 The translation task

As mentioned above, all the evaluations have in common the specific SLT task based on the scenario of a tourist using the SLT system as a speaking phrase-book. We selected 30 phrases from the "tourist" side of a Japanese–English phrase-book *English: Wagamama Aruki Travel Conversation Series 1* (i.e. phrases the tourist would use, not replies from the informant). The 30 phrases were grouped in six sets of five, each related to a specific scene, e.g. reporting a theft, getting directions, at the bank. One can also group the questions on syntactic criteria: nine were *wh*-questions, nine *yes–no* questions, seven were requests and the remaining five declarative statements. In the fidelity evaluations, judges are asked to paraphrase what they understand to be the speaker's *intention* (more about this below). The choice of phrases was not entirely random. Care was taken to choose different types of phrases (e.g. *How do I get to the theatre?* and *Where is the police station?* are functionally almost identical), and to avoid culturally loaded phrases which our subjects might not understand (e.g. *Is there a major league game tonight? Can I take the subway?*). However, *no* consideration was made of the likely ease or difficulty for SR, translation or SS each phrase might involve. The selection is of course crucial, especially for the phase of judging the results, which we discuss below.

3.2 Evaluating SR

The evaluation of SR on its own (Test 1) differs from the other evaluations in several respects. The most important one is that it is not necessary to make a subjective judgment, since the output of the speech-to-text conversion can easily be compared with the "gold standard" of the target text. As mentioned above, because written Japanese does not indicate word boundaries, we adapted the standard WER measure to count morpheme errors. We used the Japanese morphological analyser ChaSen to segment the phrases, and checked the resulting segmentation for errors along the method adopted in Verbmobil (cf. Waibel, et al., 2000). We also applied the more punitive SER evaluation.

The SR system used in this evaluation depends on prior training ("Speaker registration") of the system which the experimenter (co-author YS) had previously done. In the case of SmartVoice, this is achieved by reading out 150 test sentences, a relatively quick procedure.

It is well known that voice quality is subject to change due to fatigue, excitement, health and other factors. We decided to perform multiple evaluations of the SR under differing conditions, and running through the 30 phrases in a randomised order. Nine tests in all were made (three each at three different times of day). The results are presented below.

The multiple tests served a second purpose: in the normal use of the SR system either for dictation purposes, or together with the MT system, the user is invited to confirm the accuracy of the text output, to edit it, or to repeat the input. This feature is obviously significant in our evaluation of an SLT chain involving SR, since in real use a misrecognised input will be corrected before it is passed on to the next stage in the chain. For our experiments involving SR+MT+… we always used the "best" SR result obtained for any of the inputs: thus if any of the nine trials had resulted in error-free recognition, that result was subsequently used. Otherwise, the result with the lowest MER was used. In any case, for the sake of consistency and convenience, in the experiments involving speech input, the SR element was simulated, and a pre-stored text fed into the MT stage.

3.3 Evaluating SS

In the evaluations involving SS (Tests 2, 5 and 6), the subjects were told that they were going to hear synthetic speech, and the tourist scenario was explained. The subjects were in a relatively quiet room, and the sound was played through extension speakers. The test was run at the subject's pace: each new item was presented only when the subject said they were ready. However, requests to hear an item again were refused.

Subjects were provided with a test script which included general instructions followed by a header for each of the six scenarios (e.g. *You are working behind the counter in a bank*) with space for each

[3] Actually, as we will see, it is a product rather than a sum.

of the five utterances in each section. They were instructed to write in the space an indication of what they thought the tourist was asking. They were explicitly asked not to write down exactly what they heard, but to paraphrase it, and above all to write down something that made sense, even if they had not heard easily. One way they could do this was by using reported speech (e.g. *She's asking about the exchange rate*).

In Test 2, the target English phrases were put through the SS system. In Tests 5 and 6, the English text resulting from translation by the MT system was synthesized.

3.4 Tests 3 and 4: Evaluating text output

In Tests 3 and 4, text output from the MT system was evaluated. Since the subjects were not required to listen to anything, the evaluations could be done by the whole group in one sitting, with minimal supervision. The test scripts were similar to those for the test involving SS, but instead of being asked to listen and write down a paraphrase, the translated texts were presented on the page, and subjects were asked to write underneath what they thought was meant.

