



## LC-STAR Deliverable D4.5

Project ref. no.	IST-2001-32216
Project acronym	LC-STAR
Project full title	Lexica and Corpora for Speech-to-Speech Translation Technologies
Security (distribution level)	Public
Contractual date of delivery	M27 = Apr 2004
Actual date of delivery	M36 = Jan 2005
Deliverable number	D4.5
Deliverable name	Results on different structured language resources for speech-to-speech translation systems
Type	Report
Number of pages	15
WP contributing to the deliverable	WP4, WP5
WP / Task / Deliverable responsible	WP4 - Task 4.4 - D4.5 RWT
Other contributors	Saša Hasan, Stephan Kanthak, Klaus Macherey, Evgeny Matusov, Maja Popović, David Vilar, Richard Zens (RWT)
Author(s)	Nicola Ueffing, Hermann Ney (RWT)
EC Project Officer	Kimmo Rossi
Project Coordinator	Name: Harald Höge
Company:	Siemens AG, CT IC 5
Address:	Otto-Hahn-Ring 6, 81739 München, Germany
Phone:	+49-89-636-53374
Fax:	+49-89-636-49802
E-mail:	harald.hoege@mchp.siemens.de
Project web site:	<a href="http://www.lc-star.com">http://www.lc-star.com</a>
Keywords	Machine Translation, Speech Translation, Language Resources, Corpora, Lexica, Base Forms, POS-tags
Abstract (for dissemination)	We describe the experimental results using the baseline speech-to-speech translation systems created in D4.3 and compare them to an enhanced translation system taking different language resources into account. Experiments were performed on the trilingual corpus (English, Spanish, Catalan) built within the project in WP5. This corpus consists of spontaneous dialogues in the domain of tourism, travel arrangement and appointment scheduling. It is enriched with base form and POS information. Additional language resources that were used are a phrasal lexicon (the LOS created in WP5) and a verb full form list for Spanish and Catalan.

**Document evolution:**

Version	Date	Security	Notes
V0.1	May 06, 2004	Project internal	First draft version – work document, to be discussed by WP partners
V0.2	Oct 19, 2004	Project internal	Second draft version (500k words corpus), to be discussed by WP partners
V1.0	Jan 10, 2005	Project internal	Pre-final version, including experiments with LOS
V1.1	February 4, 2005	Public	Final version

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Experimental Setting</b>	<b>1</b>
2.1	Language Resources . . . . .	1
2.2	Evaluation Metrics . . . . .	3
2.3	Description of the SMT Systems . . . . .	4
2.3.1	Alignment Template Approach . . . . .	4
2.3.2	Finite State Transducer Approach . . . . .	5
2.4	Characteristics of the Different Translation Pairs . . . . .	6
<b>3</b>	<b>Baseline Results</b>	<b>6</b>
<b>4</b>	<b>Integration of Phrasal Lexica as Additional Language Resource</b>	<b>7</b>
<b>5</b>	<b>Sentence-class-based Language Models</b>	<b>9</b>
5.1	Approach . . . . .	9
5.2	Experimental Results . . . . .	10
<b>6</b>	<b>Lexicon Smoothing with Base Forms</b>	<b>10</b>
<b>7</b>	<b>Coupling of Speech Recognition and Translation</b>	<b>11</b>
<b>8</b>	<b>Related Experiments on Other Tasks</b>	<b>12</b>
8.1	Selection of Additional Training Data Using $n$ -gram Coverage . . . . .	12
8.2	Part-of-Speech Information for Source Sentence Reordering . . . . .	13
8.3	Morphological Information for Word Alignment . . . . .	13
<b>9</b>	<b>Use of WordNet</b>	<b>14</b>
9.1	Coverage of Words by WordNet . . . . .	14
9.2	Rescoring Using WordNet-based Features . . . . .	14
<b>10</b>	<b>Conclusion</b>	<b>15</b>

## 1 Introduction

The LC-STAR project aims at the creation of bilingual aligned text corpora and large lexica for speech-to-speech translation. The main task of WP4 is the analysis of the amount of training data and the kind of language resources needed to obtain a good speech-to-speech translation technology. Since the acquisition of language resources is expensive, the relation between data acquisition costs and translation quality has been investigated. For this purpose, experimental speech-to-speech translation systems in which different kinds of language resources are integrated have been set up in WP4.

In this report, we describe the experimental results using these speech-to-speech translation systems. Experiments were performed on the trilingual corpus comprising the languages English, Spanish, and Catalan which has successively been built within the project. Furthermore, we will describe methods of integration of different types of language resources into those systems and study their effect on speech-to-speech translation quality.

This report is organized as follows: Section 2 describes the experimental settings, like the current state of the trilingual corpus and the additional language resources (section 2.1), the evaluation metrics for machine translation (section 2.2), and the statistical machine translation systems used for the experiments (section 2.3). The baseline results of the translation system will be presented in section 3. In sections 4 to 8, different methods of integrating additional language resources such as POS-tags and base forms into the statistical machine translation (SMT) process will be described, and experimental results will be given. Section 9 contains an excursus on the use of WordNet for SMT. This report will be concluded in section 10.

## 2 Experimental Setting

In this section, we will describe the corpora that are used for the translation experiments, give a short overview over the evaluation metrics, and depict the SMT systems that we employed. Then, we will shortly explain the characteristics of the different language pairs that were investigated.

### 2.1 Language Resources

In LC-STAR task T5.6, a trilingual corpus comprising English, Spanish and Catalan has been successively created. It consists of spontaneous dialogues in all three languages which are transcribed and translated into the two other languages. At the time of this report, there was a pre-final version of the corpus available. Out of the corpus, we randomly selected sentences for the develop and test corpora; the rest was taken for training.

