Project no.:	FP6-506738
Project Acronym:	TC-STAR
Project Title:	Technology and Corpora for Speech to Speech Translation
Instrument:	Integrated Project
Thematic Priority:	IST

Deliverable no.: D5
Title: SLT Progress Report

Due date of the deliverable:	March 31, 2005
Actual submission date:	May 23, 2005
Start date of the project:	April 01, 2004
Duration:	36 months
Lead contractor for this deliverable:	RWTH
Authors:	RWTH: H. Ney, V. Steinbiss, R. Zens, E. Matusov; IBM: J. Gonzalez, Y.-S. Lee, S. Roukos; ITC-irst: M. Federico; UKA: M. Kolss; UPC: R. Banchs.

Revision: final, 23-May-05

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium(including the Commission Services)	
CO	Confidential, only for members of the consortium(including the Commission Services)	

Table of Contents

1 Introduction

The most important goal of the workpackage *Spoken Language Translation* was to build and test operational research systems with the following two requirements:

- The systems should be able to translate speech input rather than text; in other words, the systems must be robust with respect to speech recognition errors.
- The systems should work on real-life data rather than data collected in the research laboratories.

The translation tasks selected were:

- Chinese-to-English broadcast news :
  broadcast news of the 'Voice of America' in Chinese.
- Spanish-English EPPS:
  speeches of the European Parliament Plenary sessions (EPPS).
- English-Spanish EPPS:
  speeches of the European Parliament Plenary sessions (EPPS).

To arrive at operational systems, research along various dimensions was performed during the first 12 months of the project. Examples are:

- Models of handling word groups and phrases were studied.
- To account for the different word order in source and target languages, re-ordering models were developed.
- To cope with real-life data, specific methods were developed to handle long training sentences and to cope with the large amount of training data.
- Algorithms for the generation of word graphs and n-best lists were developed.
- Refined language models for the target language were designed that were used in a re-scoring framework.

The remainder of this deliverable document is organized as follows. In Section 2, we will describe the individual tasks in work package 1 (SLT). Section 3 will give an overview over the language resources and the evaluation campaign. The systems of the contributing partners will be described in detail in Section 4. The results of the evaluation campaign will be presented and discussed in Section 5. Finally, in Section 6, we will present some conclusions and ideas for the next evaluation.

2 The Tasks in WP1 (SLT)

In the following, the work done in months 1-12 will be reported in detail for each of the five tasks as defined on pages 50ff. and 68f. of the technical annex.

2.0.1 ITC-irst

The ITC-irst (Istituto Trentino di Cultura) baseline system from October 2004 was improved as follows. The language model (LM) software was improved in order to efficiently process very large corpora (1 billion words) and vocabularies (1 million words). Moreover, state-of-the-art LM smoothing methods were integrated and a new binary representation of the LM was defined for use by the SLT modules. Data structures and algorithms of the phrase-based models were improved in order to manage significantly larger sets of phrase-pairs.

In order to improve efficiency of training of the word alignment models, the parallel training corpus is now segmented into small chunks.

The search algorithm has been improved in order to efficiently produce word-graphs and n-best lists for re-scoring purposes. The already available minimum error procedure to estimate parameters of log-linear models was extended to search optimal parameters within n-best lists.

2.0.2 RWTH Aachen

The RWTH baseline system from October 2004 was improved in several ways. For the Chinese–English task, we made use of additional training data for both the translation model (now 200M running words in each language, from 70M) and the language model (now 660M running words, from 150M).

The translation process in our system is all lowercase. The case information is restored during the post-processing (so-called truecasing). We improved the truecasing using an n-gram-based tool trained on 660M running words.

We implemented a new decoder which has several advantages compared to our old system. It implements a flexible framework for reordering. The new decoder is more efficient and at the same time shows better performance than our old system under the same conditions. In addition, the new decoder is easy to extend, e.g. it is straightforward to use additional language models or phrase scores.

2.0.3 LIMSI

LIMSI started this task from scratch and a prototype decoder was implemented during this period. Bearing in mind their objective of integration with ASR, a word-based IBM-4 model was selected. No word classes are used for now but this is a planned extension to the decoder. Training is performed by the public tool Giza++ and a non-heuristic multi-stack decoder using one stack per source coverage set has been implemented.

This decoder is currently being tested on the EPPS tasks. Our preprocessing includes detection and separation of punctuation, as well as correction of a few inconsistencies (numbers, usage of quotes, etc.). Text is processed and produced in true-case form. First promising results on the EPPS data have been obtained.

2.0.4 UKA

The starting points for UKA's (University of Karlsruhe) baseline systems were the STTK toolkit for statistical machine translation which is being developed at our institute (Interactive Systems Laboratories), as well as corpora and various processing tools from earlier projects. Our framework combines phrase pair extraction from bilingual data, word-alignment based scoring and some minor additional models with standard n-gram language models in a reordering-enabled stack decoder.

As UKA participates in all translation tasks and languages, the baseline setup involved preparing relevant parallel corpora and training new initial translation and language models for the Chinese-English Broadcast News task and the Spanish-English EPPS task. Some changes to our main training and decoding programs were also necessary to enable training on very large amounts of data. The resulting systems, already incorporating a new phrase extraction approach developed during the first months, were used in the baseline evaluation, and in the TC-STAR evaluation at the end of the first year, where we participated in all major tracks.

2.0.5 UPC

UPC's (Universitat Politcnica de Catalunya) baseline system consists of a log-linear combination of five feature functions: i.- tuple n-gram model (translation model), ii.- target language model, iii.- word penalty function, iv.- source to target lexicon model (based on IBM model 1), and v.- target to source lexicon model (based on IBM model 1). Weights for the log-linear combination are computed via an optimization procedure, which is based on a simplex algorithm, that maximizes the BLEU over a given development set.
UPC's decoding engine for the baseline system implements a beam-search strategy based on dynamic programing. It takes into account all the five features mentioned above simultaneously and performs a monotonic search guided by the source. It also allows for using three different pruning methods: i.- threshold pruning, ii.- histogram pruning, and iii.- hypothesis recombination.

2.0.6 IBM

For the 1st TC-STAR evaluations in March 2005, we have developed 8 sub-systems according to various input conditions, all based on phrase translation models and DP-based beam search decoder.
We deploy various techniques to achieve performance improvement over the baseline system of October 2004. We incorporate IBM Model 1, word-level distortion model and word/block count penalty parameters into the decoder cost function. We apply automatically acquired reordering rules to the source language corpus for translation model training and decoding. These techniques have proven effective for translation quality improvement when evaluated on the development test data.
We have participated in 12 evaluation tracks for Spanish-to-English and English-to-Spanish translations (three input conditions in the primary and the secondary tracks) in the 1st TC-STAR evaluations. We are continuing our effort to enhance spoken language translation technology in both core translation areas and interface areas with speech recognition.

2.1 Innovative Methods

2.1.1 RWTH Aachen

During the first year of the TC-STAR project, the RWTH group has been working to improve the quality of its translation system. Beginning with the acquisition and preparation of the data of the European Parliament debates up to the final tuning of the system, the existing methods have been revised and improved, and new techniques have been developed which allow for a better translation quality.
The first step in training of the translation system is the word alignment of the sentence aligned texts. The main algorithms used for this task are described in the seminal paper of Brown et al. 1993 and several extensions were described in the years following its publication. There is software (originally developed at RWTH) publicly available to accomplish this task. In the TC-STAR project, however, we are dealing with a large amount of data and, as it is well-known to the machine translation community, the word-alignment process needs high computational resources, both in computing time and main storage. At RWTH, the first steps towards a parallel implementation of the sentence alignment algorithms have been undertaken in order to better use the capabilities offered by the increasing number of interconnected computers available in most research groups.
One of the main difficulties of machine translation is that, for most language pairs, the word order of a source sentence and its translation are not the same. For the language pair Spanish-English this mismatch is only a minor one limited only to local contexts of a few words, but for the task of Chinese-English translation, whole groups of words can change their position in a sentence which makes the generation of a correct translation a much more difficult task. A possible approach for solving this problem consists in producing a reordering in one of the languages (source or target) in the translation system in order to

better match the word order of the other language. However, allowing arbitrary reorderings is computationally not possible and some restrictions to the allowed word permutations have to be applied. At RWTH extensive investigations of several reordering constraints have been undertaken, applying them at single word level or at word-group level and an efficient, on-demand implementation of them has been produced. This is also the topic of ongoing research activity to improve the probability model underlying the permutations, which can further be enhanced by including morphosyntactic information.

Currently, there are three main systems used at RWTH for machine translation: the alignment template approach, the phrase based approach and the finite state transducer based approach. All of these go beyond single word based models as a broader context is taken into account while translating a sentence. It is a well known fact that the output of current translation systems is far from perfect, and currently, one approach that improves the quality of the produced translations is to examine not only the first translation proposed by the system (i.e. the translation with maximum probability) but to produce a list of n best translations (i.e. the n translations with highest probability) and to use additional models in order to decide which one between them is the most appropriate one. Efficient algorithms to extract these n-best lists have been investigated and implemented at RWTH for each of the available translation systems.

Once the n-best lists have been produced, we must choose the appropriate models in order to better select the best translation for a given source sentence. Experiments conducted at RWTH show that two of the better models for this task consist in the inclusion of an additional IBM-Model1 score and the use of clustered language models. These try to cluster the different type of sentences according to their syntactic structure and apply a different type of language model to each of the sentence classes. This approach originates from the observation that different kinds of sentences have a different syntactic structure and thus specialized language models perform better than a global one which tries to cover all types of sentences.

2.1.2 UPC

UPC's contribution to task 1.2 may be summarized as follows. In activity 1.2.i, UPC has invested a significant effort on developing and evaluating its own defined translation model. This model, which is referred to as the tuple n-gram model, constitutes actually a language model of bilingual units called tuples. The innovative issue about our implementation has to do with using this translation model in the log-linear approach.

Regarding activity 1.2.ii, UPC has mainly been working on developing a strategy for handling verbal structures in SLT. In this way, a rule-based algorithm for detecting verbal forms in both Spanish and English have been developed and is fully operational. At this moment, research in this area is devoted to evaluate different strategies for classifying and representing the identified verbal forms. These strategies for handling verbal structures will be studied and compared by evaluating their impact on both alignment quality and translation accuracy.

Finally, regarding activity 1.2.iii, a decoding engine for our tuple n-gram translation model has been developed and is fully operational. This decoder implements a beam-search strategy based on dynamic programming and allows for three different pruning methods: i.- threshold pruning, ii.- histogram pruning, and iii.- hypothesis recombination. It also allows for multiple feature simultaneous decoding and, at this moment, performs a monotonic search guided by the source.

2.1.3 UKA

UKA's statistical machine translation system has seen substantial improvements from the use of new innovative methods developed during the first year of TC-STAR. We developed a new approach to extract phrase translation pairs from bilingual data, where a constrained word-to-word alignment is optimized for an entire sentence pair. In this approach, which we call PESA, candidate phrase translations are scored by summing, for each source phrase word, only over the IBM1 probabilities for words inside the target phrase candidate, and for words outside the source phrase only over IBM1 probabilities for the words outside the target phrase candidate. This new phrase extraction method significantly improved the translation quality across all domains and has superseded previous phrase translation sources based on maximum mutual information or Viterbi pathes of word-to-word alignments. Another notable improvement came from splitting training sentences at selected splitpoints during statistical lexicon training, based on the incremental lexical probabilities. This approach improved statistical lexicon quality over standard IBM-style training.

2.1.4 IBM

IBM spoken language translation system deploys several new techniques for decoding and corpus processing: We incorporated IBM Model 1, word-level distortion model and word/block count penalty parameters into the decoder cost function. We also apply automatically acquired reordering rules to the source language corpus for translation model training and decoding. We have also developed and implemented a novel sentence alignment algorithm which uses the IBM Model 1 Viterbi alignment score as the key component, enabling us to quickly develop a new system on various parallel corpora.

2.2 Integration of Recognition and Translation

The goals of this task are to analyse speech related aspects of the translation process and to define a suitable interface between recognition and translation, and to investigate methods for making the translation robust against speech recognition errors. Contributing partners are ITC-irst, RWTH, UKA, LIMSI, and IBM.

So far, work has focused on the information to be passed from the recognizer to the translation components, where partners experimented with various forms of n-best and word graph input, using additional data from recognizers (rich transcription metadata), and global decoding strategies.