3.5 Judging the answers

Apart from Test 1, where the evaluation is mechanical, the tests require a judgment of the match between the subjects' responses and the expected answers. All the scripts were independently marked by two judges: the experimenter (YS) and the co-author (HS). The answers were rated on a seven-point scale as follows:

Useful
A (6) Clearly useful to communicate the intention of the utterance: the response matches what is intended in the original utterance. It contains the same concepts and all the necessary arguments.
A– (5) Generally useful: the response nearly matches what is intended in the original utterance; may misrepresent or omit some detail that is not fatal.
Borderline
B (4) Useful but less informative compared with the above: basic match with what is intended, but some accompanying arguments are incomplete or inadequate.
B– (3) Useful but not wholly adequate: as B but some arguments are missing.
Useless
C (2) Almost useless but still informative and useful: the response doesn't match what is intended but nevertheless contains some partially useful information.
C– (1) Clearly useless: the response doesn't match what is intended in the original utterance at all.
No response
D (0) Blank or garbage.

The judges worked through the 30 test items to agree beforehand which elements were essential or additional information in each. This process was aided by the fact that a small pilot of the experimental design had been run with four subjects, which identified some potential pitfalls.[4] For example, in one item, *I'd like an automatic sports car*, it was agreed that both *automatic* and *sports* must be mentioned for an A. After the initial judging, the results were compared, and cases of discrepancy reconciled by discussion. In this way, we tried to make our marking procedure more like an assessment of "precision" in an information retrieval task (cf. Carter et al.'s (2000) evaluation of their SLT system, also discussed below). In fact, scores were never out by more than one point, and these cases were concentrated on six of the test items, affecting less than one fifth of the results. Probably, the judges' raw scores could have been left intact, with little difference to the overall results.

4. Results

It is appreciated that the small number of subjects for each evaluation diminishes the value of the results. We are more interested in presenting the methodology here, though the results such as they are, reveal some interesting issues.

The scores for all the tests are summarized in Table 1. For Tests 2 to 6, there were 30 phrases, and five subjects, giving a total of 150 test items. The "rate" shown in the bottom row is the "success rate" of the process, calculated by awarding points for each response as shown above (6 for A, 5 for A–, and so on). A maximum score is thus 900 (=150×6). Scores for Test 1 are calculated differently, but included here for completeness. Let us look at the results in more detail.

[4] Some aspects of the experimental design were changed after the pilot, which also helped to identify one or two misleading phrases, e.g. the American term *bill* used in translation where *note* is preferred in British English.

4.1 Test 1: SR

SR was evaluated over nine occasions, with randomised presentation order of the 30 phrases. Using the MER measure, expressed as a measure of accuracy (AR = 100 – ER) the results were remarkably consistent, ranging from 94.8% to 91.4% with an average of 93.2%, indicating that SR is not sensitive to time of day or fatigue. The SAR as expected shows lower scores over a broader range: 53.3% to 70.0%, average 61.9%. The total number of morphemes in the 30 phrases was 233. The MAR indicates on average a single error every other sentence.

Where the results of SR are used as input to MT, we took the best-scoring result for each phrase. The MAR and SAR for this optimised set of phrases are a comfortable 97% and 80% respectively.

4.2 Tests 2 and 3: SS and MT

For Test 2, the seven-point rating scale described above was used. 138 (92%) of the items were rated A, the remaining 12 (8%) spread evenly over the other categories (including one D).

The scores for Test 3 show that just over 70% of the output of the text translation system is "useful", itself an encouraging result.

The bottom row of Table 1 indicates that the three processes when considered in isolation are ranked SR>SS>MT. The fact that SR outperforms SS might be surprising, until one considers that 97.0% is an optimised score (the average was 93.2%, slightly worse than SS), justified by the fact that SR is aided by the user in this set-up.

When we combine the processes, the performance deteriorates further, more or less proportionately.

	Test 1 SR	Test 2 SS		Test 3 MT		Test 4 SR+MT		Test 5 MT+SS		Test 6 SR+MT+SS	
		No.	%	No.	%	No.	%	No.	%	No.	%
A		138	92.0	93	62.0	94	62.7	91	60.7	76	50.7
A–		2	1.3	13	8.7	7	4.7	14	9.3	13	8.7
B		3	2.0	13	8.7	14	9.3	14	9.3	14	9.3
B–		2	1.3	5	3.3	9	6.0	5	3.3	3	2.0
C		3	2.0	9	6.0	5	3.3	9	6.0	12	8.0
C–		1	0.7	13	8.7	8	5.3	15	10.0	30	20.0
D		1	0.7	4	2.7	13	8.7	2	1.3	2	1.3
Rate (%)	97.0	95.6		80.1		77.8		79.4		71.1	

Table 1. Summary of scores in all tests

4.3 Pipelining the processes

Tests 4 to 6 show the results of pipelining or concatenating the processes. As we might have expected, the scores are lower, confirming the prediction that when we take the output of a noisy channel as input to another noisy channel, the result is worse than the lesser of the two channels.

