Statistical Machine Translation of European Parliamentary Speeches

David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens and Hermann Ney
Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University
52056 Aachen, Germany
{vilar,matusov,hasan,zens,ney}@cs.rwth-aachen.de

Abstract

In this paper we present the ongoing work at RWTH Aachen University for building a speech-to-speech translation system within the TC-Star project. The corpus we work on consists of parliamentary speeches held in the European Plenary Sessions. To our knowledge, this is the first project that focuses on speech-to-speech translation applied to a real-life task. We describe the statistical approach used in the development of our system and analyze its performance under different conditions: dealing with syntactically correct input, dealing with the exact transcription of speech and dealing with the (noisy) output of an automatic speech recognition system. Experimental results show that our system is able to perform adequately in each of these conditions.

Paper type: (R) Research **Keywords:** Speech Translation, Methodologies for MT, Text and speech corpora for MT, MT evaluation results.

1 Introduction

Speech-to-speech translation is an outstanding reasearch goal in the machine translation community. Up to now, most of the projects dealing with this issue have dealt only with artificial or very limited tasks (Wahlster, 2000; EuTransProject, 2000; Lavie et al., 2001; Ueffing and Ney, 2005). The goal of the TC-Star project is to build a speech-to-speech translation system that can deal with real life data. For this purpose we have collected data from parliamentary speeches held in the European Parliament Plenary Sessions (EPPS) to build an open domain corpus. There are three different versions of the data, the official version of the speeches as available on the web page of the European Parliament, the actual exact transcription of the speeches produced by human transcribers and the output of an automatic speech recognition system. We evaluate our system under these three conditions.

The structure of the paper is as follows: In Section 2 we will describe the statistical approach to machine translation and in Section 3 further methods used in our translation system. The EPPS databases and experimental results will be presented in Section 4. We will draw conclusions in the last Section.

2 Statistical Machine Translation

In a machine translation framework we are given a sentence $f_1^J = f_1 \dots f_J$ in a source language that is to be translated as sentence $e_1^I = e_1 \dots e_I$ into a target language (f and e stand for 'French' and 'English' in the original paper (Brown et al., 1993)). For the statistical approach, we use Bayes decision rule which states that we should choose the sentence that maximizes the posterior probability

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} p(e_1^I|f_1^J) \tag{1}$$

$$= \operatorname*{argmax}_{e_1^I} p(e_1^I)p(f_1^J|e_1^I)\,, \tag{2}$$

where the argmax operator denotes the search process. The transformation from (1) to (2) using Bayes rule allows us to use two sources of information, the translation model $p(f_1^J|e_1^I)$ and the target language model $p(e_1^I)$. The translation model can be further decomposed into a lexicon model, which gives the probability for word translations, and an alignment model, which connects the words in the source and target sentences. Let us consider the HMM Alignment model as presented in (Vogel et al., 1996) in order to illustrate this decomposition. This model decomposes the translation probability as follows:

$$p_\vartheta(f_1^J|e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p_\vartheta(a_j|a_{j-1}, I, J) p_\vartheta(f_j|e_{a_j})]\,, \tag{3}$$

where the term $p_\vartheta(a_j|a_{j-1}, I, J)$ is a first-order model for the alignment, and the term

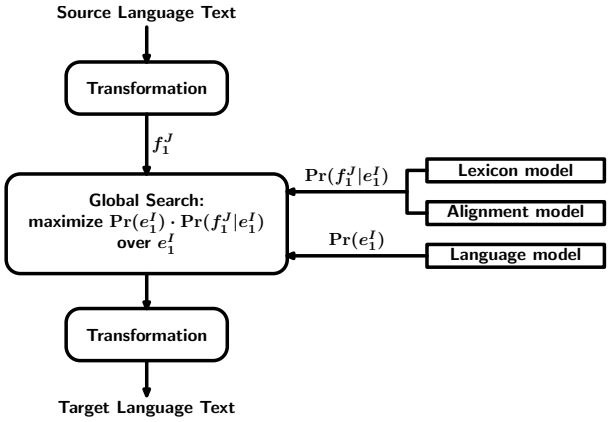

Figure 1: Architecture of the translation approach based on Bayes decision rule.