In the remaining time scheduled for this task, we intend to produce a formal, extensible interface specification in the context of the overall TC-STAR architecture, to allow pluggability of components from different groups. In particular, automatic and fixed segmentation of audio streams, score normalization, and generic lattice and metadata formats will need to be addressed.

The following sections describe in more detail activities carried out by the individual partners until month 12.

2.2.1 ITC-irst

During the first year, we started to develop statistical models and search algorithms capable to translate multiple input hypotheses, represented either by n-best lists or word-graphs. In particular, two approaches have been investigated for integrating ASR and SLT. 1) A list of N-best transcription hypotheses is generated by the ASR system and translated by a text-based SLT system; the set of translation hypotheses are re-scored and re-ranked according to the scores provided by both the ASR and SLT systems. 2) A word-graph is generated by the ASR system and decoded by means of a specialized search algorithm which extends the decoder implemented for the text-based SLT system. Specific log-linear models are defined for both cases.

Experiments on speech translation based on n-best and word-graph based decoding have been performed, until now, on the BTEC domain, from Italian to English. In addition, optimization techniques including speech recognition and machine translation were investigated in order to optimize global performance.

2.2.2 RWTH Aachen

In our first experiments on the integration of speech recognition and machine translation we followed a statistical translation approach implemented with weighted finite-state transducers.

The training can be summarized as follows: we produce word alignments for the training corpus using the IBM 1-5 translation models. Then, the target sentences are reordered to monotonize the word alignment. Now, we produce a sequence of new tokens, each of them consisting of a source word and the aligned target words. On this 'bilanguage' sequence, we train an n-gram language model represented as a weighted finite-state transducer.

This transducer is used to produce monotone translations. The resulting hypothesis is permuted with suitable constraints and rescored with a target language model. The great advantage of this transducer-based approach is that the translation of lattices is straightforward, at least theoretically. In translation, we included scaled acoustic scores of the ASR lattices in the global decision process.

With the goal of tighter coupling of the ASR and the SLT systems we also optimized the interface between the systems. To this end, the corpus preprocessing (which included tokenization, spelling of digit numbers, dates, expansion of abbreviations, etc.) was identical for both systems.

RWTH participated in the first official TC-STAR evaluations with this integrated system. On tasks with smaller training corpora we were able to significantly improve the translation results by using ASR lattices instead of single-best ASR hypothesis as input. In the future, we will implement the lattice translation for our primary phrase-based SLT system.

2.2.3 UKA

We implemented a first prototype of an integrated speech translation system to study the optimal coupling between speech recognition and translation components. On the translation side, we experimented with incorporating more information from the recognizer, such as source language and acoustic scores, into the translation decoder. This can be done for single-best hypotheses as well as n-best lists and word lattice input. While using lattice input improved translation results over using single-best input when using monotonic decoding, more design and implementation work is needed to enable our current decoder to handle arbitrary word graph input and word reordering at the same time.

To improve the robustness of an integrated system, we also developed a module capable of automatically cleaning some disfluencies from spoken (conversational) language. Based on a statistical source-channel approach, this module tries to remove repetitions, hesitations and false starts from speech hypotheses before passing them on to the translation module.

2.3 Human-Supplied Knowledge

The main goal of task 1.4 is to develop specific methods for incorporating human-supplied knowledge into SLT systems in order to either improve translation accuracy, or reduce the amount of required training data. This particular kind of knowledge may be extracted from available bilingual and monolingual information sources such as conventional bilingual dictionaries, morphologic analyzers, morpho-syntactic analyzers and syntax models, among others. Although recent efforts in this direction have not yet produced satisfactory results [Och & Gildea^{+} 04b], integration of human-supplied knowledge into SLT systems continues to be a promising approach but a challenging task.

Efforts in the human-supplied knowledge task are organized into two specific sub-tasks: 1.4.i.- investigation about how existing resources might be exploited for SLT, and 1.4.ii.- integration of human-supplied knowledge into SLT. Four of the partners of TC-STAR consortium (UPC, RWTH, UKA and ITC) have participated and dedicated a significant amount of work in this area, UPC being the institution responsible for this task. Contributions of each of these four partners are present in detail here.

2.3.1 UPC

During the first five months (M7 to M11) of research, most of the UPC's effort has been focused on sub-task 1.4.i; more specifically, in collecting and generating human-supplied knowledge resources. From this activity, the following three results can be reported:

- Bilingual resources extraction from WordNet: this includes a semantic-based bilingual dictionary of 172,260 entries, and multi-word expression lists for both English (66,940 entries) and Spanish (19,633 entries). Theses resources contain only lemma forms so they must be used along with morphologic analyzers and synthesizers.

- Morphologic and syntactic annotation of EPPS data: an annotated version of the EPPS corpus has been generated for which chunk, lemma and part of speech information has been added to each token. Some language analysis tools from the TALP research center [FreeLing 05, SVMTool 05] were used for annotating the corpus.

- Verbal form identification and classification: a rule-based algorithm for detecting verbal forms in both Spanish and English have been generated. These detected verbal forms are classified into groups according to the main verb basic form [Gispert & Mariño 05].

Regarding sub-task 1.4.ii (integration of human-supplied knowledge into SLT), no final results have been obtained yet. At this point, efforts in this area are still in progress. More specifically:

- Use of bilingual dictionary in SLT: the impact of using a bilingual dictionary at each of three steps of the translation process is being tested. These steps are: i.- training, by using the bilingual dictionary for smoothing and/or modifying the translation model probabilities; ii.- decoding, by using the bilingual dictionary as a feature function; and iii.- post-processing, by using the bilingual dictionary for re-ranking n-best lists.

- Use of verbal form classes: the impact of replacing verbal forms with their corresponding classes is being tested on both alignment quality and translation accuracy. Notice that this procedure requires selecting a surface verbal form from the corresponding class during (or after) decoding. This last problem is going to be handled statistically too, except for those cases for which final verbal forms are not present in the training data. In these particular cases, a morphologic synthesizer will be used.

- Use of multi-word expressions: the impact of identifying and using multi-word expressions as single tokens is being tested on both alignment quality and translation accuracy. Notice that this procedure requires morphologic analysis and synthesis since multi-word expressions extracted from WordNet contain only lemma forms.

Finally, some further ideas for incorporating human-supplied knowledge into SLT are to be tested. The following can be mentioned among the most relevant: i.- decoding or re-ranking n-best lists by incorporating a POS-tag translation model feature, ii.- decoding or re-ranking n-best lists by incorporating a POS-tag target language model feature, and iii.- extending the verbal form identification and classification method to other common syntactic constructions.

2.3.2 RWTH Aachen

In the past months, the RWTH research concentrated on using morpho-syntactic information to improve word alignment quality and translation results. Such resources are especially valuable if the amount of the available training data is rather small. In our preliminary experiments on a Spanish to English translation task [Popovic & Ney 05], we used manually created phrasal lexica, as well as base forms of Spanish words and part-of-speech information. The phrasal lexicon was used to increase the existing training data. The morpho-syntactic information was used for Spanish verb expansions. With these additional knowledge sources, reasonable translation quality was achieved with only 1,000 sentence pairs of in-domain data in training.

For the first official TC-STAR evaluation relatively few human-supplied language resources were available. We made use e.g. of the lists of parliamentary speakers in order to translate them with correct spelling.
Analysis of translation errors showed that a substantial number of errors in translation from Spanish to English resulted from omission of Spanish pronouns like ”we” or ”I”, since the verb form (given by a suffix like ”mos”) already gives a hint on the person, to which the verb relates. As a result, a Spanish verb is often aligned either only to the English pronoun or English verb, which causes frequent omissions of verbs or incorrect ”guessing” of the pronouns in the resulting translations. The correct solution for this problem would be to artificially insert Spanish pronouns only where this is needed. However, to this end a reliable part-of-speech tagger is required. A statistical tagger was not yet available, since no manually-tagged data from the EPPS domain exists to train such a tagger.
When translating from English into Spanish, the error rates are higher, since the system often produces correct words but with wrong morphology, e.g. wrong suffixes and endings of verbs and adjectives. In the future, we plan to handle this problem by rescoring/modifying the translation output using manually provided morpho-syntactic information.

2.3.3 UKA

UKA’s main phrase-based SLT system makes limited use of human-supplied knowledge in the form of shallow translation rules which cover mainly expressions involving numbers, such as date and time expressions and money amounts. These rules are currently applied as a preprocessing step, before the actual decoder sees the source language input.
While this setup has already improved translation quality, we have started to investigate how to better integrate this and other additional human or linguistic knowledge into our statistical framework. To this end, we have experimented, in a currently separate translation system, with using a standardized and linguistically enriched English as an intermediate step in the translation process. This enables us to exploit sources of morphologic and semantic information such as WordNet to provide additional features which are useful for various purposes, such as word reordering and word sense disambiguation.

2.3.4 ITC-irst

The ITC-irst phrase-based log-linear SLT system has been extended to accept lexical rules provided by an external source of knowledge. A rule provides suggestions for alternative translations of phrases to the statistical model. The rules, which are weighted according to their reliability, can be used at decoding-time to improve translation quality. These rules are set at run-time according to the input sentence and can be context-dependent. Simple translation rules were developed for Spanish-to-English to model fixed expressions, such as time expressions, money amounts, news headings, etc.
It is worth noticing that, from the point of view of the model, rules are considered as an additional feature function. Hence, the decoder can be trained in a way to optimize the weight associated to rules as well as to other statistical models.

2.4 Technology Support

Although the main activity for this task had been planned for the months 13-18, RWTH has put some effort into this task. For IBM and ITC-irst, technology support has been limited to providing feedback on the UIMA architecture (WP5).

2.4.1 RWTH Aachen

For languages with only small (local) differences in word order, we have a fully monotonic search organization, which, under moderate pruning, allows us to produce translations at a speed of about 0.5 second

per sentence using less than 1 GB of memory.
For languages with significant differences in word order we translate monotonically, but take a permutation graph as input. The permutation graph is constrained using an efficient and flexible reordering framework. We can then additionally automatically identify source word sequences which should always be translated monotonically and keep the word order of these sequences in search. Such sequences can be determined using word alignment information. This allows us to make the translation process significantly faster without a loss in translation quality.

3 Evaluation Campaign and Language Resources

3.1 Overview

In order to support the performance progress and the exchange of knowledge across partners, an evaluation campaign was carried out in month 12 of the project.
The translation tasks selected were: Chinese to English for broadcast news ('Voice of America') and Spanish to English and English to Spanish for speeches of the European Parliament Plenary Sessions (EPPS).
To study the effect of both recognition errors and spontaneous speech phenomena, particularly for the EPPS task, three types of input to the translation system were studied and compared:

- **ASR:** the output of automatic speech recognizers, without using punctuation marks.
- **verbatim:** the verbatim (i.e. correct) transcription of the spoken sentences including the phenomena of spoken language like false starts, ungrammatical sentences etc. (again without punctuation marks).
- **text:** the so-called final text editions, which are the official transcriptions of the European Parliament and which do *not* include the effects of spoken language any more. (Here, punctuation marks are included.)

By comparing the translations of ASR and verbatim input, we can see the effects of speech recognition errors on the translation quality. Similarly by comparing the translations of verbatim and text input, we can see the effects of spoken language in the translations. For the Chinese broadcast news task, the only difference between verbatim and text input concerns the punctuation marks.

In addition to the TC-Star partners (IBM, ITC-irst, RWTH, UKA, UPC), two external research groups participated in the evaluation campaign:
– JHU: a joint team of Johns-Hopkins University and Cambridge University,
– UPV: Universidad Politecnica de Valencia.
The results of the evaluation campaign and the research systems involved were presented at a TC-Star workshop in Trento (in April 2005), in which also the external groups participated.

The evaluation campaign was organized as follows. The project partners agreed on a list of LDC corpora to be used as training data for the Chinese-English task. For the EPPS task, the training data mainly included final text editions of parliamentary speeches and their translations (see also Section 3.3). These were extracted from the WWW and sentence-aligned by RWTH. The preparation and distribution of development and test data, as well as evaluation was handled by ELDA (for more details, refer to the deliverable D7). In the first stage of the evaluation campaign, automatic speech recognition systems were evaluated. After the evaluation for automatic speech recognition was over, ELDA provided test data for the SLT evaluation. The input for the ASR evaluation conditions was the combination of the outputs from different ASR systems using ROVER. At the end of the SLT evaluation, the project partners submitted the outputs of their systems to ELDA. The submitted translations were scored by ELDA using the same set of evaluation tools and reference translations. In such way, all translation systems are evaluated in a consistent way so that the SLT systems can be directly compared with respect to their performance.