The trilingual corpus is annotated with POS-tags for all three languages and base forms for Spanish and Catalan. The creation of those data is described in detail in [Arranz & Castell<sup>+</sup> 04]. The corpus statistics are given in Table 1. We see that the Spanish and Catalan parts of the corpus contain a lower number of running words, but have a much larger vocabulary than the English part. Moreover, the number of singletons in the training corpus and that of unknown words (OOV) in the develop and test corpora are much higher for Spanish and Catalan. The reason for these effects is that Spanish and Catalan have a richer morphology than English. If we consider only the number of base form instead of full form words for Spanish and Catalan, we obtain numbers that are more similar to those for English.

One of the objectives of the LC-STAR project was to analyze the kind and amount of language resources necessary to obtain reasonable end-to-end speech translation quality. Pre-studies have

Table 1: Statistics of the trilingual dialogue corpus for English, Spanish, and Catalan.

		Spanish	Catalan	English
Train	Sentences	40 574		
	Running Words	482 290	485 741	516 717
	Running Words without Punctuation Marks	399 607	402 668	436 140
	Vocabulary	14 327	12 773	8 116
	Singletons	6 743	5 931	3 081
Dev.	Sentences	972		
	Running Words	12 883	13 048	13 983
	Running Words without Punctuation Marks	10 733	10 888	11 920
	Vocabulary	1 988	1 943	1 584
	OOVs (running words)	214	184	100
	Trigram perplexity	49.2	47.8	34.5
Test	Sentences	972		
	Running Words	12 771	12 991	13 922
	Running Words without Punctuation Marks	10 654	10 844	11 871
	Vocabulary	1 997	1 913	1 583
	OOVs (running words)	213	174	124
	Trigram perplexity	48.2	47.7	34.4

shown that the quality of machine translation improves significantly with the amount of bilingual and monolingual training data [Ueffing & Ney 02a].

To investigate the gain of using additional language resources as opposed to more bilingual (full-form only) data, we constructed a reduced training corpus. This was done by randomly extracting 1k sentences from the trilingual dialogue training set. The corpus statistics of this reduced training corpus are given in Table 2. The number of OOVs in the develop and test corpora rise by a factor of up to 10.

An additional language resource has been created for those three languages: a phrasal lexicon or *list of sentences (LOS)* (LC-STAR deliverable D5.8) as described in LC-STAR deliverable D5.3 [Arranz & Castell<sup>+</sup> 04]. The corpus consists of a US-English reference phrase list and

Table 2: Statistics of the *reduced* trilingual dialogue corpus for English, Spanish, and Catalan. The number of OOVs for the develop and test corpora described in Table 1 are included for comparison.

		Spanish	Catalan	English
Train	Sentences	1 014		
	Running Words	12 138	12 215	12 972
	Running Words without Punctuation Marks	10 102	10 179	10 978
	Vocabulary	1 880	1 823	1 436
	Singletons	1 150	1 070	744
Dev.	OOVs (running words)	1 386	1 317	1 002
Test	OOVs (running words)	1 362	1 271	1 053

Table 3: Statistics of the list of sentences (LOS) for English, Spanish, and Catalan. The number of OOVs for the develop and test corpora described in Table 1 are included for comparison.

	Spanish	Catalan	English
LOS Entries	10 520		
Running Words	44 289	46 002	41 850
Running Words without Punctuation Marks	43 753	45 542	41 293
Vocabulary	10 797	10 460	11 167
Singletons	6 573	6 218	7 153
Vocabulary entries <i>not</i> in dialogue corpus	6 153	5 855	6 756
Dev. OOVs (running words)	1 081	918	478
Test OOVs (running words)	1 128	892	599

its translation into Spanish and Catalan. The US-English phrases have been extracted from dialogue corpora and websites, and parts of them have been manually created. The LOS can be taken as additional data in training.

The corpus statistics of the LOS are given in Table 3. We see that the vocabularies of those short phrases are very large for all three languages, especially for English. Thus, the LOS are a good resource for extending an already existing system. The number of additional vocabulary entries with respect to the dialogue corpus is around 6k for each of the languages. The average phrase length is short with around 4 words per entry. The number of OOVs of the develop and test corpora with respect to the LOS vocabulary is about five times as high as the ones for the full training corpus for all three languages.

For Spanish and Catalan, there was also a monolingual language resource available: a list of word base forms and the full forms that can be derived from them. This list was extracted from the morphological analyzer by UPC [Carmona & Cervell<sup>+</sup> 98]. From this list, we used the verbs only which are about 12k base forms per language.

## 2.2 Evaluation Metrics

The question of how to measure quality and compare performance of different translation systems on a certain corpus in a fast and cheap way is still open; MT research suffers from the lack of suitable, consistent, and easy-to-use criteria for the evaluation of the experimental results. In the LC-STAR project, we applied automatic evaluation, since it is fast and generates reproducible results. The quality of the output of the machine translation systems is measured automatically by comparing the generated translation to a given (manually created) reference translation. The three following criteria are used:

- WER (word error rate):

The word error rate is based on the Levenshtein distance. It is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated string into the reference string.

- PER (position independent word error rate):

A shortcoming of the WER is the fact that it requires a perfect word order. The word order of the automatically generated target sentence can be different from that of the given target sentence, but nevertheless acceptable. To overcome this problem, the position independent word error rate (PER) was introduced as additional measure. It compares the words in the two sentences *without* taking the word order into account. Words that have no matching

counterparts are counted as substitution errors, missing words are deletion and additional words are insertion errors. The PER is a lower bound for the WER.

- BLEU (bilingual evaluation understudy) [Papineni & Roukos<sup>+</sup> 02]: BLEU is based on the notion of modified  $n$ -gram precision, for which all candidate  $n$ -gram counts in the translation are collected and clipped against their corresponding maximum reference counts. These clipped candidate counts are summed and normalized by the total number of candidate  $n$ -grams. Since BLEU expresses quality, we determine  $100 - \text{BLEU}$  to transform it into an error measure.