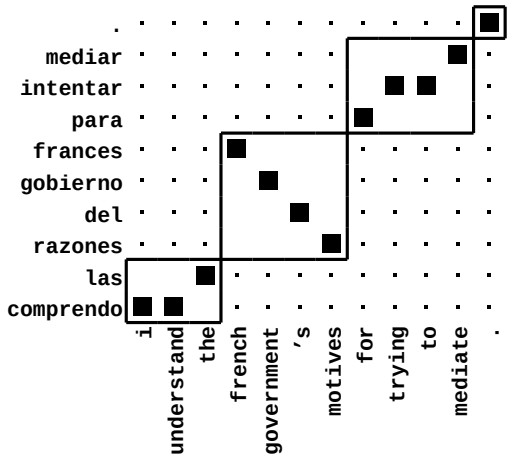

Figure 2: Example of alignment and phrases.

$p_\vartheta(f_j|e_{a_j})$ is the lexicon model[1], both dependent on a set of parameters θ. The architecture of such a system is depicted in Figure 1.

2.1 Training

The parameters of the statistical models are estimated using the available training data, normally using the *maximum likelihood* criterion. The maximum likelihood criterion states that we should choose those set of parameters $\hat{\vartheta}$ that maximizes the term

$$\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmax}} \sum_s \log p_\vartheta(f_{s1}^{J_s}|e_{s1}^{I_s}) \,, \qquad (4)$$

where s is a running index over the training sentence pairs $(f_{s1}^{J_s}, e_{s1}^{I_s})$. That is, we try to maximize the probability that our statistical model gives to the observed training sentences. In the case of the HMM alignment model (3) the set of parameters is composed of the alignment probabilities $p_\vartheta(a_i|a_{i-1}, I, J)$ and the lexicon probabilities $p_\vartheta(f_j|e_{a_j})$. The maximization process is carried out with a modified version of the EM algorithm (Brown et al., 1993).

3 Refined Methods for Machine Translation

3.1 Phrase Based Models

The HMM model discussed in Section 2, as well as the IBM models introduced in (Brown et al., 1993), constitute the so-called *single-word based* models because the lexicon probabilities depend only on one word. One major disadvantage of these approaches is that contextual information is not taken into account. The translation of a given word, however, depends heavily on the surrounding words. In the single-word based translation approach, this disambiguation is addressed by the language model only, which is not often capable of doing this.

One way to incorporate the context into the translation model is to learn translations for whole phrases instead of single words. Here, a phrase is simply a sequence of words, no other linguistic meaning is required. So, the basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations.

First an alignment between source and target sentence is found by using a chain of single-word based models[2] in both directions (source-to-target and target-to-source) and combining the two obtained alignments (Och and Ney, 2003). Given this alignment an extraction of contiguous phrases is carried out and their probabilities are computed by mean of relative frequencies (Zens and Ney, 2004). An example of an alignment between two sentences and a (possible) set of phrases to be extracted is shown in Figure 2.

3.2 Log-linear Model

As an alternative to the traditional source-channel approach given in Equation (2) we can model the translation probability directly using a log-linear model (Och and Ney, 2002):

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)} \,, \qquad (5)$$

[1] For a more exact formulation we should also include a length model $p_\vartheta(J|I)$. For simplicity it is left out in this presentation.

[2] Usually the used models are IBM-1, HMM and IBM-4.

with h_m different models, λ_m scaling factors and the denominator a normalization factor that can be ignored in the maximization process. We choose the λ_m by optimizing a performance measure over a development corpus using the downhill simplex algorithm as presented in (Press et al., 2002). The source-channel model (2) is a special case of (5) with appropriate feature functions. The log-linear model, however, has the advantage that additional models can be easily included. In particular the inclusion of phrase translation probabilities in both directions, additional word based models and heuristics like word and phrase penalty have proven adequate in practice (Zens and Ney, 2004).