3.2 Data for Chinese Broadcast News

In this and the next section, we will give some statistics of the training, development and test corpora used in the evaluation campaign. To describe the corpora, we will make use of the following quantities:

- **Sentence Pairs**: the number of bilingual sentence pairs,
- **Running Words + Punct. Marks**: the number of running words (tokens) with punctuation marks counted as words,
- **Running Words**: the number of running words (tokens) without punctuation marks,
- **Vocabulary**: the number of distinct words (types),
- **Singletons**: the number of words that occur only once in the training corpus,
- **OOV rate**: out-of-vocabulary rate which is the percentage of running words in the test corpus that did not occur in the training corpus.

The bilingual training data for the Chinese-English task consists of a large variety of corpora from LDC. The data are quite heterogeneous. The most important domains include some newswire texts, the Hong Kong Hansards, and documents of the United Nations. Table 1 shows the statistics of the training data. There are about seven million sentence pairs with about 200 million running words in each language.
The English side of the bilingual training corpus is used to train the target language model. As additional monolingual data, the English Gigaword corpus is available via LDC. In Table 2, the corpus statistics and the OOV rates are shown. Also, the language model perplexities using the SRI toolkit are shown for the text and verbatim development data.
The development and the evaluation data are broadcast news of 'Voice of America' radio programs in Chinese. For the verbatim task, the punctuation marks are removed. The corpus statistics for the development and the evaluation data are shown in Table 3.

Table 1: Chinese-to-English: Corpus Statistics of the Bilingual Training Data.

		Chinese	English
Train	Sentence Pairs	7.1M	
	Running Words + Punct. Marks	199M	213M
	Running Words	173M	191M
	Vocabulary	223K	351K
	Singletons	100K	162K

Table 2: Chinese-to-English: Corpus Statistics of the Monolingual Training Data (Target Language).

	Verbatim	Text
Running Words	537M	607M
OOV rate	0.8%	0.6%
Trigram perplexity	242	241
Fourgram perplexity	229	231

Table 3: Chinese-to-English: Corpus Statistics of the Development and the Evaluation Data.

		Chinese	English
Dev	Sentences	525	
	Running Words + Punct. Marks	15 173	14 513
	Running Words	13 265	13 235
Test	Sentences	494	
	Running Words + Punct. Marks	13 920	-
	Running Words	12 508	-

3.3 Data for EPPS Task

European parliamentary speeches and their translations were used as training data for the EPPS task. The training corpus consisted of final text editions of parliamentary sessions from April 1996 to October 2004. In addition, the corpus of verbatim transcriptions of the sessions from May to October 2004 was made available. This corpus of verbatim transcriptions corresponds to the data that was used to train the speech recognition systems in WP2. The corpus statistics for the training corpus of final text editions is given in Table 4. The bilingual sentence-aligned training corpus includes over 30M running words of text in each language with large vocabularies of about 140K words for Spanish and over 90K words for English.

Table 4: Training, development and test corpus statistics for the EPPS Task (Final Text Editions).

		Spanish	English
Train	Sentence Pairs	1 207 740	
	Running Words + Punct. Marks	34 851 423	33 335 048
	Running Words	31 360 260	30 049 355
	Vocabulary	139 587	93 995
	Singletons	48 631	33 891
Dev	Sentences	1011	1011
	Running Words + Punct. Marks	25 717	26 009
	Running Words	22 946	23 173
	Distinct Words	3668	2982
	OOV Words	41	24
Test	Sentences	840	1094
	Running Words + Punct. Marks	22 756	26 885
	Running Words	20 427	24 103
	Distinct Words	3844	3744
	OOV Words	40	102

No sessions beyond October 16, 2004 were allowed to be used for training. The sessions of October 26 and 27, 2004 were used as development data. Most of the development data was made available by ELDA about one month before the evaluation period. The test data contained excerpts of parliamentary sessions of November 16 through 18, 2004. Table 4 also presents the statistics for the development and test corpora which were used for the text input evaluation.

Table 5 gives an overview of the development and test data used for verbatim input evaluation. ASR transcripts of these corpora were used as input for the ASR evaluation condition.

Note that, in the development and test sets, the English and Spanish sentences are *not* translations of

Table 5: Development and test corpus statistics for the EPPS task (Verbatim Transcriptions).

		Spanish	English
Dev	Sentences	2643	1750
	Words	20 289	23 407
	Vocabulary	2932	2566
	OOV Words	46	59
Test	Sentences	1073	792
	Words	18 896	19 306
	Vocabulary	3302	2772
	OOV Words	145	44

each other (this is different for the training set, where we have *bilingual sentence pairs*). Therefore, the number of English and Spanish sentences can be different.

4 The Systems Used in Evaluation Campaign

4.1 The ITC-irst System

The ITC-irst Spoken Language Translation (SLT) system [Bertoldi & Cattoni$^+$ 04] implements an extension of the IBM Model 4 as a log-linear interpolation of statistical models, which apply probabilities at the level of *phrases*. The interpolation involves the following models: the lexicon, the distortion, the fertility and the target language model. The use of phrases rather than words is a mean to cope with the limited context that Model 4 exploits to guess word translations (lexicon model) and word positions (distortion model).

The architecture of the system at run time is shown in Figure 1. After a preprocessing step, the sentence in the source language is given as input to the decoder, which implements a dynamic programming algorithm and outputs the best hypothesis in the target language; the actual translation is obtained by a further postprocessing.

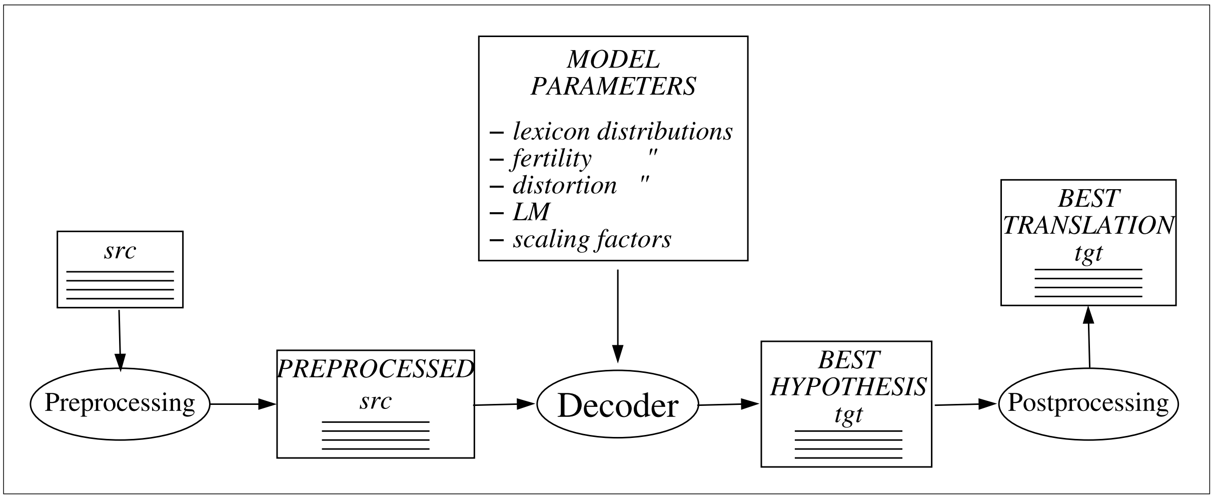

Figure 1: The architecture of the ITC-irst SMT system at run time: after preprocessing, the input sentence is sent to the decoder that, given the model parameters, searches for the best hypothesis. A final postprocessing step provides the actual translation.

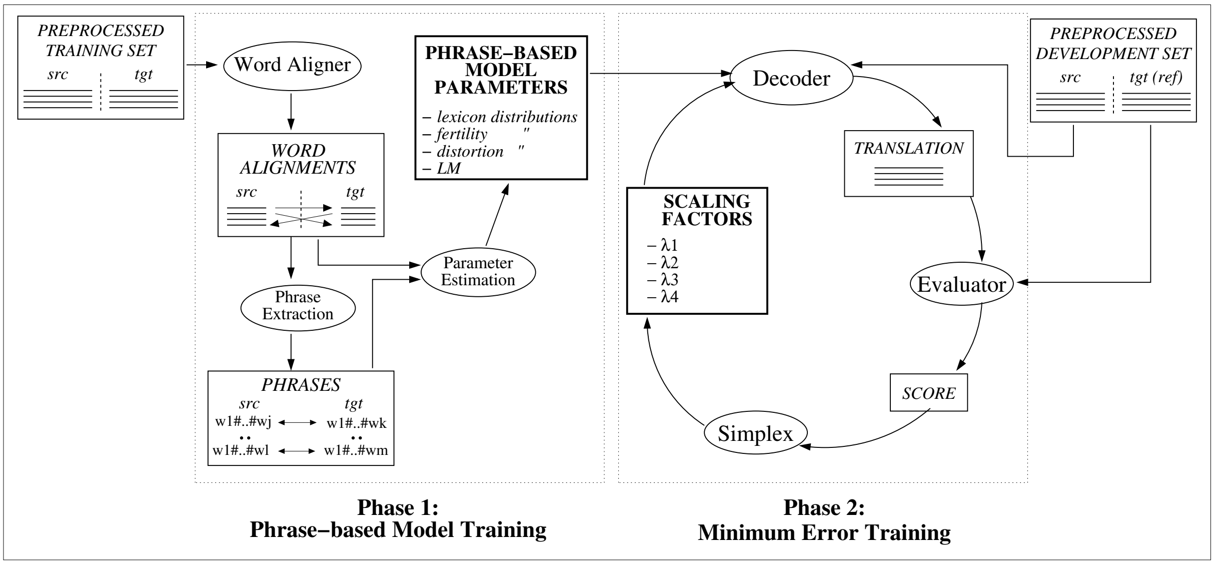

Figure 2: The two-phase architecture of the training system: first, the distributions of the components of the phrase-based model are estimated by means of alignments (left side). Then, the scaling factors of the components are computed by a minimum error training loop (right side).

Preprocessing and postprocessing consist of a sequence of actions aiming at normalizing text and are applied for preparing both training data and text to be translated (see Section 4.1.1). The same steps are applied to both source and target sentences, according to the language. Input strings are tokenized, and put in lowercase. Casing information on the output text is recovered through a maximum entropy-based text tagger.
Parameters of the statistical translation model can be divided into two groups: the parameters of each basic phrase-based model and the weights of their log-linear combination. Accordingly, the training procedure of the system, shown in Figure 2, consists of two separate phases.
In the first phase, distributions of the components of the phrase-based models are computed starting from a parallel training corpus. After preprocessing, Viterbi alignments from source to target words, and vice-versa, are computed by means of the GIZA++ toolkit [Och & Ney 00]. Phrase pairs are then extracted taking into account both direct and inverse alignments, and the phrase-based distributions are estimated.
In the second phase the scaling factors of the log-linear model are estimated by a *minimum error training* procedure [Cettolo & Federico 04].

4.1.1 Corpus Preprocessing

The same preprocessing steps are applied to the three languages (Chinese, Spanish and English) involved in the two translation directions:

Tokenization. Words are separated from punctuation. Acronyms and abbreviations are also managed in this step.

Splitting. Long sentences represent a problem for training as well as decoding, because sentence length impacts on computational complexity and affects precision in estimation of parameters. To overcome this problem, aligned long sentences are first split into shorter portions (chunks) and then aligned taking into account the IBM Model-1 statistics. Candidate splitting points are selected according to strong punctuation and sentence length.

Case normalization. All words are put in lowercase. For Asian languages like Chinese this step is applied only to non-Asian words (for example English words included in the sequence of Chinese ideograms).