Although these measures are only approximations, they seem to be sufficient at the present level of performance of machine translation systems. For a more detailed description of evaluation metrics in machine translation see LC-STAR deliverable D4.1 [Ueffing & Ney<sup>+</sup> 02b] and [Leusch & Ueffing<sup>+</sup> 03].

## 2.3 Description of the SMT Systems

We performed experiments using two different SMT systems. One is based on groups of words, so-called *alignment templates*, and the other one uses finite state transducers for modelling the translation process. Both of them choose the target string that has maximal probability given the source string:

$$\hat{e} = \arg \max_{e_1^I} Pr(e_1^I | f_1^J) .$$

### 2.3.1 Alignment Template Approach

In this section, we give a brief description of the phrase-based statistical translation system. The key elements of this translation approach are the *alignment templates* [Bender & Zens<sup>+</sup> 04, Och & Ney 04]. These are pairs of source and target language phrases with an alignment within the phrases. The alignment templates are build at the level of word classes. This improves the generalization capability of the model.

The system takes a number of different sub-models into account which are combined log-linearly. Those sub-models are:

- a phrase translation model
- a word translation model
- two language models: a word-based  $n$ -gram model and a class-based  $n$ -gram model. The orders of both language models have been optimized for each translation direction individually, where the values range between 3 and 6.
- two heuristics: the word penalty and the alignment template penalty which assign costs to each word/alignment template that is generated
- three feature functions that model reordering on the level of alignment templates and on the word-level

A dynamic programming beam search algorithm is used to generate the translation hypothesis with maximum probability for a given source sentence. This search algorithm allows for arbitrary reorderings at the level of alignment templates. Within the alignment templates, the reordering is learned in training and kept fix during the search process. There are no constraints on the reorderings within the alignment templates.

This is only a brief description of the alignment template approach. For further details, see [Bender & Zens<sup>+</sup> 04, Och & Ney 04].

The system applied here has been continuously improved in the duration of the LC-STAR project. There has been a number of changes compared to the system described in LC-STAR deliverable D4.4 [Ueffing & Ney 03], such as

- the number of sub-models considered in the translation process has been increased,
- the language models have been extended to account for longer histories,
- the word alignment process has been improved,
- the training of the combination of the different sub-models has been optimized.

An overview of the current system is given in [Bender & Zens<sup>+</sup> 04].

### 2.3.2 Finite State Transducer Approach

Statistical machine translation may be viewed as a weighted language transduction problem [Vidal 97]. This makes it possible to build a machine translation system with the use of weighted finite-state transducers.

The translation problem can be solved by estimating a language model on a bilanguage defined over source and target language (see also [Bangalore & Riccardi 00, Casacuberta & Llorens<sup>+</sup> 01]). This bilanguage is learned from the training corpus that has been (automatically) word aligned beforehand. An example of sentences from this bilanguage is given in Figure 1 for the Catalan to English translation task in the LC-STAR corpus (in this figure, Catalan is simplified and does not contain accents). The bilanguage words are of the form `source|target`. The `$` symbol denotes transitions that contain a source but no target word. Those can occur if the source word does not have any correspondance in the target sentence. Another reason can be that due to the fixed segmentation given by the word alignments, phrases in the target language are moved to the last source word of an alignment block, as the example 'donar per favor' / 'please give me' shows.

```
podria|could_you donar|$ per|$ favor|please_give_me el|$ seu|your
      numero|number ?|?

si|yes ,|, be|well ,|, no|I_do_not se|know .|.

```

Figure 1: Examples of bilanguage sentences (Catalan → English).

Thus, the steps for training the translation system are the following:

1. Perform word alignment (using the publically available GIZA++ toolkit<sup>1</sup>),
2. transform the training corpus with a given alignment into the corresponding bilingual corpus,
3. train a language model on the bilingual corpus,
4. build an acceptor  $A$  from the language model.

Then the translation problem from above can be solved by composition of finite state transducers. A more detailed description of the system can be found in [Kanthak & Ney 04].

<sup>1</sup>The GIZA++ toolkit for word alignment can be downloaded from <http://www-i6.informatik.rwth-aachen.de/web/Software/index.html>

## 2.4 Characteristics of the Different Translation Pairs

When considering the six translation directions from the trilingual LC-STAR corpus, we see that the language pairs can be divided into three different classes:

1. The language pair Spanish/Catalan is by far the easiest to translate, because the languages are very similar in word order and morphology.
2. The translation from Spanish or Catalan into English is harder, due to the difference in word order and morphology.
3. The hardest translation direction is from English into Spanish or Catalan. This is due to the poorness of the English morphology compared to the other two languages. The word English full forms contain less information which makes it difficult for an automatic system to generate the correct inflection in the target language.

We will see these facts reflected in the translation results presented in the following sections.

## 3 Baseline Results

In this section, we present the baseline results for all six translation directions on the development and test corpora using the alignment template system described in section 2.3.1. Please note that the baseline system has been improved in the duration of the LC-STAR project (cf. section 2.3.1), such that the translation quality of this system lies beyond that of the system described in deliverable D4.4 [Ueffing & Ney 03].

We see that the translation from Spanish into Catalan and vice versa is by far the easiest one: The WER lies around 11-12%. Considering the fact that there are only single references available for the develop and test corpora, we can assume that the actual error rate lies even below this value.

As explained in section 2.4, the translation from Spanish into English and from Catalan into English are approximately of the same complexity. This fact is reflected in the error rates which are very similar (around 40% WER).

On the hardest translation tasks, the translation from English into Spanish and Catalan, the baseline system achieves a translation quality of 41% to 43% in WER.

In the following sections, we will present methods making use of additional LRs that improve the performance of the translation system.

Table 4: Translation error rates [%] on the development and test set for all six language pairs. Optimized baseline system, single best hypothesis.