The results of this set of experiments allow us to attempt to quantify the multiplication effects of this chaining process.

We can illustrate this by looking at Test 4 (SR+MT) compared to Tests 1 (SR) and 3 (MT). MT on its own is rated at 80.1% which goes down to 77.8% when it is combined with SR. One can say that SR "degrades" the MT by a factor of 0.971; that is to say, SR+MT is 0.971 of the quality of MT on its own. Curiously, this is almost exactly the same as the reliability score achieved by SR in Test 1.

A similar calculation can be made for SS. Comparing Tests 3 and 5, we can say that SS degrades MT by 0.991, which compares favourably to the Test 2 score for SS of 95.6%.

Finally, we can say that the combined degradation factor of SR and SS on MT is 0.888.

These derived scores are summarized in Table 2. This table confirms the more intuitive result that SS, which is on the whole very robust, hardly degrades MT output at all (by 0.009 in fact), even though the individual success rate for this module is slightly worse than that for SR. The same data are shown graphically in Figure 3 which indicates

that the success rates and the degradation factors are proportional.

	Rate	Degradation
SR	97.0	0.971
SS	95.6	0.991
SR+SS		0.888

Table 2. Success rates and degradation factors

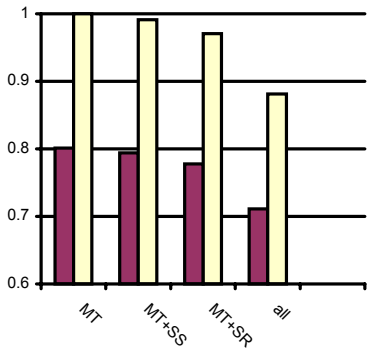

Figure 3. Success rates (the darker shorter bars) and degradation factors shown graphically.

5. Previous Studies and Discussion

For SLT much more than text translation it is the content rather than the form of the message that is important: an ungrammatical but (correctly) understood translation is perfectly acceptable in the communicative scenario for which this type of SLT system is envisaged. As Carter et al. (2000) state:

"Our goal … is to measure objectively the ability of subjects to understand the content of speech output. This must be the key criterion …: if apparent deficiencies in syntax or word choice fail to affect subject's ability to understand content, then it is hard to say that they represent real loss of quality." (p. 300)

Their system handles dialogues in the ATIS domain, so it is reasonable for them to evaluate its precision and recall on information retrieval tasks.

Jain et al. (1993) performed a glass-box evaluation of an early version of the JANUS system. Gates et al. (1997) evaluated a later version in a fairly traditional subjective manner using bilingual judges. Levin et al. (2000) report an evaluation of the same system similar to ours in that they attempt to evaluate separately the effect of the SR and MT functions. Interestingly, Jain et al. found "errors in speech recognition [to be] the primary cause of incorrect translation" (*op. cit.*:159). It is open to speculation whether our contrary finding is due to better SR or inferior MT.

The purpose of this paper is to report our *methodology* for evaluating this form of SLT system much more than the results we happened to get with our small number of subjects. With only five subjects in each mode, it is obvious that one wayward score could completely derail our results. Let us concentrate therefore in this final section on the evaluation methodology itself.

We were relatively happy with the choice of task for our evaluation. The "tourist abroad" scenario seems quite a natural use for this software (cf. Ward, 2002), and subjects for the most part quickly understand their part in the role play. It was felt to be important in both the SS tests, and even more so in the text-output tests, to require the subjects to indicate what they understood in their own words: "communicative intent" is the key notion here, which is why we label our evaluation as one of "fidelity" rather than, say, intelligibility. A bad translation could be clearly synthesised and transcribed verbatim by the subject, but this would give no indication that the communicative intent had been translated.