Sentence filtering. Parallel sentences in the corpus are filtered according to their length. Two criteria are applied: (1) absolute-length pruning (when the length of one of the source/target sentences overcomes a given threshold), and (2) relative-length pruning (when the ratio between the length of the source and target sentences overcomes a given threshold). The former criterion is used to reduce alignment errors and training time, the latter to remove ill-formed sentence pairs.
Preprocessing of Chinese includes also a *segmentation* step in which the sequence of ideograms is separated into words. The tool used for segmentation is *ICTCLAS* (Institute of Computing Technology, Chinese Lexical Analysis System), a publicly available monolingual resource [ICTCLAS 05].
Concerning preprocessing of Spanish data, first the software tools provided by RWTH in November 2004 have been applied [RWTH Tools 04]; then, some refinements and an integration of our local preprocessing tools have been performed.

4.1.2 Ongoing Research Work

With respect to the system employed for the baseline evaluation held in October 2004, new versions of software modules and the decoder have been used which are able to handle huge amounts of data. This has made possible the estimation and the use at run time of models much larger, and then better, than those used in the past.
Concerning research topics, two main issues are currently investigated at ITC-irst: (i) handling multiple hypotheses and (ii) adding rules to the pure statistical translation process.
The ITC-irst SLT system employed in this first evaluation shows a typical cascaded structure, speech recognition followed by machine translation, where the translation module takes a single best recognition hypothesis as input and performs standard text-based translation. Last updates of the translation decoder allows to receive from speech recognition supplementary information in terms of n-best lists, word graphs and confusion networks. In addition, it is now able to output multiple translation hypotheses as word graphs. This kind of information can be effective for improving translation quality if employed properly [Zhang & Kikui^{+} 04, Och & Gildea^{+} 04b], and in the next future we are going to move in that direction.
Recently, the ITC-irst SLT decoder has also been made able to cope with explicit translation rules. This should improve the quality of the translation in general, but especially for the Chinese-English pair, since it will allow to force both verbatim translation of proper names and the correct translation of some specific patterns whatever segmentation of the Chinese source string is provided.

4.1.3 System Setup for the Evaluation

ITC-irst submitted runs for the Spanish-to-English (Es→En) and for the Chinese-to-English (Chi→En) translation directions, and for all the three input sets, namely the final text edition (FTE), the verbatim transcription (VRB) and the recognizer output (ASR).
Four different best systems have been developed for the six above-mentioned primary conditions. Table 6 summarizes their main differences; the table legend follows.

Models: the FTE and VRB/ASR systems of each language pair share the training data (see Section 4.1.4), but different alignments have been used for the training of their models. In fact, for the FTE systems the output of the GIZA++ tool has been employed, while for the VRB/ASR systems, the punctuation tokens have been removed from the original alignments.

Scaling factors: the scaling factors of the log-linear translation model of the "Es→En" pair have been estimated through the minimum error training procedure, applied for maximizing the BLEU score

Table 6: ITC-irst systems employed for the first TC-star evaluation.

Language pair	Set	Models	Scaling factors	Penalty length	Rules
Es→En	FTE	punct	FTE	N	Y
	VRB/ASR	no punct	VRB	N	Y
Chi→En	FTE	punct	no	Y	N
	VRB/ASR	no punct	no	Y	N

Table 7: Bilingual resources used by ITC-irst.

LDC Code	**Title**
LDC2002E17	English Translation of Chinese Treebank
LDC2004E09	Hong Kong Hansard Parallel Text, aligned at the sentence level
LDC2003E25	Hong Kong News Parallel Text, sentence-aligned (1997-2003)
LDC2002L27	Chinese English Translation Lexicon version 3.0
LDC2002E58	Sinorama Chinese-English Parallel Text
LDC2002E18	Xinhua Chinese-English Parallel News Text Version 1.0 beta 2
LDC2003E14	FBIS
LDC2003E04	Multiple-Translation Chinese Corpus Part 3
LDC2000T47	Hong Kong Laws Parallel Text (1997-2000)

on development sets. The FTE development set has been employed for the FTE system; the VRB development set has been employed for the VRB and ASR systems. For the “Chi→En” pair, no estimation procedure has been performed, and weights equal to 1 have been used.

Penalty length: in addition to the baseline models (target language, lexicon, distortion and fertility models) a penalty length model has been introduced in the definition of the log-linear translation model. It has been activated only for the “Chi→En” pair (weight equal to 1).

Rules: so far, only a limited number of rules has been shown to be effective for the “Es→En” pair, while no rule has been designed for the “Chi→En” pair.

4.1.4 Training Data

Chinese→English translation.
The bilingual resources used are shown in Table 7. In addition, the `LDC2004E12` corpus *UN Chinese-English Parallel Text Version 2* has been included to train the IBM Model-1 used only to split long sentences.
As far as the monolingual resources are concerned, the language model for English has been trained on the English side of the same bilingual resources described above plus the Xinhua portion of the `LDC2003T05` *Gigaword* corpus.
Spanish→English translation. For this direction the only corpus used for both bilingual and monolingual training has been the `EPPS` corpus, the version provided by RWTH in November 2004, Latin-1 encoding [RWTH Corpus 04].

4.1.5 System Development

On the development sets, different system setups have been evaluated, in order to choose the one with the best performance to be used for the evaluation.

Spanish→English translation. Since the ASR development set contains less segments than the FTE and VRB sets, the BLEU score have been measured at the document level instead of segment level; moreover, the computation was case-insensitive.

Table 8 reports performance of the tested systems on the respective development sets.

Table 8: Development of ITC-irst systems for the "Es→En" pair. Performance is given in terms of BLEU score.

scaling factors:	1 . . . 1	1 . . . 1	min. err. train.	min. err. train.
rules:	no	yes	yes	yes
search:	monotone	monotone	monotone	non-monotone
FTE	54.95	55.73	58.23	58.60
VRB	45.31	45.58	47.38	47.86
ASR	na	44.12	44.15	44.28

Chinese→English translation. Actually, for the evaluation on the "Chi→En" pair we have not performed any real development stage. With respect to the baseline evaluation, we have employed much more training data, but no rules have been been applied and no estimation through minimum error training has been performed. Table 9 reports performance of the FTE and VRB/ASR systems on the respective development sets, for both the case-insensitive and case-sensitive conditions.

Table 9: Performance (BLEU score) of ITC-irst systems for the "Chi→En" pair on the development sets.

	case-insensitive	case-sensitive
FTE	15.10	13.42
VRB	13.59	11.97

4.1.6 Evaluation

A brief description follows of the systems employed for the runs ITC-irst submitted in the first evaluation.

Spanish→English direction:

FTE best: scaling factors estimated on the FTE development set by maximizing the BLEU score; use of rules; non-monotone search

FTE alternative: as FTE best but with no estimation of scaling factors

VRB best: as FTE best but with scaling factors estimated on the VRB development set

VRB alternative: as VRB best but with no estimation of scaling factors

ASR best: scaling factors estimated on the VRB development set by maximizing the BLEU score and normalized; use of rules; non-monotone search

ASR alternative 1: as ASR best but with non-normalized scaling factors

ASR alternative 2: as ASR best but with no estimation of scaling factors

Chinese→English direction:

FTE best: baseline system for data with punctuation + penalty length model

FTE alternative: as FTE best but no penalty length model

VRB/ASR best: baseline system for data without punctuation + penalty length model

VRB/ASR alternative: as VRB/ASR best but no penalty length model

Table 10 allows a comparison of the baseline system from October 2004 and the system employed in the March 2005 evaluation. Results are given for the March 2005 development set and the official March 2005 test set, both for case-insensitive and true-case conditions, on the Chinese-English direction. Improvements range from 15 to 26%.

Table 10: Chinese-English: translation results (BLEU score) for the TC-Star development and the March 2005 evaluation test sets.

Chinese → English		Oct04 system		Mar05 system	
		case-sens.	case-insens.	case-sens.	case-insens.
DEV05	FTE	10.91	11.99	13.42 (+23.0%)	15.10 (+25.9%)
	VRB	10.04	11.49	11.97 (+19.2%)	13.59 (+18.3%)
TST05	FTE	10.89	12.28	12.56 (+15.3%)	14.41 (+17.3%)
	VRB	10.08	11.68	12.03 (+19.3%)	14.18 (+21.4%)
	ASR	9.69	11.27	11.49 (+18.5%)	13.65 (+21.1%)

4.1.7 Conclusion

With respect to the baseline evaluation, for this first evaluation we have exploited the following progresses:

- new data structures employed by training scripts and by the decoder able to handle much more training data and larger models
- use of explicit translation rules (only in Spanish→English direction)

This has allowed us to improve the quality of our Chinese→English translation system, and to build from scratch a good-quality Spanish→English translation system.
Our current decoder is already able to receive as input multiple hypotheses from the speech recognizer (both as n-best list and confusion network) and to output word graphs. This will allow the rescoring of multiple translation hypotheses on the basis of new and/or more refined information sources. Moreover, for the final evaluation we will design new translation rules for the various language pairs.

4.2 The RWTH System

In this section, we will describe the RWTH statistical machine translation system that was used in the evaluation. We participated in the Chinese → English task and the Spanish ↔ English (EPPS) tasks. For each task there were three conditions: Text, Verbatim and ASR.

Table 11: Corpus statistics of the Chinese–English bilingual training data.

		Text		Verbatim/ASR	
		Chinese	English	Chinese	English
Train	Sentences	7.1M			
	Running Words	199M	213M	173M	191M
	Vocabulary	223K	351K	223K	351K
Dictionary	Entries	82K			

4.2.1 Translation Model

In this section, we give a brief description of the translation system. The source language ('French') sentence will be denoted as $f_1^J = f_1 \dots f_j \dots f_J$ and the target language ('English') sentence will be denoted as $e_1^I = e_1 \dots e_i \dots e_I$.

We use a phrase-based translation system similar to [Zens & Ney 04]. The posterior probability $Pr(e_1^I | f_1^J)$ is modeled directly using a weighted log-linear combination of various models: a trigram language model, a phrase translation model and a word-based lexicon model. The latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally, we use a word penalty and a phrase penalty.

We extended the monotone search algorithm from [Zens & Ney 04] such that reorderings are possible. In our case, we assume that local reorderings are sufficient. Within a certain window, all possible permutations of the source positions are allowed. These permutations are represented as a reordering graph, similar to [Zens & Och$^+$ 02] and [Kanthak & Vilar$^+$ 05]. Once we have this reordering graph, we perform a monotone phrase-based translation of this graph.

Our search algorithm generates a word graph of the most likely translation hypotheses. Out of this word graph we extract the N-best translation candidates and compute additional model scores for each of them. For this evaluation, we performed rescoring with IBM model 1 and additional language models [Hasan & Ney 05]. Similar to [Och 03], the model scaling factors are optimized with respect to the final translation quality measured with the BLEU score [Papineni & Roukos$^+$ 02].

4.2.2 Data Resources

The preprocessing consists of several steps: The corpus is tokenized and converted to lowercase. We split long sentences using the algorithm described in [Xu & Zens$^+$ 05]. A rule-based categorization and translation of number and date expressions is performed. During the training procedure, number and date expressions are replaced with special symbols. During the translation process, the rule-based translation of the actual number or date is inserted via the alignment information.

During the postprocessing we restore the case information (truecasing) and do some text normalization. Our truecaser uses the SRI disambig tool [Stolcke 02]. The text normalization that is done during the postprocessing includes removing double commas etc. and unifying monetary amounts, abbreviations and so on.

Some task specific preprocessing is done: The Chinese text is segmented using the LDC segmentation tool. For the EPPS verbatim task, categorization is not used.

For the Chinese–English task we make use of almost all bilingual corpora provided by LDC. For language modeling in the Chinese–English task, we use the English part of the bilingual training corpus and in addition some parts of the English GigaWord corpus. The total language model training data consists of about 600 million running words.

The corpus statistics of the preprocessed Chinese–English training corpus are shown in Table 11. We show the statistics for the text condition as well as for the verbatim/ASR condition.

The corpus statistics of the preprocessed Spanish–English EPPS training corpus are shown in Table 12. Again, we distinguished the corpora for the final text editions (Text) and the verbatim transcriptions

Table 12: Corpus statistics of the EPPS Spanish–English training data.

		Text		Verbatim/ASR	
		Spanish	English	Spanish	English
Train	Sentences	1.2M		1.7M	
	Running Words	35M	33M	33M	31M
	Vocabulary	134K	94K	124K	80K

Table 13: Chinese–English: Effect of Truecasing.

evaluation	truecasing	NIST	loss [%]	BLEU[%]	loss [%]
case-insensitive	–	6.09	–	16.9	–
case-sensitive	Oct04	5.74	5.7	15.3	9.5
	Mar05	5.85	3.9	15.8	6.5

(Verbatim).