System	dev			test		
	WER	PER	100-BLEU	WER	PER	100-BLEU
Catalan → Spanish	11.4	10.7	19.8	12.6	11.7	21.1
Spanish → Catalan	10.9	10.2	21.2	11.7	11.1	21.6
Catalan → English	39.7	32.2	58.7	41.6	33.1	60.2
English → Catalan	41.6	35.9	59.9	43.6	37.1	61.7
Spanish → English	39.9	32.6	59.2	41.3	33.8	61.1
English → Spanish	41.3	35.4	56.6	42.8	35.8	58.1

## 4 Integration of Phrasal Lexica as Additional Language Resource

In this section, we present the results for the full training corpus and for the reduced training corpus introduced in section 2.1 with and without use of additional phrasal lexicon. We also present the results obtained using only the phrasal lexicon. An extended form of the phrasal lexicon containing expanded verb forms has also been investigated for the translation with scarce training material.

Extensions of the phrasal lexicon have been done using base forms and POS information for the Spanish and Catalan verbs and expansion tables for those two languages containing base forms and all their full forms with POS tags (cf. section 2.1). Each base form seen in the lexicon more than 5 times has been expanded with all POS tags seen in the lexicon. The English equivalents are chosen using lexical probabilities and then manually checked and eventually corrected. Finally, the base forms and POS tags are mapped to the corresponding full forms by using the lists, e.g. “ir VMIF1P0” is mapped to “iremos”. In this way, the lexicon was enriched with some previously unseen full forms of the verb.

The translation with the system trained only on the phrasal lexicon is done without language model, as well as with the language model trained on the full target language corpus.

The language pairs in these experiments are Spanish-English and Catalan-English, and translation is performed in all four directions using the alignment template system with optimized scaling factors. The results are shown in Tables 5 and 6.

Table 5: Translation error rates [%] on the development and test set for Spanish  $\leftrightarrow$  English. Results are presented for different training corpora; 40k and 1k sentences and LOS only. Effect of adding the LOS to the training material, verb expansions and a full language model.

			dev			test		
			WER	PER	100-BLEU	WER	PER	100-BLEU
S $\rightarrow$ E	<b>40k</b>	(full training)	39.9	32.6	59.2	41.3	33.8	61.1
		+ LOS	39.5	32.1	58.9	40.8	32.8	59.9
	<b>1k</b>	(reduced training)	53.6	44.9	75.9	54.8	46.0	76.8
		+ LOS	49.8	40.9	70.9	50.7	41.7	71.7
		+ verb expansions	48.7	39.4	69.8	50.1	40.3	70.9
		+full LM (40k)	48.2	39.1	68.9	49.2	39.8	69.8
	<b>LOS</b>		57.8	47.1	78.8	59.3	47.8	80.7
		+ verb expansions	56.7	45.9	78.1	57.6	46.5	78.8
		+ full LM (40k)	53.9	44.0	75.0	56.0	45.7	76.6
<hr/>								
E $\rightarrow$ S	<b>40k</b>	(full training)	41.3	35.4	56.6	42.8	35.8	58.1
		+ LOS	40.7	34.4	55.8	42.9	35.9	57.8
	<b>1k</b>	(reduced training)	57.5	49.2	74.3	58.4	49.9	76.5
		+ LOS	52.4	43.9	68.0	53.6	44.9	68.7
		+ verb expansions	51.4	43.0	67.7	52.8	44.0	68.4
		+full LM (40k)	50.4	42.3	66.2	52.1	43.6	67.6
	<b>LOS</b>		61.3	51.8	74.6	62.3	52.7	75.7
		+ verb expansions	58.1	49.7	73.2	59.5	50.4	74.7
		+ full LM (40k)	57.6	49.4	72.5	59.1	50.1	73.9

Table 6: Translation error rates [%] on the development and test set for Catalan  $\leftrightarrow$  English. Results are presented for different training corpora; 40k and 1k sentences and LOS only. Effect of adding the LOS to the training material, verb expansions and a full language model.

			dev			test		
			WER	PER	100-BLEU	WER	PER	100-BLEU
C $\rightarrow$ E	<b>40k</b>	(full training)	39.7	32.2	58.7	41.6	33.1	60.2
		+ LOS	38.9	31.8	58.6	41.1	33.2	60.2
	<b>1k</b>	(reduced training)	53.4	44.8	76.4	54.2	45.0	76.6
		+ LOS	49.2	40.4	71.8	50.9	41.4	72.7
		+ verb expansions	48.9	39.3	70.9	50.1	39.7	71.4
		+ full LM (40k)	48.1	38.7	69.8	49.4	39.3	70.4
	<b>LOS</b>		57.2	47.3	79.2	59.0	47.9	80.2
		+ verb expansions	55.0	44.6	76.9	56.1	45.1	76.7
		+ full LM (40k)	54.0	44.0	75.2	55.6	44.7	75.9
<hr/>								
E $\rightarrow$ C	<b>40k</b>	(full training)	41.6	35.9	59.9	43.6	37.1	61.7
		+ LOS	41.0	35.0	57.8	43.3	36.3	60.1
	<b>1k</b>	(reduced training)	57.2	49.3	75.2	57.9	49.8	75.1
		+ LOS	52.3	44.2	69.8	53.2	44.9	69.4
		+ verb expansions	51.6	43.7	69.3	53.0	44.8	70.0
		+ full LM (40k)	49.7	42.0	66.2	51.2	43.0	67.2
	<b>LOS</b>		63.0	53.8	76.3	65.0	55.1	78.2
		+ verb expansions	58.5	49.8	73.6	59.8	50.6	74.6
		+ full LM (40k)	57.7	49.3	73.4	59.1	50.2	74.6

As can be seen from the tables, the best results are obtained using the full training corpus with additional phrasal lexicon. When a large bilingual corpus from the domain is available, the improvements from the additional phrasal lexicon are relatively small.