3.3 N-best Lists

It is a well known fact that the quality of the output of any current machine translation system is far from being perfect. For efficiency reasons in most tasks, the whole search space can not be treated directly. So some pruning has to be carried out in the search process, which can lead to the rejection of valid translations (so-called search errors). The state-of-the-art algorithms used in current systems, however, allow to minimize these kinds of errors, so the main source of errors still lies in the probability models, i.e. sentences which are better translations do *not* get a better score (a higher probability).

In order to alleviate this effect, we can make use of word graphs and n-best lists (Ueffing et al., 2002). These are representations of different possible translations for a given sentence. Once we have this representation we can use further models in order to compute an additional score for each of the possible candidates and then choose the one with the best score. Ideally these additional models would be integrated into the generation algorithm, but most of them are too costly to include in the search procedure or do not have a structure which allows this kind of coupling. How to efficiently compute n-best lists and word graphs for the phrase-based approach is presented in (Zens and Ney, 2005).

3.4 IBM1 Rescoring

Although the IBM1 model is the easiest one of the single-word based translation models and the phrase based models clearly outperform this approach, the inclusion of the scores of this model, i.e.

$$h_{\mathrm{IBM1}}(f_1^J|e_1^I) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(f_j|e_i) \quad (6)$$

has been shown experimentally to improve the performance of a machine translation system (Och et al., 2003).

3.5 LM Rescoring

During the generation process, a single language model is used. However, additional language models specific to each sentence to be translated can help to improve the machine translation quality (Hasan and Ney, 2005). The motivation behind this lies in the following observation: the syntactic structure of a sentence is influenced by its type. It is obvious that an interrogative sentence has a different structure from a declarative one due to non-local dependencies arising e.g. from *wh*-extraction. As an example, let us consider the syntax of the following sentences: "Is the commissioner ready to give an undertaking?" and "The commissioner is ready to give an undertaking." If we look closer at the first four words of each sentence (*is*, *the*, *commissioner* and *ready*), the trigrams observed are quite different, leading to the hypothesis that a language model that can discriminate between these cases also performs better than the traditional approach.

We apply a method based on regular expressions to cluster the sentences into specific classes. A very simple trigger for an interrogative class is e.g. a question mark "?". This information is then used to train class-specific language models which are interpolated with the main language model in order to elude data sparseness.

4 Experimental results

4.1 EPPS Databases

The *European Parliament* (EP) usually holds plenary sessions six days each month. The major part of the sessions takes place in Strasbourg (France) while the residual sessions are held in Brussels (Belgium). Today the European Parliament consists of members from 25 countries, and 20 official languages are spoken. The sessions are chaired by the President of the European Parliament. Simultaneous translations of the original speech are provided by interpreters in all official languages of the EU.

It is possible to categorize speakers in two ways: Firstly there are native speakers as well as non-native speakers who have more or less pronounced accent. Secondly there are original speakers and interpreters. Although most of the speeches are planned, almost all speakers exhibit the usual effects known from spontaneous speech (hesitations, false starts, articulatory noises). The interpreters' speaking style is somewhat choppy: dense speech intervals ("bursts") alternate with pauses when the interpreter is listening to the original speech.

The European Union's TV news agency, *Europe by Satellite (EbS)*[3] broadcasts the EP Plenary Sessions live in the original language and the simultaneous translations via satellite on different audio channels: one channel for each official language of the EU and an extra channel for the original untranslated speeches. These channels are also available as 30 minute long internet streams for one week after the session.

A preliminary version of the texts of the speeches given by members of the European Parliament are published on the EUROPARL website[4] on the day after the EPPS. After a time period (about two months) in which the politicians are allowed to make corrections the speeches are available in its final form, known as Final Text Edition (FTE), in all official languages of the EU. The website also provides all previous reports since April 1996. We worked with the available reports building an English-Spanish parallel text corpus for the TC-Star project. The FTE aims for high readability, and therefore does not provide a strict word-by-word transcript. Notable deviations from the original speech include removal of hesitations, false starts and word interruptions. Furthermore transposition, substitution, deletion and insertion of words can be observed in the reports. An example is given in Table 1. Additionally human transcribers were asked to produce an accurate verbatim transcription of the speeches of the politicians.