4.2.3 Results

In this section, we present some translation results of our system.

In Table 13, we show the effect of the new true case mapping. The March 2005 truecaser, based on the disambig tool from the SRI toolkit, results in an improvement of 0.5% of the BLEU score compared to the maximum entropy based truecaser from October 2004.

In Table 14, a comparison of the baseline system from October 2004 and the March 2005 evaluation system is shown. The results are presented for the TC-Star development set and for the official March 2005 evaluation test set. On the development set as well as on the evaluation test set, we observe an improvement for all tasks and for all four evaluation criteria. For the verbatim and the ASR task, we observe a large improvement of more than 2% absolute for the BLEU score on the evaluation test set.

In Table 15, the translation results for the TC-Star EPPS development set are presented. The word error rates of the speech recognizer in the ASR conditions are 12% for Spanish and 15% for English.

4.2.4 Summary

We use a phrase-based translation model that memorizes all phrasal translations that have been observed in the training corpus. We use a weighted log-linear combination of various models. This allows the optimization of the model scaling factors with respect to the final evaluation criterion. Additional models, for example IBM model 1, can be easily integrated via rescoring of N-best lists.

4.3 The UPC System

4.3.1 Corpus Preprocessing

The EPPS corpus [EPPS 05] has been preprocessed by using standard tools for tokenizing and filtering. In the filtering stage, sentence pairs with a word ratio larger than 2.4 have been removed, as well as sentence pairs with at least one sentence of more than 100 words in length.

In addition to this basic preprocessing, a verbatim-oriented preprocessing, was performed for the verbatim system. In this specific case, punctuation marks have been removed from the training data sets. However, this was not exactly a preprocessing procedure since it was actually performed after aligning the corpus. It was done in such a way in order to take advantage of the better alignment quality obtained when including punctuation marks.

Table 14: Chinese–English: Translation Results for the TC-Star Development and the March 2005 Evaluation Test Set (case-sensitive evaluation).

	Task	System	WER[%]	PER[%]	NIST	BLEU[%]
Dev	Text	Oct04	77.1	57.5	5.74	15.3
		Mar05	74.2	54.7	5.99	16.2
	Verbatim	Oct04	81.6	61.2	5.72	14.5
		Mar05	76.6	55.8	6.27	16.4
Eval	Text	Oct04	78.3	58.1	5.66	14.7
		Mar05	75.8	55.4	5.95	16.5
	Verbatim	Oct04	82.3	61.2	5.57	14.2
		Mar05	79.5	57.7	6.15	16.7
	ASR	Oct04	82.6	61.9	5.43	13.6
		Mar05	78.1	57.8	5.87	16.2

Table 15: Spanish–English (EPPS): Results for the TC-Star Development Set.

Direction	Condition	WER[%]	PER[%]	NIST	BLEU[%]
Spanish→English	Text	31.5	22.8	11.3	60.5
	Verbatim	36.3	29.3	10.5	53.5
	ASR	42.1	35.2	9.6	47.1
English→Spanish	Text	38.2	29.5	10.0	52.5
	Verbatim	42.6	34.3	9.7	47.3
	ASR	49.8	41.2	8.5	39.4

4.3.2 System Setup for the Evaluation

According to the maximum entropy framework [Berger & Della Pietra$^+$ 96], the corresponding translation hypothesis T, for a given source sentence S, is defined by the target sentence that maximizes a log-linear combination of feature functions $h_i(S,T)'s$.

$$\underset{T}{\operatorname{argmax}} \left\{ \exp\left(\sum_i \lambda_i h_i(S,T) \right) \right\} \tag{1}$$

where the λ_i's constitutes the weighting coefficients of the log-linear combination. These λ_i's are computed via an optimization procedure which maximizes the translation BLEU score [Papineni & Roukos$^+$ 02] over a given development set. This optimization is performed by using an inhouse developed optimization algorithm, which is based on a simplex method [Press & Teukolsky$^+$ 02]. The UPC's translation system implements a total of five feature functions. The first of these feature functions is a tuple 3-gram model. This model, which actually constitutes the translation model of UPC's system, approaches the joint probability between source and target languages by using 3-grams [Crego & Mariño$^+$ 04].

$$p(T,S) = \prod_{n=1}^{N} p((t,s)_n | (t,s)_{n-2}, (t,s)_{n-1}) \tag{2}$$

where $(t,s)_n$ refers to the n-th tuple of a given bilingual sentence pair. It is important to notice that, since both languages are linked in tuples, the context information provided by this model is bilingual.

The tuple 3-gram model is indeed a language model of bilingual units called tuples [Gispert & Mariño 02]. Tuples are extracted from a word-to-word aligned corpus. The word-to-word alignment is performed in both directions, source to target and target to source, by using GIZA++ [Och & Ney 00]. Then, tuples are extracted from the union of both alignments according to the following constraints: i.- tuple extraction should produce a monotonic segmentation of bilingual sentence pairs, and ii.- the produced segmentation is maximal in the sense that no smaller tuples can be extracted without violating the previous constraint [Crego & Mariño$^+$ 04].

Once tuples have been extracted, the tuple vocabulary is pruned by using histogram pruning. More specifically, this pruning is performed by keeping the N-most frequent tuples with same source sides. In the case of the EPPS data [EPPS 05], a value of N=20 happened to provide a good trade off between translation quality and computational expenses for Spanish to English translations. In the case of English to Spanish, a value of N=30 happened to provide the best trade off. Finally, the tuple 3-gram model is trained by using the SRI Language Modeling toolkit [Stolcke 02], and the improved back-off smoothing method proposed by [Kneser & Ney 95] is used.

Two important issues regarding this translation model must be considered: i.- a relative important amount of single-word translation probabilities are left out of the model; and, ii.- some words linked to NULL end up producing tuples with NULL source sides. On the one hand, for all those words that always appear embedded into tuples containing two or more words, no translation probability for an independent occurrence of such words exist. To overcome this problem, the tuple 3-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words detected during the tuple extraction step [Gispert & Mariño$^+$ 04]. These 1-gram translation probabilities are computed from the intersection of both, source to target and target to source, alignments.

On the other hand, tuples with NULL source sides cannot be allowed. This problem is solved by preprocessing the union of alignments before the tuple extraction step. During this preprocessing any target word that is linked to NULL is attached to either its precedent word or its following word. In this way, no target word remains linked to NULL, so tuples with NULL source sides will not occur during tuple extraction.

The second feature function implemented in the UPC's translation system is a target language model. This feature actually consists of a word 3-gram model.

$$p(w_n|w_1, w_2, ..., w_{n-1}) = p(w_n|w_{n-2}, w_{n-1}) \tag{3}$$

This model is trained from the target side of the bilingual corpus by using the SRI Language Modeling toolkit [Stolcke 02], and again, the improved back-off smoothing method proposed by [Kneser & Ney 95] is used.

Extended target language models might also be obtained by adding additional information from other available monolingual sources. These extended target language models are computed by performing a log-linear combination of independently computed target language models. The weights of the log-linear combination are adjusted so perplexity, with respect to a given development set, is minimized. In the case of this first evaluation campaign no extended target language models were used.

The three remaining feature functions implemented in the UPC's translation system are the following:

- A word penalty model: which introduces a sentence length penalization in order to compensate the system preference for short output sentences.
- Source to target lexicon model: which introduces a lexicon translation probability based on word to word IBM model 1 probabilities [Och & Gildea$^+$ 04b]. IBM model 1 is trained from the bilingual training data by using GIZA++ source to target alignments.
- Target to source lexicon model: this feature function is very similar to the previous one, but in this case, GIZA++ target to source alignments are used instead.

Finally, the search engine of UPC's translation system consists of an in-house developed decoder. This decoder implements a beam-search strategy based on dynamic programming and allows for three different pruning methods: i.- threshold pruning, ii.- histogram pruning, and iii.- hypothesis recombination.

In the first case, hypotheses for which scores are below a predetermined threshold value are eliminated. This type of pruning, although available in the decoder, was not actually used in experiments performed for this first evaluation campaign.

In the second kind of pruning, the maximum number of hypotheses to be considered is limited the K-best ranked ones. During decoding, competing hypotheses are arranged in different lists according to the amount of source words covered by each of the hypotheses. Histogram pruning is then used to restrict the size of such lists. In the case of the EPPS data [EPPS 05], a value of K=50 happened to provide a good trade off between translation quality and computational expenses for both directions, English to Spanish and Spanish to English.

In the third case, a method for recombining hypotheses, which constitutes a risk free pruning method, is implemented [Koehn 04]. At any step of the search, two or more hypotheses are recombined if they agree in both the present tuple and the tuple 3-gram history.

The decoding algorithm takes into account all the five features described above simultaneously. All the results generated for this first evaluation campaign were performed by using monotonic search (i.e. without including reordering capabilities) and the decoding was always guided by the source. Although an effort was done to take advantage of the word reordering capability of the decoder, no improvement has been achieved yet.

4.3.3 Evaluation and Conclusions

The following tables present the official results (provided by ELDA) corresponding to the first evaluation campaign, as well as some internal results obtained for both UPC's original baseline system and UPC's final system. In the original baseline system only one feature function, the tuple 3-gram model, was used; while in the final system all five feature functions previously described were taken into account.

Table 16: Internal and official results for Spanish-to-English translations

INPUT	Condition	Evaluation	System	WER	BLEU	NIST	CER
final text edition	primary	internal	baseline	36.07	0.57	11.01	28.90
final text edition	primary	internal	final	30.62	0.65	11.99	24.97
final text edition	primary	official	final	35.12	0.53	10.55	27.80
verbatim	primary	internal	baseline	43.97	0.44	9.32	33.90
verbatim	primary	internal	final	39.85	0.50	10.00	30.91
verbatim	primary	official	final	44.06	0.42	9.26	32.60
ASR (LIMSI)	primary	official	final	48.75	0.38	8.56	35.52

Table 17: Internal and official results for English-to-Spanish translations

INPUT	Condition	Evaluation	System	WER	BLEU	NIST	CER
final text edition	primary	internal	baseline	41.17	0.51	10.05	31.43
final text edition	primary	internal	final	37.16	0.57	10.67	28.82
final text edition	primary	official	final	41.20	0.46	9.66	31.73
verbatim	primary	internal	baseline	50.98	0.36	8.32	36.84
verbatim	primary	internal	final	47.10	0.40	8.82	34.35
verbatim	primary	official	final	49.46	0.38	8.73	35.75
ASR (LIMSI)	primary	official	final	54.03	0.34	7.99	38.44

It is important to mention that in the case of internal results presented in the tables, the development data provided by ELDA was used as test data for computing the evaluation measurements, while only the first half of each development set were used as actual development data. In other words, in the case of internal results, the development data used for optimizing and the test data used for computing the measurements overlap. This explains the fact, which becomes evident from the tables, that for all comparable cases, internal results are consistently and slightly better than the official evaluation results.

Notice that PER values have been omitted. This was because, by the time this report was required, some inconsistencies were still present in the available script for computing the PER. Also, for all entries in both tables, only case sensitive evaluations are considered.

From comparing internal results, it can be seen that UPC's final system presents significant improvements in translation quality with respect to the original baseline system. These results clearly indicate that the four additional feature functions that were implemented (target language model, word penalty model, source to target lexicon model and target to source lexicon model) do indeed provide valuable information during the decoding process and improve translation quality.

A more detailed evaluation about the relative impact of each of these four feature functions revealed that the model with the most impact on translation quality was the source to target lexicon model. The target language model, the target to source lexicon, and the word penalty also contributed to improve translation quality, but in less degree, being the latter the one with the lowest impact. Another important observed fact was that the target language model becomes more relevant when lexicon models are used. Indeed, when the lexicon models are not present, the weight of the target language model becomes neglectable during the optimization process. Although reasons of this behavior are not clear at all yet, it seems that including lexicon models tends to favor short tuples over long ones, so the target language model becomes more important for providing target context information. However more evaluation and research is required for fully understanding this interesting result.