However, for the reduced bilingual corpus, the importance of the phrasal lexicon is significant. The degradation in terms of WER and PER by using only 2.5% of the original corpus is not higher than 25% relative if the additional phrasal lexicon and language resources containing morphological information are available. This can be further improved by up to 1.9% absolute in WER if monolingual in-domain data is available and a better language model can be trained. The advantage of using such a small corpus is that its acquisition should not require any particular effort since in the matter of fact 1 000 parallel sentences in two or three languages can be also produced manually.

Using only the phrasal lexicon and additional resources, the obtained error rates are similar to those for the small training corpus.

Some translation examples for the direction English  $\rightarrow$  Spanish using only the phrasal lexicon with and without verb expansions are shown in Table 7. It can be seen that the extension of the lexicon enables the system to find the correct full form of the verb in the inflected language more often. For the translation Spanish  $\rightarrow$  English, Table 8 shows that for the extended lexicon the system is able to produce correct or approximatively correct translations also for full forms that have not been seen in the original training corpus. In the baseline system, those words remain untranslated and are marked by "UNKNOWN\_".

Table 7: Translation examples English  $\rightarrow$  Spanish using only the LOS as training data. Results are presented without and with additional verb expansions.

source sentence	well I am pretty interested .
train: LOS	bien estoy bastante interesa .
LOS + verb expansions	bien estoy bastante interesado .
source sentence	there is no problem .
train: LOS	hay no es problema
LOS + verb expansions	no hay problema .
source sentence	I do not know , what kind of sport do you like best ?
train: LOS	no me sabe , qué tipo de deporte te gusta mejor ?
LOS + verb expansions	no lo sé , qué tipo de deporte te gusta mejor ?

Table 8: Translation examples Spanish  $\rightarrow$  English using only the LOS as training data. Results are presented without and with additional verb expansions.

source sentence	si , esto , cuántas personas serán ?
train: LOS	if , this , how many people UNKNOWN_serán ?
LOS + verb expansions	if , that , how many people will be ?
source sentence	queríamos un poco de fruta , yogur , este tipo de cosas , algunas galletas .
train: LOS	UNKNOWN_queríamos a little fruit , yogurt , this kind of things , some salted .
LOS + verb expansions	we wanted a little fruit , yogurt , this kind of things , some salted .
source sentence	sí , los hoteles , nos movemos entre las tres y las cinco estrellas .
train: LOS	yes , hotels , we UNKNOWN_movemos between the three and the five-star .
LOS + verb expansions	yes , the hotels , we are moving within the three and the five-star .

## 5 Sentence-class-based Language Models

### 5.1 Approach

In this approach, a standard  $n$ -gram language model is interpolated with a special sentence-dependent language model. Each sentence is filtered according to various regular expressions. In this setup, the triggers are punctuation marks such as “?”, “!” and “.”, which separate the sentences into three classes: questions, exclamations and normal sentences. Another possibility is to add language dependent triggers that usually introduce subordinate clauses, e.g. “because” in English and “porque” in Spanish, or to base the triggers on POS-tag information of the verbs (which can be a coarse indicator of the upper-level syntactic structure of a sentence, cf. verb sub-categorization).

We use several different regular expression triggers based on punctuation marks and POS verb tags for additional language models which are interpolated with a base language model using five-grams and modified Kneser-Ney discounting. The interpolation lambdas are optimized on the development set for each class. The perplexities can be reduced around 10% relative

Table 9: Translation error rates [%] on the development and test set for all six language pairs. Results are presented for  $N$ -best list rescoring with the IBM-1 translation model and the sentence-class-based language models.

		dev			test		
		WER	PER	100-BLEU	WER	PER	100-BLEU
Cat → Spa	Baseline	11.4	10.7	19.8	12.6	11.7	21.1
	Rescoring	10.8	10.1	18.1	11.7	10.8	18.9
Spa → Cat	Baseline	10.9	10.2	21.2	11.7	11.1	21.6
	Rescoring	10.3	9.7	18.0	11.1	10.5	19.4
Cat → Eng	Baseline	39.7	32.2	58.7	41.6	33.1	60.2
	Rescoring	38.4	31.3	57.7	40.5	32.7	58.4
Spa → Eng	Baseline	39.9	32.6	59.2	41.3	33.8	61.2
	Rescoring	39.4	32.2	58.3	40.7	33.2	59.8

for Spanish (perplexity 48.2 → 42.9) and Catalan (perplexity 47.7 → 42.6), and 7% relative for English (34.4 → 31.9) with this kind of setting. Rescoring experiments were performed on  $N$ -best lists using the sentence-class-based language models and the IBM-1 translation model [Brown & Della Pietra<sup>+</sup> 93]. Section 5.2 shows the obtained results on several combinations of Catalan, Spanish and English.

## 5.2 Experimental Results

**Catalan-Spanish** For the Catalan-Spanish language pair, the standard alignment template training procedure has been followed (cf. section 3). The best results were obtained generating a 3,000-best list and applying the rescoring methods described in section 5.1. The results can be seen in Table 9 for the development and test sets of both translation directions.

**Catalan-English and Spanish-English** For the translation experiments from Catalan/Spanish to English, we applied the same procedure as for the language pair Catalan-Spanish. The only difference is that we used 10,000-best lists (instead of 3,000) for the rescoring experiments. This difference yields from the fact that the translations Catalan ↔ Spanish are nearly monotonous and thus shorter  $N$ -best lists are sufficient for representing the search space.

For all four translation directions, we see that the rescoring experiments on  $N$ -best lists slightly improve the overall translation performance.

## 6 Lexicon Smoothing with Base Forms

The standard lexicon model is based on full form words only. For highly inflected languages such as Spanish and Catalan this might cause problems, because many full form words occur only a few times in the training corpus. Compared to English, the token/type ratio for Spanish and Catalan is usually much lower (e.g. English 99.4, Catalan 66.6, Spanish 63.2). The information that multiple full form words share the same base form is not used in the lexicon model. To take this information into account, we smooth the lexicon model with a backing-off lexicon that is based on word base forms. The smoothing method we apply is absolute discounting with interpolation.