We can be self-critical about some of the finer points of our experimental design. As we discovered, choice of test items could be quite crucial. Although we filtered out culturally sensitive items, some of the phrases were difficult to understand out of context, even under the most favourable conditions (hence, perhaps, the ten responses scoring B, C or D in Test 2). Interestingly, it could be argued that the gold standard (100% understanding) is an unreasonable target: even two humans face to face might be expected to misunderstand one another from time to time. In this case one strategy is to ask your dialogue partner to repeat themselves, an option that we denied to subjects participating in Tests 2, 5 and 6 (all involving SS output), where a scraping chair or lack of concentration could lead the subject to underperform – another reason to have a much larger subject population.

On the whole however, we would feel confident that our methodology is suitable for replication with a much larger population, for example in a

comparative evaluation of several systems, and look forward to an opportunity to do so.

References

Amengual, J.C., J.M. Benedí, F. Casacuberta, A. Castaño, A. Castellanos, D. Llorens, A. Marzal, F. Prat, E. Vidal and J.M. Vilar (1997) "Using Categories in the EUTRANS System", *Spoken Language Translation: Proceedings of a Workshop Sponsored by the Association for Computational Linguistics and the by the European Network in Language and Speech (ELSNET)*, Madrid, pp. 44–53.

Arnold, Doug, Louisa Sadler and R. Lee Humphreys (1993) "Evaluation: An Assessment", *Machine Translation* **8**, 1–24.

Carter, David, Manny Rayner, Robert Eklund, Catriona MacDermid and Mats Wirén (2000) "Evaluation", in Manny Rayner, David Carter, Pierrette Bouillon, Vassilis Digalakis and Mats Wirén (eds) *The Spoken Language Translator*, Cambridge: Cambridge University Press, pp. 297–312.

Gates, Donna, Alon Lavie, Lori Levin, Marsal Gavaldà, Monika Woszczyna and Puminhg Zhan (1997) "End-to-End Evaluation in JANUS: A Speech-to-Speech Translation System", in E. Maier, M. Mast and S. Luperfoy (eds) *Dialogue Processing in Spoken Language Systems*, Berlin: Springer, pp. 195–206.

Jain, A.N., A.E. McNair, A. Waibel, H. Saito, A.G. Hauptmann and J. Tebelskis (1993) "Connectionist and Symbolic Processing in Speech-to-Speech Translation: The JANUS System", in Sergei Nirenburg (ed.) *Progress in Machine Translation,* Amsterdam: IOS Press and Tokyo: Ohmsha, pp.153–160.

Kay, Martin, Jean Mark Gawron and Peter Norvig (1994) *Verbmobil: A Translation System for Face-to-Face Dialog*, Stanford, CA: CSLI.

Krauwer, Steven (2000) "Introduction: Special Issue on Spoken Language Translation", *Machine Translation* **15**, 1–2.

Levin, Lori, Alon Lavie, Monika Woszczyna, Donna Gates, Marsal Gavaldà, Detlef Koll and Alex Waibel (2000) "The JANUS-III Translation System: Speech-to-Speech Translation in Multiple Domains", *Machine Translation* **15**, 3–25.

Matsumoto, Yuji, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka and Masayuki Asahara (2000) *Morphological Analysis System ChaSen version 2.2.1*, Technical Report, NAIST, Nara, Japan; available at http://chasen.aist-nara.ac.jp/.

Seligman, Mark (2000) "Nine Issues in Speech Translation", *Machine Translation* **15**, 149–185.

Tillmann, Christoph, Stephan Vogel, Hermann Ney and Hassan Sawaf (2000) "Statistical Translation of Text and Speech: First Results with the RWTH System", *Machine Translation* **15**, 43–74.

van Heuven, Vincent and Renie van Bezooijen (1995) "Quality Evaluation of Synthesized Speech", in K.K. Paliwal (ed.) *Speech Coding and Synthesis*, Amsterdam: Elsevier Science, pp. 707–738.

Waibel, Alex, Hagen Soltau, Tanja Schultz, Thomas Schaaf and Florian Metze (2000) "Multilingual Speech recognition", in Wolfgang Wahlster (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer, pp. 33–45.

Ward, Nigel (2002) "Machine Translation in the Mobile and Wearable Age", *Proceedings of the MT Roadmap Workshop at TMI-2002*, Keihanna, Japan, pages not numbered.

Wilks, Yorick (1994) "Keynote: Traditions in the Evaluation of MT", in Muriel Vasconcellos, (ed.) *MT Evaluation: Basis for Future Directions. Proceedings of a workshop sponsored by the National Science Foundation*, San Diego, California, pp. 1–3.