From comparing official results, it can be observed that best results were achieved for the final text edition set, followed by verbatim transcripts and, finally, the ASR output. It is interesting to notice that our text system outperformed our verbatim system. While the former was used for translating the final test edition set, the latter was used for for translating both the verbatim and the ASR sets, for which measurements degraded significantly. This can be explained by the existence of tokens such as ”uhm”, ”ah” (among others), in the verbatim references, which were not taken into account by our verbatim system during both training and translation.

Also from the tables, it can be seen that in all the cases Spanish to English translation results are consistently and significantly better than English to Spanish translation results. This is clearly due to the more inflected nature of Spanish vocabulary. For example the single English word ”the” can generate any of the four Spanish words ”el”, ”la”, ”los” and ”las”. Similar situations occur with nouns, adjectives and verbs which may have many different forms in Spanish depending on gender, number, tense and mode. According to this, significant efforts should be dedicated for properly exploiting morphological analysis and synthesis methods for improving English to Spanish translation quality.

Additionally, a detailed observation of translated sentence pairs was performed. This exercise resulted to be very useful since it allowed to identify the most common errors and problems related to the translation systems that were evaluated. A brief analysis of this observation revealed that most of translation problems encountered are mainly related to four basic issues: verbal forms, translations resulting into NULL, reordering and concordance. Regarding reordering, the two specific cases that most commonly occurred were problems related to adjective-noun and subject-verb structures. In the case of concordance, inconsistencies related to gender and number were the most commonly found. More evaluation and discussion is required in this area for fully understanding the most common translation failures and, then, implementing appropriate solutions.

Finally, regarding our system efficiency, we can mention that in both translation directions the memory required during decoding was around 1.2 Giga-bytes. Translation times, on the other hand, vary according to the number of sentences and translation direction. However, as illustrative examples we can give the following measurements for the final text edition set: i.- Spanish to English, 840 sentences, total translation time = 1300 seconds; and 2.- English to Spanish, 1094 sentences, total translation time = 2300 seconds.

4.3.4 Ongoing and Further Research Work

UPC's ongoing research work is moving into three principal directions: i.- incorporating human-supplied knowledge into SMT, ii.- developing a flexible and efficient decoding tool, and iii.- improving translation unit (tuple) determination.
Regarding the incorporation of human-supplied knowledge into SMT, the following four activities can be mentioned:

- Bilingual dictionary: the impact on translation accuracy of using a bilingual dictionary during training, decoding and/or post-processing is under study.
- Multi-word expressions: the impact of identifying and using multi-word expressions (extracted from WordNet and other bilingual sources) as single tokens is being tested on both alignment quality and translation accuracy.
- Verbal form classes: the impact of identifying and replacing verbal forms with corresponding classes is being tested on both alignment quality and translation accuracy.
- POS-tag models: the impact of incorporating new feature functions based on POS-tag translation probabilities and/or target POS-tag n-gram probabilities will be studied and evaluated.

Regarding the development of a flexible and efficient decoding tool, the following five activities can be mentioned:

- Beam-search decoder: a decoder has been implemented by using a beam-search strategy. This decoder allows for combining multiple models or features and selecting from different driving conditions to guide the translation. This decoding tool is already fully operational.
- Optimization tool: an optimization tool, which is based on a simplex algorithm, has been implemented. This tool allows for calculating optimal weights (in either a minimum WER or maximum BLEU sense) for a given log-linear combination of models and feature functions.
- N-best list generation: a decoding tool for generating n-best lists of translation hypothesis will be implemented.
- Word graph input: in order to improve translation accuracy from spoken language input the SMT system should be able to handle word graph inputs. Since word graph error rates are lower that single best error rates in ASR systems, the use of word graphs as SMT inputs will help SMT systems to better dealing with ASR errors.
- Reordering capability: as already mentioned, an effort was done to take advantage of the word reordering capability of the decoder. However, no significant improvement over monotonic decoding has been achieved yet. A reordering strategy which makes use of the decoder's reordering capability only when the tuple translation model falls to the 1-gram will be implemented and evaluated.

Finally, regarding translation unit determination, most of the effort is being focused into two activities:

- Tuple-border definition: present criteria for tuple extraction will be reviewed in order to improve tuple n-gram model performance. Some specific problems related to the present tuple extraction strategy have been identified. This is the case of, for example, extremely long tuples resulting from long alignment links, and tuples containing NULL at either source or target side.
- Mini-chunk information: in order to reduce data sparseness a preprocessing strategy for replacing mini-chunks by lemma forms will be designed and evaluated.

4.4 The UKA System

4.4.1 Introduction

The statistical machine translation system developed in the Interactive Systems Laboratories (ISL) uses phrase-to-phrase translations as the primary building blocks to capture local context information, leading to better lexical choice and more reliable local reordering. A new approach to extract phrase translation pairs from bilingual data has been developed, which is not using the Viterbi alignment, but is based on optimizing a constrained word-to-word alignment for the entire sentence pair [Vogel & Hewavitharana+ 04].

4.4.2 Training Data and Preprocessing

For the EPPS task, we used the provided parallel Spanish-English corpus of about 35 million running words. For the Chinese-English BN task, the training data was selected from the specified LDC corpora. As a preprocessing step, the Chinese side was segmented using the LDC segmenter, and a rule-base translation of number and date expressions was performed. The English side is tokenized, separating punctuation marks from proper words, and converted to lowercase. A similar preprocessing was applied to the EPPS corpus. The Chinese-English corpus was then pruned to include sentences of up to 58 words, resulting in about 200 million running words per side. Prior to lexicon training and phrase alignment, part of the training sentence pairs were split at selected splitpoints, based on incrementally calculated lexical IBM1 probabilities.

4.4.3 Models

Phrase Alignment The ISL translation system uses word-to-word and phrase-to-phrase translations, extracted from the bilingual corpus. Different phrase alignment methods have been explored in the past, like extracting phrase translation pairs from the Viterbi path of a word alignment, or simultaneously splitting source and target sentence into phrases and aligning them in an integrated way [Zhang & Vogel+ 03].

Phrase Alignment via Constrained Sentence Alignment Assume we are searching for a good translation for one source phrase $\tilde{f} = f_1...f_k$, and that we find a sentence in the bilingual corpus, which contains this phrase. We are now interested in finding a sequence of words $\tilde{e} = e_1...e_l$ in the target sentence, which is an optimal translation of the source phrase. Any sequence of words in the target sentence is a translation candidate, but most of them will not be considered translations of the source phrase at all, whereas some can be considered as partially correct translations, and a small number of candidates will be considered acceptable or good translations. We want to find these good candidates.
The IBM1 word alignment model aligns each source word to all target words with varying probabilities. Typically, only one or two words will have a high alignment probability, which for the IBM1 model is just the lexicon probability. We now modify the IBM1 alignment model by not summing the lexicon probabilities of all target words, but by restricting this summation in the following way:

- for words inside the source phrase we sum only over the probabilities for words inside the target phrase candidate, and for words outside of the source phrase we sum only over the probabilities for the words outside the target phrase candidates;
- the position alignment probability, which for the standard IBM1 alignment is $1/I$, where I is the number of words in the target sentence, is modified to $1/(l)$ inside the source phrase and to $1/(I - l)$ outside the source phrase.

More formally, we calculate the constrained alignment probability:

$$p_{i_1,i_2}(f|e) = \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} p(f_j|e_i) \times \\ \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} p(f_j|e_i) \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} p(f_j|e_i)$$

and optimize over the target side boundaries i_1 and i_2.

$$(i_1, i_2) = \operatorname*{argmax}_{i1,i2} \{p_{i_1,i_2}(f|e)\}$$

It is well know that 'looking from both sides' is better than calculating the alignment only in one direction, as the word alignment models are asymmetric with respect to aligning one to many words.
Similar to $p_{i_1,i_2}(f|e)$ we can calculate $p_{i_1,i_2}(e|f)$, now summing over the source words and multiplying along the target words:

$$p_{i_1,i_2}(e|f) = \prod_{i=1}^{i_1-1} \sum_{j \notin (j_1...j_2)} p(e_i|f_j) \times \\ \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} p(e_i|f_j) \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1...j_2)} p(e_i|f_j)$$

To find the optimal target phrase we interpolate both alignment probabilities and take the pair (i_1, i_2) which gives the highest probability.

$$(i_1, i_2) = \operatorname*{argmax}_{i_1,i_2} \{(1-c)p_{(i_1,i_2)}(f|e) + cp_{(i_1,i_2)}(f|e)\}$$

In addition, we take not only the best translation candidate, but all candidates which are within a given margin to the best one. All candidates are then used in the decoder, when also the language model is available to score the translations. The phrase pairs can be either extracted from the bilingual corpus at decoding time or stored and reused during system tuning. Single source words are treated in the same way, i.e. just as phrases of length 1. The target translation can then be one or several words.

Phrase Translation Probabilities Most phrase pairs $(\tilde{f}, \tilde{e}) = (f_{j_1}...f_{j_2}, e_{i_1}...e_{i_2})$ are seen only a few times, even in very large corpora. Therefore, probabilities based on occurrence counts have little discriminative power. In our system we calculate phrase translation probabilities based on a statistical lexicon, i.e. on the word translation probabilities $p(f, e)$:

$$p(\tilde{f}|\tilde{e}) = \prod_j \sum_i p(f_j|e_i).$$

Language Model The language model used in the decoder is a standard 3-gram language model. We use the SRI language model toolkit [Stolcke 02] to build language models of different sizes, using the target side of the bilingual data only or using additional monolingual data.

4.4.4 Decoding

The decoding process works in two stages: First, the word-to-word and phrase-to-phrase translations are used to generate a translation lattice. Second, a first-best or n-best search is performed on this lattice, using the language model probabilities in addition to the translation model probabilities to find the overall best translation.
Once the complete translation lattice has been built, a first-best search through this lattice is performed. In addition to the translation probabilities, or rather translation costs, as we use the negative logarithms of the probabilities for numerical stability, the language model costs are added and the path which minimizes the combined cost is returned.
Starting with a special begin-of-sentence hypothesis attached to the first node in the translation lattice, hypotheses are expanded over all outgoing edges from the current node. To realize word reordering, the search algorithm allows to leave a gap and jump to a distant node in the translation lattice, filling the gap at a later time. This requires to keep track of positions already covered in the source sentence.
The search space, especially when allowing for reordering, is very large. Pruning is applied to keep decoding times within reasonable bounds. Our decoder realizes a standard beam search, where all hypotheses which are worse than the best hypothesis by some factor are deleted [Vogel 03].

4.4.5 Postprocessing

For the purposes of this evaluation, post-processing consisted of removing untranslated words, removing all punctuation marks for the ”plain transcription” conditions, and adding case information for all conditions. Case information was obtained by treating casing as a translation problem itself, training translation models on lower case/mixed case bi-text, and ”translating” using our decoder with word reordering disabled.

4.5 The IBM System

We describe the baseline phrase translation system, key performance enhancing techniques and experimental results for Spanish-to-English and English-to-Spanish translations.

4.5.1 Base Phrase Translation and Language Models

Phrase translation models are obtained via word alignment and phrase selection. We obtain word alignment between the source and the target language sentences by successive application of IBM Model 1 viterbi alignment for initialization and iterative HMM-based alignment, [Vogel & Ney$^+$ 1996], for refinement. We align a parallel corpus bi-directionally: one from the source language to the target language $(A1 : f \rightarrow e)$ and the other from the target language to the source language $(A2 : e \rightarrow f)$, where f denotes a source word position and e a target word position. We define precision (A_P) and recall (A_R) oriented alignments as follows: A_P is the intersection of A1 and A2, a high precision alignment. A_R is the union of A1 and A2, a high recall alignment. Starting from a high precision alignment A_P, we obtain phrases according to the projection and block extension algorithm detailed in [Tillmann 03].
Phrase translation probabilities are obtained differently for Spanish-to-English and English-to-Spanish translations. For Spanish-to-English translation, the translation probability of a target phrase $(\overline{e})$ given a source phrase $(\overline{f})$ is obtained directly, as shown by the formula (4):

$$p(\overline{e} \mid \overline{f}) = \text{count}(\overline{e}, \overline{f}) / \sum_{e'} \text{count}(\overline{e}, \overline{f}) \tag{4}$$

For English-to-Spanish translation, the unigram translation probability of a target-source phrase pair $(\overline{e}, \overline{f})$, called a block $(b = \overline{e}, \overline{f})$, is obtained, as shown by the formula (5):

$$p(b) = \text{count}(b) / \sum_{b'} \text{count}(b') \tag{5}$$

Trigram language model probabilities are obtained for each word included in a target phrase ($\bar{e}$), with the option of assigning different weights for words at phrase boundaries and words inside a phrase, as shown by the formula (6).

$$p(e_i | e_{i-1}, e_{i-2}) * \alpha \tag{6}$$

α is the weight assigned to the word trigram language model probability, which may be set differently for e_i inside a phrase and e_i at phrase boundaries.
For decoding, we use a DP-based beam search procedure. We start with an initial empty hypothesis. We maximize over all block segmentations $b_{1,n}$, where *n* is the number of blocks covering the input sentence. The source phrases yield a segmentation of the input sentence, generating the target sentence simultaneously. The decoder processes the input sentence 'cardinality synchronously', i.e. all partial hypotheses active at a given point cover the same number of input words. We prune out weaker hypotheses based on the cost (for the phrase translation model probability and the language model probability) they incurred so far. The final hypothesis with the minimum cost, i.e. the hypothesis with the highest probability, with no un-translated source words constitutes the translation output.