Table 10: Translation error rates [%] on the develop and test set for Catalan  $\rightarrow$  English. Results are presented with and without lexicon smoothing, and for a system that uses only Catalan base forms

System	dev			test		
	WER	PER	100-BLEU	WER	PER	100-BLEU
Baseline (only full form)	39.7	32.2	58.7	41.6	33.1	60.2
Only base form for Catalan	41.5	34.5	62.4	43.2	35.7	63.3
Smoothed full form	39.1	32.5	59.7	40.7	33.5	60.6

The effect of this smoothing method on the quality of the word alignment was evaluated on the German-English translation task, because there is manually annotated data available. It results in an improvement of the alignment error rate for both translation directions (cf. section 8.3). For the German-to-English direction the alignment error rate improves from 5.7% to 5.2% and for the English-to-German direction from 9.9% to 9.1%. These improvements of the alignment error rate are statistically significant at the 95% level. For more details, see [Zens & Matusov<sup>+</sup> 04].

We evaluated the effect of this lexicon smoothing on translation quality as well. Table 10 summarizes the results for the Catalan-to-English translation task. We compare three systems: first the baseline system which takes only full form words into account. The second system is based only on the base forms of Catalan, this means we replace each word in the Catalan source corpus with its base form and apply our standard translation system. The idea here is to ensure that taking only the base forms of Catalan does not already outperform the baseline system. The third system uses the described lexicon smoothing during the training of the word alignment. All the systems were optimized with respect to the word error rate (WER).

We see that the smoothed system slightly outperforms the baseline with respect to the word error rate on both the development and the test corpus.

## 7 Coupling of Speech Recognition and Translation

In most of the present approaches, speech translation is separated into the task of producing the speech transcript using automatic speech recognition methods and the task of automatic translation of this transcript. Since the automatically produced transcript often contains a number of speech recognition errors, this results in degradation of translation quality as compared to the translation of a perfect human transcript of the spoken utterances. We can also pursue an integrated approach which makes use of the fact that correct transcriptions of some sentences are often hypothesized, but not selected by the automatic speech recognition (ASR) system. In this case, in a statistical machine translation system, translations of such hypotheses may receive higher probability.

In the experiments with our finite-state transducer translation system (described in section 2.3.2), we followed this integrated approach and exploited the multiple hypotheses using the ASR word lattices as input. The translation process can then be represented as the composition operation of the word lattice for one sentence and the translation transducer. Each edge in the word lattice has its own acoustic score. We used a 4-gram language model for the target language to score and select constrained reorderings of the produced translations.

For the Spanish-English task, we introduced a mapping of some Spanish singletons to their morphological base forms for GIZA++ alignment training. This was done to improve the model estimation. These mappings were used also for translation. We represented the mappings as an unweighted transducer. When translating from word lattices, we mapped the singletons in the

word lattice by composing it with the mapping transducer. For the alignment training we also replaced named entities (names, city names, etc.) with labels. We also decided to reorder the target part of the training corpus based on the automatically trained word alignment to improve the estimation of a trigram language model on the level of bilingual tuples.

We used the RWTH statistical speech recognition system to produce first-best ASR hypotheses and word lattices. The experimental setup is described in Table 11. Since the test utterances were rather long, the corresponding ASR word lattices were of a large density <sup>2</sup>. However, by removing the epsilon-edges from the lattice and disposing (through determinization) of the time information which is irrelevant for translation, we reduced the lattice size and were able to translate without pruning. We optimized the scaling factor  $\lambda$  for the translation model scores on the development set. The optimal factor was 60. Table 12 shows the experimental results. The relatively bad overall performance of the system can be partially explained by a quite high ASR error rate of about 30 % and the availability of only one set of references. Nevertheless, we were still able to significantly improve most of the translation error measures when translating from multiple ASR input and using acoustic model scores.

Table 11: Experimental setup of the ASR experiments for Spanish.

	dev	test
WER [%]	27.3	31.9
Graph Error Rate [%]	13.7	16.6
Word graph density	~20	~30

Table 12: Translation error rates [%] on the development and test set for different ways of coupling speech recognition and translation. Translation system: Finite state transducer system.

input	dev			test		
	WER	PER	100-BLEU	WER	PER	100-BLEU
correct transcription	46.8	35.1	64.4	44.2	32.5	62.8
ASR single best	59.9	46.1	73.5	60.6	45.5	74.1
ASR word lattice	58.4	45.1	73.5	59.8	45.2	74.4

## 8 Related Experiments on Other Tasks

In this section, we will report on different experiments dealing with the use of additional language resources for speech translation. These are experiments that have been conducted on other domains than LC-STAR, namely on German–English speech translation tasks. The reason for this is that we did not have all the necessary data – such as manually aligned reference data – available on the LC-STAR task, whereas we do have this for the German–English language pair.

### 8.1 Selection of Additional Training Data Using $n$ -gram Coverage

When only a small sentence-aligned corpus is available for the training of the statistical translation models, it may be reasonable to include additional bilingual training data from other

<sup>2</sup>The density of a word graph is defined as the number of edges divided by the number of words in the (correct) sentence.

sources. Since this additional data may come from another domain and substantially differ from the original training corpus, a method for selecting relevant sentences is desirable. We introduced a relevance measure of *n-gram coverage* [Matusov & Popović<sup>+</sup> 04]. To this end, we computed the set  $C$  of *n*-grams occurring in the source part of the initial training corpus ( $n = 1, 2, 3, 4$ ). Then, for each candidate source sentence in the additional corpus, we computed a score based on the occurrence of the *n*-grams from  $C$  in that sentence and added only those sentence pairs to the initial training corpus, for which this score is sufficiently high. The score is defined as the geometric mean of *n*-gram precisions and provides a quantitative measure of how “out-of-domain” or “in-domain” the additional training data may be.