4.2 Experimental Setup

Before training the translation models, some preprocessing steps have been carried out in order to better adapt the training material to the evaluation conditions.

For the FTE, the goal is to produce high quality translations, as produced by human translators. This includes punctuation marks and true casing. So, hardly any preprocessing of the data is needed. In order to aid the translation system, a categorization of the text has been carried out where numbers, dates, proper names, etc. have been detected and marked. The text is also lowercased in order to reduce the vocabulary size, and the case information is restored after the translation with help of the `disambig` tool from the SRILM toolkit (Stolcke, 2002).

For the verbatim transcriptions, we did some additional preprocessing. In the texts to be translated, we did some normalization, like expanding contractions ("I am" instead of "I'm", "we will" instead of "we'll", etc.) and eliminating hesitations ("uhm-", "ah-", etc.). In addition, as current state-of-the-art speech recognition systems do not generate (reliable) punctuation marks and most of them produce only lowercased text, we eliminated the punctuation marks of the training texts. Additionally, all numbers are written out (e.g. "forty-two" instead of "42"). The statistics of the training corpus are given in Table 2.

The statistics of the corpus used for evaluation can be seen in Table 3. The texts correspond to the plenary sessions held between the 15th and 18th November 2004. It should be noted that the notion of 'sentences' is slightly different in both conditions. In the FTE the 'sentences' correspond to grammatical sentences, while for the verbatim transcriptions the 'sentences' correspond to the segmentation used for the speech recognition systems. This is guided by the silences in the audio signal more than by the grammatical constructs.

4.3 Evaluation Metrics

In a later phase of the TC-Star project, a human evaluation of the output of the translation systems will be carried out. In this early stage however, only an automatic evaluation of the results has been performed. We used the four standard evaluation metrics WER, PER, BLEU and NIST:

- WER (**w**ord **e**rror **r**ate): The WER is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference sentence.
- PER (**p**osition-independent word **e**rror **r**ate): A shortcoming of the WER is that it requires a perfect word order. The word

[3] http://europa.eu.int/comm/ebs/

[4] http://www.europarl.eu.int/plenary/

Verbatim Transcription	Final Text Edition
It is for our Parliament, as we have already marked in a symbolic ceremony *outdoor*, a special and extraordinary moment. *It was described in Dublin last Saturday captured in the words of* Ireland's Nobel literature laureate Seamus Heaney, *he talked about and I quote ...*	It is for our Parliament, as we have already marked in a symbolic ceremony *outside*, a special and extraordinary moment. *In Dublin last Saturday*, Ireland's Nobel literature laureate Seamus Heaney *captured this special event with the words ...*

Table 1: Excerpt of the verbatim transcription corpus and the corresponding Final Text Edition.

	Spanish	English
Sentence pairs	1 207 740	
Running Words	34 851 423	33 335 048
Running Words without Punct. Marks	31 360 260	30 049 355
Vocabulary	139 587	93 995
Singletons	48 631	33 891

Table 2: Statistics of the EPPS training corpus.

order of an acceptable sentence can be different from that of the target sentence, so that the WER measure alone could be misleading. The PER compares the words in the two sentences ignoring the word order.

- BLEU and NIST scores: These scores are a weighted n-gram precision in combination with a penalty for sentences which are too short, and were defined in (Papineni et al., 2002) and (Doddington, 2002). Both measure accuracy, i.e. large scores are better.

All of these metrics can be extended to the case where we have multiple references by calculating the value for each of the reference translations and choosing the best one among them. In our case we had two references per sentence.

4.4 Results

The results for the FTE corpus are given in Table 4. The baseline results refer to the output of the translation system, as described in Section 3.1, without any of the further improvements discussed in Section 3. It can be seen that the log-linear combination of models significantly improves the translation quality. For Spanish to English, the WER is reduced by 14% absolute, from 49.2% to 35.2%. Both the IBM1-rescoring and the additional language models also help to improve the quality of the translation, although in a lesser way (1% WER for both models combined). The improvements are consistent in all the evaluation metrics[5].

[5]Note that the system was trained to optimize the BLEU score.