4.5.2 Performance Enhancing Techniques

We deploy various techniques to achieve performance improvement over the baseline system:

1. Application of automatically acquired reordering rules to source sentences
2. IBM Model 1 cost
3. Word-based distortion models
4. Word/block count penalty

IBM Model 1 cost, word-based distortion models and word/block count penalty are incorporated into the decoder cost function. Reordering rules are applied as pre-processing to translation model training and decoding.

IBM Model 1 Cost

We augment the phrase translation model cost with IBM Model 1 cost. IBM Model 1 translation cost associated with each phrase is computed, as shown by the formula (7):

$$\sum_{j=0}^{m} -\log_{10}\left(\sum_{i=0}^{n} p(f_j|e_i) / \text{linkCount}\right) \tag{7}$$

j is the source word position index, *m* is the number of source words. *i* is the target word position, *n* is the number of target words. *linkCount* is the number of target words whose translation probability into the source word f_j is greater than 0.0. In cases where linkCount is zero, we assign a fixed cost β. We set β to 5.0 for Spanish-to-English translation and 5.5 for English-to-Spanish translation.

Distortion Models

Distortion model probabilities, introduced in [Al-Onaizan 2004], are computed at word level, consisting of lexical pair and out bound models, as in (8, 9)

$$\text{Lexical pair model: } p_{pair}(j-i|f_j, f_i) \tag{8}$$

$$\text{Out bound model: } p_{out}(j-i|f_i) \tag{9}$$

f_i is the source word in the i^{th} position, f_j is the source word in the j^{th} position. i is the position index of the source word f_i. j is the position index of the source word f_j whose translation word immediately follows the translation word of f_i. Distortion model cost between two sets of phrases is the minimum cost of the cross product of each word position in the two phrases.

Reordering Rules

Reordering rules are acquired from viterbi-aligned parallel corpus where the source language corpus is part-of-speech tagged and the target language corpus is not part-of-speech tagged. Schematic procedure for automatic reordering rule acquisition is given below, [Lee 04]:

- Step 1: IBM Model 1 viterbi alignment between part-of-speech tagged source language corpus and un-tagged target language corpus
- Step 2: Identification of source part-of-speech tag sequences whose corresponding target word sequences are monotone decreasing
- Step 3: Computation of reordering probabilities of each source part-of-speech tag sequence $\overline{tag_k}$ according to the formula (10)

$$p(reorder_i|\overline{tag_k}) = \text{count}(reorder_i, \overline{tag_k}) / \sum_{reorder'} \text{count}(reorder', \overline{tag_k}) \tag{10}$$

Examples of Spanish word/part-of-speech tag sequences whose corresponding English word sequences are monotone decreasing and their reordering probabilities are given in Table 18.

The reordering rule of a given source tag sequence is the reordering pattern with the highest probability, satisfying the condition in (11):

$$p(reorder_i|\overline{tag_k}) > p(reorder_j|\overline{tag_k}) + \alpha, 0 < \alpha < 0.5 \tag{11}$$

$reorder_i$ is the reordering pattern with the highest probability, and $reorder_j$ is the reordering pattern with the second highest probability.

Table 18: Examples of Spanish word/part-of-speech tag sequences whose corresponding English word sequences are monotone decreasing and their reordering probabilities.

case A: neuve/DTS propuestas/NNS legislativas/JJS presentadas/JJS
→ nine proposals legislative presented
case B: de/IN transportes/NNS especialmente/RB peligrosos/JJS
→ of transport extremely dangerous

A: $\mathrm{IN}_{\mathrm{S1}}$ $\mathrm{NNS}_{\mathrm{S2}}$ $\mathrm{RB}_{\mathrm{S3}}$ $\mathrm{JJS}_{\mathrm{S4}}$		**B:** $\mathrm{DTS}_{\mathrm{S1}}$ $\mathrm{NNS}_{\mathrm{S2}}$ $\mathrm{JJS}_{\mathrm{S3}}$ $\mathrm{JJS}_{\mathrm{S4}}$	
$reorder'$	$p(reorder'\|\overline{tag_k})$	$reorder$	$p(reorder'\|\overline{tag_k})$
S1 S2 S3 S4	0.176	S1 S2 S3 S4	0.107
S1 S2 S4 S3	0.049	S1 S2 S4 S3	0.059
S1 S3 S2 S4	0.091	S1 S3 S2 S4	0.239
S1 S3 S4 S2	0.566	S1 S3 S4 S2	0.109
S1 S4 S2 S3	0.071	S1 S4 S2 S3	0.111
S1 S4 S3 S2	0.047	**S1 S4 S3 S2**	0.375

Table 19: Development test set statistics.

	Spanish-to-English		English-to-Spanish	
input conditions	FTE	VHT	FTE	VHT
no. of segements	1,008	2,619	1,008	1,746

4.5.3 Experimental Results

Performances of Spanish-to-English and English-to-Spanish translation systems are evaluated on the TC-STAR development test set, using case sensitive BLEU [Papineni & Roukos$^+$ 02] and 2 reference translations. Development test set statistics are shown in Table 19 (FTE stands for final text edition, VHT for verbatim human transcription, hereafter).

System performances of the primary and the secondary systems (after decoder parameter tuning) are shown in the Table 20.

Table 20: System performances of the primary and the secondary systems.

	Spanish-to-English		English-to-Spanish	
input conditions	FTE	VHT	FTE	VHT
BLEU (Primary)	0.5576	0.4677	0.4621	0.3782
BLEU (Secondary)	0.5539	0.4687	0.4690	0.3805

Primary evaluation training data (for both TM and LM) consists of RWTH-distributed EPPS corpus containing 1,227,172 sentence pairs. For secondary evaluation, we add UN parallel corpus (LDC94T4A) containing 99,669 sentence pairs for translation model training. For language model traning, we add English Gigaword corpus (LDC2003T05) for Spanish-to-English translation, and Spanish News Texts (LDC95T9, LDC99T41) and Spanish Parliament Texts (UPC) for English-to-Spanish translation.

Impact of various techniques on Spanish-to-English and English-to-Spanish translations is shown in Table 21, where CURRENT stands for the primary system trained for the input condition FTE.

Table 21: Comparison of techniques for Spanish-to-English and English-to-Spanish translations.

Systems	Spanish-to-English	English-to-Spanish
CURRENT	0.5576	0.4621
IBM Model 1	0.5074	0.4295
Distortion Model	0.5315	0.4529
Reordering Rules	0.5439	0.4608
Word/block count penalty	0.5359	0.4443

Impact of various techniques is measured by setting the decoder parameter weight of the given technique to 0, retaining all other parameter weights the same as the decoder parameter weights of the CURRENT system. IBM Model 1 cost is the most effective technique. Reordering rules and distortion models together improve the system performance statistically significantly. Word and block count penalty, cf. [Zens & Ney 04], together also improve the system performance statistically significantly.

5 Results of Evaluation Campaign

In this section, we summarize the results of the evaluation campaign for the SLT systems. For each of the partners, we will consider only the best system version. Each system will be evaluated using the following four measures that are widely used and accepted:

- **WER** := word error rate.
  It is defined as in speech recognition. The disadvantage is that it does not allow any difference in word order.
- **PER** := position independent word error rate.
  It ignores the word positions and counts, independent of word positions, how many of the words are correct or wrong. The disadvantage is that it allows *any* word order.
- **BLEU** := bilingual evaluation understudy (accuracy measure).
  It measures the degree of n-gram overlap between test sentence and reference sentence. BLEU is basically a precision measure (as known from precision and recall in information retrieval) and uses some geometric averaging. There is no simple interpretation of the measure as for WER and PER.
- **NIST** := National Institute of Standardization and Technology (accuracy measure). This is a variant of the BLEU measure, in which the geometric average is replaced by an arithmetic average. As for BLEU, there is no simple interpretation.

Note that BLEU and NIST are *accuracy* measures ('high values are better') as opposed to the *error* measures WER and PER. Each of these four measures can handle *multiple* reference translations and shows a (more or less) good correlation with human judgement. In the evaluation, ELDA used two reference translations for each source sentence.

As pointed out in Section 3.1, three types of inputs to the SLT systems were used:

- **ASR:** the output of automatic speech recognizers.
  The ASR input allows us to measure the performance of *recognition* and *translation*.

- **verbatim:** the verbatim (i.e. correct) transcription of the spoken sentences.
  The verbatim input removes the effects of speech recognition errors, but preserves the effects of *spoken* language in comparison with *written* language.
- **text:** the final text editions, i.e. the official EU documents.
  The text input allows us to study the difference between *spoken* and *written* language.

Since the speech recognizers were not (yet) able to produce punctuation marks (or hypotheses thereof), punctuation marks were included in the reference translation only for the text input. For ASR input, the ROVER combination of all speech recognizer results was used – if possible. For Chinese speech recognition, there was only a single recognizer.

5.1 Results: Chinese to English

The evaluation results for the Chinese-English BN task are summarized in Table 22. In addition to the TC-Star sites (IBM, ITC-irst, RWTH, UKA), there was an external site, namely JHU (join team of Johns-Hopkins University and Cambridge University). The ASR input was provided by the the joint LIMSI/UKA speech recognizer (for details see deliverable D6), which had a *character error rate* (CER) of approxmatley 9.5%. For the IBM system with ASR input, a bug had been reported.

Table 22: Evaluation Results for Chinese-English BN.

Input	Site	BLEU [%]	NIST	PER [%]	WER [%]
ASR	RWTH	16.2	5.87	57.8	78.1
(CER $\approx$ 9.5 %)	UKA	13.5	5.46	61.8	81.2
	JHU	13.2	5.43	63.8	85.3
	ITC-irst	11.5	5.20	63.6	83.7
	IBM	5.2	2.61	90.0	104.5
Verbatim	RWTH	16.8	5.99	58.0	78.6
	IBM	13.7	5.70	62.4	86.6
	UKA	13.6	5.64	60.8	80.8
	JHU	13.4	5.58	63.1	84.9
	ITC-irst	12.0	5.37	62.8	83.6
Text	RWTH	16.5	5.95	55.4	75.8
	JHU	14.6	5.75	58.9	80.8
	UKA	14.2	5.67	58.2	78.2
	IBM	13.9	5.67	60.9	84.8
	ITC-irst	12.6	5.36	61.4	82.3

In Table 22, the systems are ordered according to the BLEU measure. However, apart from a few exceptions, each of the other three mesures produces the same ranking of the systems. This will also be true for the two EPPS tasks (see later).
Looking at the results shown in Table 22, we make the following observations:

- There is a strong correlation between the four evaluation measures.
- There is only a comparatively small degradation when we switch from text to verbatim input and from verbatim to ASR input.

- The absolute level of performance is not too good. Even for the best systems, we get only 40-45% (= 1-PER) of the words correct, independent of the word positions. (For comparison: this will be much better for the EPPS tasks, namely around 70%.)
- Looking at WER, i.e. including the word positions, the error rates are even higher, namely by about 20%. In other words, the word order seems to pose some serious problems.

We speculate that a major problem for this task is the mismatch between training and test data. The test data are broadcast news data, and this type of data does not occur in the training data.