This approach was investigated on the German-English Nespole corpus consisting of spontaneous utterances in the tourist domain. By selecting relevant sentence pairs from the Verbmobil and Zeres bilingual corpora with this method, we increased the training corpus size from 3 000 to 16 000 sentence pairs. The word error rate was reduced from 60.7% to 56.1%, the BLEU score increased from 21.2% to 23.8%. Taking all of the sentence pairs from the additional corpora resulted in degradation of translation quality.

## 8.2 Part-of-Speech Information for Source Sentence Reordering

The training data sparseness may not allow for reliable word alignment training and phrase extraction, especially for language pairs with significantly different word order. In such cases, it is always of benefit to have monotonous alignments. This is not always possible due to word order differences. These differences can be reduced through initial re-ordering of the source training sentences. We proposed to perform such re-ordering based on the POS information for the source words and information about the typical sentence structure of the target language [Matusov & Popović<sup>+</sup> 04]. We used a statistical POS tagger to annotate the source (German) sentences. The re-ordering was done with context-dependent rules which are specific to the involved language pair and are based on the information about typical syntactic structures. For German-English translation we derived the rules which e.g. put verb parts (infinitives, participles, verb prefixes) standing at the end of a sentence directly after the first verb or noun/pronoun in the sentence.

More thorough linguistic considerations would have probably produced better rules, but even with these heuristics we achieved a significant improvement in the translation quality. On the Nespole corpus of spontaneous utterances we improved the WER from 56.1% to 53.7% and the BLEU score from 23.8% to 27.0%. When training on Verbmobil with only 8K sentence pairs, the improvement on the Verbmobil test corpus was from 56.3% to 52.3% WER. The fluency of the resulting translations was improved dramatically.

## 8.3 Morphological Information for Word Alignment

In [Matusov & Popović<sup>+</sup> 04], we presented two ways to solve the data sparseness problem by including morphological information into the EM training of word alignments. Existing statistical translation systems usually treat different derivations of the same base form as they were independent of each other. The information that multiple full forms of the words share the same base form is not used in the lexicon model. In our approach, these interdependencies are taken into account during the EM training of the statistical alignment models.

One approach to the exploitation of this information is the smoothing of the lexicon model with base forms [Matusov & Popović<sup>+</sup> 04, Zens & Matusov<sup>+</sup> 04] (cf. section 6). The smoothing method we apply is well known from language modelling - absolute discounting with interpolation. Another way to incorporate morphological information into the process of automatic word alignment is to make use of hierarchical representation of the statistical lexicon model [Matusov & Popović<sup>+</sup> 04, Popović & Ney 04]. Each full form is enriched with its base

form and the sequence of part-of-speech tags such that the interdependencies between the full forms sharing the same base form and POS-tags are taken into account.

We evaluated our methods on the German-English Verbmobil corpus for which manually created word alignment data is available. We obtained an improvement in alignment quality for the full training corpus containing 34k sentence pairs as well as for reduced training corpora containing 8000 and 500 sentence pairs. The Alignment Error Rate (AER) for the full training corpus has been reduced from 5.7% to 5.5% using the hierarchical lexicon approach and to 5.2% using lexicon smoothing for the direction German  $\rightarrow$  English. For the other direction, the AER is reduced from 9.9% to 9.7% using the hierarchical lexicon approach and to 9.1% using lexicon smoothing.

On the corpus of 8000 sentences the application of lexicon smoothing lead to a reduction in AER for German  $\rightarrow$  English direction from 6.2% to 6.0% and for the other direction from 11.5% to 11.1%. Using hierarchical lexicon for the training corpus of 500 sentences reduced AER from 16.7% to 15.6% for German  $\rightarrow$  English and from 21.1% to 20.9% for English  $\rightarrow$  German.

## 9 Use of WordNet

In this section the use of WordNet for statistical machine translation is investigated. WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. As mentioned above, WordNet only covers four different part of speech: nouns, verbs, adjectives and adverbs. Since WordNet was constructed from resources different then those used in this project one cannot expect that WordNet can be used for all translation tasks. It highly depends on the domain of the corpus and whether the lexicon of the corpora used in this project matches the dictionary of WordNet. Therefore, the first experiment is to determine the percentage of words in a corpus that are covered by WordNet.

### 9.1 Coverage of Words by WordNet

In the first experiment we determine the percentage of words that are covered by WordNet for the English part of the Catalan-English corpus. Since this would require the part of speech information for each word, we do an optimistic search and try for each word of the English corpus all four parts of speech. If at least one entry is found, the word counts as being covered by the dictionary of WordNet. Additionally, words are reduced to their base form using some morpho-syntactic tools provided with WordNet. This yields the following results:

Table 13: Coverage of words from the English corpus by WordNet

English Corpus	# Sentences	$\frac{\text{covered Words}}{\text{\# Words}}$
development	972	8637/13983
test	972	8418/13922

As can be derived from Table 13, the coverage of the words occurring in the English corpus by WordNet is around 60%.

### 9.2 Rescoring Using WordNet-based Features

For the following experiments we have defined different WordNet-based feature functions. Beside other features these features are used in order to rescore the sentence hypotheses of an

$N$ -best list. This allows for an easy integration of WordNet in the translation process without modifying the search criterion. The drawback of this approach is that the framework for rescoring used only allows for sentence-based features. Therefore, the feature functions must be defined appropriately on a sentence level. For each sentence hypothesis, the associated features are combined in a log linear manner. Each feature has a parameter  $\lambda$ . The optimal parameter set  $\Lambda = \{\lambda_1, \dots, \lambda_I\}$  is determined on a development set beforehand and applied to the test set afterwards.