The results for the verbatim transcriptions can be found in Table 5. We can observe a slight degradation in the performance of the system. This is mainly due to ungrammatical structures in the sentences. Additionally Table 6 shows the *oracle error rate*; the error rate of the *best* translations contained in an n-best list selected by comparing with reference translations, depending of the size of the list. These values were computed in a separate EPPS development corpus on which we optimized the scaling factors. It can be seen that the WER decreases as the size of the list increases. In our experiments we used a 10 000-best list.

The last line in each of the blocks of Table 5 shows the performance of the translation system when the input of the system is not the verbatim transcription of the speeches but the output of the speech recognizer. It is worth noting that the loss in performance is much smaller than the word error rate of the speech recognition system. For example, for English to Spanish translation, the degradation of the translation quality is about 4% WER while the WER of the speech recognition system is 9.5%. This shows that the statistical approach to speech-to-speech translation is robust with respect to errors in the speech recognition system. Table 7 shows some translation examples and the effect of speech recognition errors.

5 Conclusions

In this paper we have presented our ongoing work on a speech-to-speech translation system within the TC-Star project. The task, trans-

		Spanish	English
FTE	Sentences	840	1094
	Running Words	22 756	26 885
	Vocabulary	3644	3744
	OOVs (running words)	40	102
Verbatim & ASR	Sentences	1073	792
	Running Words	18 896	19 306
	Vocabulary	3302	2772
	OOVs (running words)	145	44
	Number of Politicians*	36	11
	Input WER (ASR only)	10.1%	9.5%

* Unknown number of interpreters.

Table 3: Statistics of the EPPS test corpus.

		WER [%]	PER [%]	BLEU [%]	NIST
English to Spanish	Phrase based model	50.7	44.7	29.4	6.35
	Log-linear model	40.9	32.0	46.3	9.80
	+ IBM-1 rescoring	40.1	30.9	47.8	9.88
	+ LM rescoring	39.9	30.6	48.6	9.95
Spanish to English	Phrase based model	49.2	43.4	29.3	6.53
	Log-linear model	35.2	26.7	53.8	10.48
	+ IBM-1 rescoring	34.5	25.9	54.8	10.65
	+ LM rescoring	34.3	25.9	55.0	10.68

Table 4: Results for the Final Text Editions (FTE) corpus.

lation of parliamentary speeches, is a difficult translation task, as the domain is a broad one with a big vocabulary and long sentences. The results obtained are competitive or superior to the ones presented by other groups on this and similar tasks. We have also shown that, with these methods, it is possible to directly translate the output of a speech recognition system, and the statistical approach to translation is able to recover from errors produced by the speech recognizer.

There are still open questions that will be subject of further research within the project. In order to draw more relevant conclusions we have to further analyze the training and testing data. It is important to measure the overlap of topics between both corpora. Due to the chronological nature of the selected data, this overlap is probably high. This is proved for example by the small number of out of vocabulary words in the test data. It is also interesting to investigate the overlap of speakers between the training and testing corpora. It is possible that taking information about the speaker into account helps to adapt the translation model to the specific speaking style of the politicians.

Due to the limited audio data available, which has to be manually transcribed, for these experiments we have also included the text spoken by the interpreters as part of our test data. Of course, the ultimate goal will be to translate only the politicians' speeches. New data is being collected and for the next evaluation within the project we expect to be able to make experiments on a politician-only evaluation corpus.

Another research topic is to include additional sources of information, like morphosyntactic information into the translation process. Having Spanish as one of the languages, this should especially improve the translation quality, as this language has a very rich verb morphology. By examining the generated translations, one can see that for the direction English to Spanish the tense of the Spanish verb is often not chosen correctly. This can partly explain the difference in performance between the two translation directions.

Furthermore, a tighter coupling between the speech recognition and translation system is being worked on. For the experiments reported in this paper, we used a serial architecture, i.e. the speech recognition system receives the audio signal and produces a transcription which is passed on to the translation system. Ongoing

		WER [%]	PER [%]	BLEU [%]	NIST
English to Spanish	verbatim	46.1	35.4	42.5	9.33
	ASR	49.8	38.6	38.7	8.73
Spanish to English	verbatim	42.5	31.7	45.9	9.75
	ASR	46.6	35.4	41.5	9.12

Table 5: Results for Verbatim and ASR.

size	WER[%]
1	37.0
5	27.8
10	25.2
25	22.3
50	20.7
100	19.4
500	17.3
5 000	16.1

Table 6: Oracle word error rate (on development data) for different sized n-best lists.

research aims to integrate both systems so that the translation system can benefit from multiple hypothesis produced by the first system.

Acknowledgments

This work has been partly funded by the integrated project TC-STAR – Technology and Corpora for Speech-to-Speech Translation – (IST-2002-FP6-506738), and by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistische Textübersetzung" (Ne572/5).

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology.*

EuTransProject. 2000. Final report of esprit research project 30268 (EuTrans): Example-based language translation systems. Technical report, Instituto Tecnológico de Informática (ITI, Spain) and Fondazione Ugo Bordone (FUB, Italy) and RWTH Aachen and Lehrstuhl für Informatik VI (Germany).

Saša Hasan and Hermann Ney. 2005. Clustered language models based on regular expressions for SMT. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, May. To appear.

A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei, and F. Balducci. 2001. Architecture and design considerations in NESPOLE!: a speech translation system for e-commerce applications. In *HLT 2001 - Human Language Technology Conference*, San Diego, CA, March.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Syntax for statistical machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Johns Hopkins University 2003 summer workshop final report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++.* Cambridge University Press, Cambridge, UK.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int.*

VERBATIM	los proyectos de enmienda deberán presentarse con la firma **de** treinta y siete diputados como mínimo **o en** nombre de una Comisión
ASR	los proyectos de enmienda deberán presentarse con la firma **del** treinta y siete diputados como mínimo **buen** nombre de una comisión
TRANS VERBATIM	the amendment of projects must be made with the signing of thirty seven Members as minimum **or on behalf** of a **committee**
TRANS ASR	the amendment of projects must be made with the signing of **the** thirty seven Members as minimum **good name** of a **Commission**
REFERENCE 1	the draft amendments must be tabled with the signatures of at least thirty-seven Members or on behalf of a Committee
REFERENCE 2	amendment projects must be presented with the signature of at least thirty-seven members or on behalf of a Committee
VERBATIM	espero que la reunión **del** día veintiuno sea suficientemente interesante para justificar el trabajo y el coste de **convocarla**
ASR	espero que la reunión **el** día veintiuno sea suficientemente interesante para justificar el trabajo y el coste de **convocar**
TRANS VERBATIM	I hope that the meeting **of** twenty one is sufficiently interesting to justify the work and the cost of **around**
TRANS ASR	I hope that the meeting **the** twenty one **day** is sufficiently interesting to justify the work and the cost of **convene**
REFERENCE 1	I hope that the meeting on the twenty-first will be interesting enough to justify the work and expense of convening it
REFERENCE 2	I hope the meeting of the coming twenty-first is sufficiently interesting to justify the work and the cost of convening it
VERBATIM	**tendremos** media hora para atender a este turno de preguntas
ASR	**debemos** media hora para atender a este turno de preguntas
TRANS VERBATIM	we **have** half an hour to meet this question time
TRANS ASR	we **must** half-hour to meet this question time
REFERENCE 1	we have half an hour for this questions and answers session
REFERENCE 2	we will have half an hour to attend to this turn of questions

Table 7: Examples of translation quality for different conditions.

Conf. on Spoken Language Processing, volume 2, pages 901–904, Denver.

Nicola Ueffing and Hermann Ney. 2005. Results on different structured language resources for speech-to-speech translation systems. Technical Report Deliverable D4.5, LC-STAR project by the European Community (IST project ref. no. 2001-32216), January.

Nicola Ueffing, Franz Josef Och, and Hermann Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 156–163, Philadelphia, PA, July.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of speech-to-speech translations*. Springer Verlag, Berlin, Germany, July.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 257–264, Boston, MA, May.

Richard Zens and Hermann Ney. 2005. Word graphs for statistical machine translation. In *Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*, Ann Arbor, Michigan, June. To appear.