Using WordNet entries directly as features is not feasible since this would enlarge the set of parameters such that they can not be estimated reliably (over-fitting). Instead we reduce the information provided by the WordNet entries as follows: we introduce two features functions  $f_c$  and  $f_{nc}$ . If a word hypothesis of a sentence hypothesis is covered by WordNet, we will simply increase the feature count of feature  $f_c$  for this given sentence hypothesis. If the word hypothesis is not covered, the feature count for feature  $f_{nc}$  is increased. Another set of features takes into account the part of speech information for words that are covered by WordNet. If word hypothesis is for example a verb and it is covered by WordNet, the corresponding feature function will fire. This leads to the following results. Here, 'wncov' represents the coverage feature and 'wncpv' represents the coverage feature together with the POS-WordNet feature. The leading '+' symbol means that the feature functions are combined with the baseline features.

Table 14: Translation error rates [%] on the development and test set for Catalan  $\rightarrow$  English. Results are presented for the general WordNet coverage feature (wncov) and the WordNet coverage feature together with the POS-WordNet feature (wncpv).

	dev			test		
	WER	PER	100-BLEU	WER	PER	100-BLEU
Baseline	39.7	32.2	58.7	41.6	33.1	60.2
+ wncov rescoring	38.9	31.8	58.7	40.8	32.8	59.3
+ wncpv rescoring	39.8	32.5	59.3	43.6	35.2	62.1

As we see in Table 14, the coverage feature shows slight but non-significant improvements for the word error rate and the position independent error rate. Adding the POS-WordNet feature does not improve the results; it even slightly increases the error rates.

## 10 Conclusion

We have presented several methods to incorporate additional language resources into the process of statistical machine translation. Those language resource were POS-tags, base forms, mappings of base forms to their possible full forms and POS-tags, and phrasal lexica. A state-of-the-art SMT system has been extended to exploit those resources in different steps: in training, during the translation process itself and also as an additional step in rescoring on the system output. Furthermore, we have undertaken steps towards the tighter coupling of automatic speech recognition and machine translation. We found ways to take both the acoustic and the translation models into account in the decision process.

## References

- [Arranz & Castell<sup>+</sup> 04] V. Arranz, N. Castell, J. Giménez, A. Moreno: Description of raw corpora. Technical Report Deliverable D5.3, LC-STAR project by the European Community (IST project ref. no. 2001-32216), 2004.

- [Bangalore & Riccardi 00] S. Bangalore, G. Riccardi: Stochastic Finite-State models for Spoken Language Machine Translation. In *Proc. Workshop on Embedded Machine Translation Systems, NAACL*, pp. 52–59, Seattle, WA, May 2000.
- [Bender & Zens<sup>+</sup> 04] O. Bender, R. Zens, E. Matusov, H. Ney: Alignment Templates: the RWTH SMT System. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pp. 79–84, Kyoto, Japan, September 2004.
- [Brown & Della Pietra<sup>+</sup> 93] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer: The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, June 1993.
- [Carmona & Cervell<sup>+</sup> 98] J. Carmona, S. Cervell, L. Màrquez, M. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulée, J. Turmo: An environment for morphosyntactic processing of unrestricted Spanish text. In *Proc. of the First Int. Conf. on Language Resources and Evaluation (LREC)*, Granada, Spain, 1998.
- [Casacuberta & Llorens<sup>+</sup> 01] F. Casacuberta, D. Llorens, C. Martínez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Picó, A. Sanchis, E. Vidal, J.M. Vilar: Speech-to-speech translation based on finite-state transducers. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 613–616, Salt Lake City, UH, May 2001.
- [Kanthak & Ney 04] S. Kanthak, H. Ney: FSA: An Efficient and Flexible C++ Toolkit for Finite State Automata Using On-Demand Computation. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 510–517, Barcelona, Spain, July 2004.
- [Leusch & Ueffing<sup>+</sup> 03] G. Leusch, N. Ueffing, H. Ney: A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In *Proc. MT Summit IX*, pp. 240–247, New Orleans, LA, September 2003.
- [Matusov & Popović<sup>+</sup> 04] E. Matusov, M. Popović, R. Zens, H. Ney: Statistical Machine Translation of Spontaneous Speech with Scarce Resources. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pp. 139–146, Kyoto, Japan, September 2004.
- [Och & Ney 04] F.J. Och, H. Ney: The alignment template approach to statistical machine translation. *Computational Linguistics*, Vol. 30, No. 4, 2004.
- [Papineni & Roukos<sup>+</sup> 02] K. Papineni, S. Roukos, T. Ward, W.J. Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, Philadelphia, PA, July 2002.
- [Popović & Ney 04] M. Popović, H. Ney: Improving Word Alignment Quality using Morpho-Syntactic Information. In *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pp. 310–314, Geneva, Switzerland, August 2004.
- [Ueffing & Ney 02a] N. Ueffing, H. Ney: Training Corpus Size and Statistical Machine Translation Quality. In *Poster at Informatiktage 2002 der Gesellschaft für Informatik*, Bad Schussenried, Germany, November 2002. [http://www-i6.informatik.rwth-aachen.de/ueffing/papers/Ueffing\\_TrainingCorpora\\_InfTage2002.ps](http://www-i6.informatik.rwth-aachen.de/ueffing/papers/Ueffing_TrainingCorpora_InfTage2002.ps).
- [Ueffing & Ney<sup>+</sup> 02b] N. Ueffing, H. Ney, V. Arranz, N. Castell: Overview of speech centered translation. Technical Report Deliverable D4.1, LC-STAR project by the European Community (IST project ref. no. 2001-32216), 2002.

- [Ueffing & Ney 03] N. Ueffing, H. Ney: First experimental results on baseline speech-to-speech translation systems. Technical Report Deliverable D4.4, LC-STAR project by the European Community (IST project ref. no. 2001-32216), 2003.
- [Vidal 97] E. Vidal: Finite-State Speech-to-Speech Translation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 111–114, Munich, Germany, April 1997.
- [Zens & Matusov<sup>+</sup> 04] R. Zens, E. Matusov, H. Ney: Improved Word Alignment Using a Symmetric Lexicon Model. In *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pp. 36–42, Geneva, Switzerland, August 2004.