5.2 Results: English to Spanish

The evaluation results for the English-Spanish EPPS task are summarized in Table 23. In addition to the TC-Star sites (IBM, RWTH, UKA, UPC), there was an external site, namely UPV (Universidad Politecnica de Valencia). The ASR input was the ROVER combination of all English speech recognizers, which had a (recognition) word error rate (WER) of 9.9% (for details see deliverable D6). There was an exception for the UPC system which used the case-sensitive output of the LIMSI speech recognizer (UPC* in Table 23). For the RWTH system with text input, a bug had been reported.

Table 23: Evaluation Results for English-Spanish EPPS.

Input	Site	BLEU [%]	NIST	PER [%]	WER [%]
ASR	RWTH	38.7	8.73	38.6	49.8
(WER= 9.9%)	IBM	34.3	8.13	42.0	54.5
	UPC*	33.8	8.00	43.1	54.0
	UKA	33.0	7.94	43.4	55.9
	UPV	19.1	5.46	53.3	62.5
Verbatim	RWTH	42.5	9.32	35.4	46.1
	UPC	38.1	8.72	39.2	49.5
	IBM	36.8	8.55	39.6	51.8
	UKA	33.4	8.29	41.8	53.2
Text	UPC	46.2	9.65	32.7	41.2
	IBM	45.2	9.44	32.8	43.2
	RWTH	38.9	8.72	36.1	48.4
	UKA	37.6	8.46	38.3	49.6
	UPV	34.1	7.51	40.8	48.7

Looking at the results shown in Table 23, we make the following observations:

- There is a strong correlation between the four evaluation measures (as for Chinese-English).
- There is only a comparatively small degradation when we switch from text to verbatim input and from verbatim to ASR input (as for Chinese-English).
- The absolute level performance is rather good (but still needs improvements). For the best systems, we get 60-70% (= 1-PER) of the words correct, independent of the word positions. (This is much better than for the Chinese-English task.)
- Looking at WER, i.e. word errors including the word positions, the error rates are only slightly increased over PER, namely by about 10% absolute.

5.3 Results: Spanish to English

The evaluation results for the Spanish-English EPPS task are summarized in Table 24. In addition to the TC-Star sites (IBM, ITC-irst, RWTH, UKA, UPC), there was an external site, namely UPV (Universidad Politecnica de Valencia). The ASR input was the ROVER combination of all Spanish speech recognizers, which had a (recognition) word error rate (WER) of 10.1% (for details see deliverable D6). Again, there was an exception for the UPC system which used the case-sensitive output of the LIMSI speech recognizer (UPC* in Table 24). For the RWTH system with text input, a bug had been reported.

Table 24: Evaluation Results for Spanish-English EPPS.

Input	Site	BLEU [%]	NIST	PER [%]	WER [%]
ASR	RWTH	41.5	9.12	35.4	46.6
(WER= 10.1%)	IBM	39.7	8.81	37.7	48.6
	UPC*	37.7	8.56	39.2	48.7
	ITC-irst	34.7	7.97	42.8	53.8
	UKA	32.3	7.85	43.1	55.0
	UPV	16.0	4.35	57.1	63.6
Verbatim	RWTH	45.9	9.75	31.7	42.5
	IBM	44.1	9.47	33.4	43.9
	UPC	42.1	9.26	34.9	44.1
	ITC-irst	38.1	8.46	39.8	50.0
	UKA	33.4	7.96	43.3	54.5
Text	UPC	53.3	10.55	27.1	35.1
	IBM	53.1	10.38	27.0	35.9
	ITC-irst	47.5	9.60	31.3	40.6
	RWTH	46.1	9.68	29.7	40.5
	UKA	40.5	8.96	34.4	44.8
	UPV	32.7	6.80	41.3	47.5

Looking at the results shown in Table 24, we make the following observations, some of which are similar to the remarks for the English-Spanish EPPS task:

- There is a strong correlation between the four evaluation measures.
- There is only a comparatively small degradation when we switch from text to verbatim input and from verbatim to ASR input.
- The absolute performance is rather good, even slightly better than for the English-Spanish EPPS task.
- Looking at WER, i.e. word errors including the word positions, the error rates are only slightly increased, namely by about 10% absolute.

For both EPPS tasks, the overall performance is much better than for the Chinese-English BN task. There are several possible reasons like good match of training and test data for the EPPS tasks or the closer 'structural' similarity of Spanish and English as opposed to Chinese and English. To better understand the reasons, some analysis of the data and of the translation errors will be needed.

Comparing the two EPPS tasks, the direction Spanish-to-English shows consistently better performance than the direction English-to-Spanish. At the moment, the reason is not clear, in particular when we take into account that the we use exactly the same training corpus for both directions. Again, some analysis of the results will be needed.

5.4 Summary

In summary, the evaluation campaign has shown that the first 12 months of the project have been very successful:

- There has been an evaluation of the full system *speech recognition + translation* as opposed to the evaluation of individual components *speech recognition* and *speech translation*.
- To the best of our knowledge, this evaluation campaign was the first worlwide to cover speech translation for *real-life* data and a *large-vocabulary* task.
- Although the overall performance is not perfect, the error rates are unexpectedly low for such a *real life task*. E.g. the best system for the EPPS task can get 65% of the words correct, ignoring the word positions.
- There are now five operational research systems (one by each major partner in this workpackage); both these systems and the evaluation campaign are ahead of planning. (The original plan of the first evaluation did *not* cover: the whole ASR-SLT system, the EPPS task, external participants.)

There were two other international evaluation campaigns on machine translation:

- May 2004: NIST-DARPA evaluation of text translations:
– *Chinese-to-English* and *Arabic-to-English*
– newswire text, large vocabularies (50 000 words and more)
- Sep. 2004: CSTAR/ATR international workshop on spoken language translation:
– *Chinese-to-English* and *Japanese-to-English*
– travelling domain, limited vocabularies (about 8 000 words)

In each of these evaluations (apart from *Arabic-to-English*), many of the TC-Star partners participated with very good results.

6 Conclusions for Next Evaluation

For the next evaluation (in project month 23), we plan to stick to the translation tasks as defined for the first evaluation and extend them:

- **Broadcast news: Chinese to English.**
  This translation task will address two goals: First, we will have a direct comparison with ongoing DARPA projects. Second, the language pair Chinese-English will serve as an example of a non-European language pair to which the algorithms will be applied.
- **EPPS: English to Spanish.**
  Unlike the previous evaluation, we will focus on politicians' speeches. If possible, the politicians in the test data should be different from the politicians in the training data. In addition to EPPS, we will select a speech translation task with more *spontaneous* effects, e.g. EU hearings.
- **EPPS: Spanish to English.**
  It is not clear whether there will be enough speeches by Spanish EU politicians. If necessary, as a remedy, we will consider speeches from the national Spanish parliament (already recorded), which might result in a mismatch between training and test data and thus requires improved generalization capabilities.

We will preserve the three types of input to the translation system: ASR output, verbatim transcription and final text edition. Unlike the first evaluation campaign, the ASR output will include a *word graph* (along with probability scores) in addition to the single-best recognized sentence. In such a way, we will be able to study methods for counteracting speech recognition errors.

As for the evaluation procedure, there will be a couple of new issues to be studied:

- The manual segmentation of the speech input (in work package ASR) will be replaced by an *automatic segmentation* so that phrase boundary hypotheses (including punctuation marks) will be passed on to the translation systems. The translation systems will have to cope with this type of ambiguous segmentation. In addition, the evaluation measures (BLEU, NIST, PER, WER) will have to handle *non-segmented* translations.
- The outputs of multiple translation systems (both within and across sites) will be combined to improve the overall translation quality.
- A comparison with SYSTRAN (or other state-of-the-art translation products) will be included in the evaluation. In addition to the automatic evaluation, we will also consider *human evaluation* for selected translation results.

More details will be given in the updated implementation plan.

7 References

[Al-Onaizan 2004] Y. Al-Onaizan. Distortion-Based Word Reordering. In *Proceedings of DARPA Machine Translation Evaluation Workshop: IBM Site Report*, Alexandria, VA, USA, June 22–23 2004.

[Berger & Della Pietra$^+$ 96] A. Berger, S. Della Pietra and V. Della Pietra. A maximun entropy approach to natural language processing, In *Computational Linguistics*, vol.22, no.1, pp.39-71, 1996.

[Bertoldi & Cattoni$^+$ 04] N. Bertoldi, R. Cattoni, M. Cettolo, and M. Federico. The ITC-irst Statistical Machine Translation System for IWSLT-2004. In *Proc. of International Workshop on Spoken Language Translation, IWSLT'04*, Kyoto, Japan, 2004.

[Cettolo & Federico 04] M. Cettolo, and M. Federico. Minimum Error Training of Log-Linear Translation Models. In *Proc. of International Workshop on Spoken Language Translation, IWSLT'04*, Kyoto, Japan, 2004.

[Crego & Mariño$^+$ 04] J.M. Crego, J. B. Mariño and A. de Gispert. Finite-state-based and phrase-based statistical machine translation. In *Proceedings of the 8th International Conference on Spoken Language Processing, ICSLP'04*, pp. 37-40, October 2004.

[EPPS 05] EPPS: The European Parliament Corpus, which comprises Spanish and English transcriptions of the European Parliament sessions. The training and development data sets used are available on line at http://www.elda.org/en/proj/tcstar-wp4/tcs-slt-data.htm, 2005.

[FreeLing 05] FreeLing: an open source language analysis tool developed at the TALP research center, UPC. This tool is available on-line at http://garraf.epsevg.upc.es/freeling/, 2005.

[Gispert & Mariño 02] A. de Gispert and J. B. Mariño. Using X-grams for speech-to-speech translation. In *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP'02*, September 2002.

[Gispert & Mariño$^+$ 04] A. de Gispert, J. B. Mariño and J.M. Crego. TALP: Xgram-based spoken language translation system. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT'04*, pp. 85-90. Kyoto, Japan, October 2004.

[Gispert & Mariño 05] A. de Gispert and J. Mariño. Phrase linguistic classification and generalization for improving statistical machine translation. To appear in *Proccedings of the ACL 2005 Student Research Workshop*, Ann Arbor, Michigan, June 2005.

[Hasan & Ney 05] S. Hasan, H. Ney. 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 2005. To appear.

[ICTCLAS 05] www.nlp.org.cn/project/project.php?proj_id=6, 2005.

[Kanthak & Vilar$^+$ 05] S. Kanthak, D. Vilar, E. Matusov, R. Zens and H. Ney. Novel Reordering Approaches in Phrase-Based Statistical Machine Translation. In *Proc. of theACL 2005 Workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*, Ann Arbor, Michigan, June 2005. To appear.

[Kneser & Ney 95] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, MI, v.1, pp. 49-52, May 1995.

[Koehn 04] P. Koehn. Pharaoh: a beam search decoder for phrase-based SMT. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 2004.

[Koehn & Och$^+$ 2003] P. Koehn, F. J. Och and D. Marcu. Statistical phrase-based translation. In *Proceedings of HLT-NAACL Conference*, pp 48–54, 2003.

[Lee 04] Y-S. Lee. N-Best Reordering of Arabic for Statistical Machine Translation. In *Proceedings of DARPA Machine Translation Evaluation Workshop: IBM Site Report*, Alexandria, VA, USA, June 22–23 2004.

[Lee & Roukos 04] Y-S. Lee and S. Roukos. IBM Spoken Language Translation System Evaluation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)* 2004.

[Och 03] F.J. Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, Sapporo, Japan, July 2003.

[Och & Gildea$^+$ 04] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin and D. Radev. Syntax for statistical machine translation. *Final report on Johns Hopkins 2003 Summer Workshop*, Revised version: February 2004. Available on-line at: http://www.clsp.jhu.edu/ws2003/groups/translate/

[Och & Gildea$^+$ 04b] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference, NAACL'04*, Boston, MA, pp.161-168, May 2004.

[Och & Ney 00] F. J. Och and H. Ney. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China, 2000.

[Papineni & Roukos$^+$ 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.

[Popovic & Ney 05] M. Popovic, H. Ney. Exploiting Phrasal Lexica and Additional Morpho-syntactic Language Resources for Statistical Machine Translation with Scarce Training Data. In *Proceedings of EAMT 2005 (10th annual Conference of European Association of Machine Translation)*, Budapest, Hungary, May 2005. To appear.

[Press & Teukolsky$^+$ 02] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery. Numerical Recipes in C++: the Art of Scientific Computing, Cambridge University Press, 2002.

[RWTH Corpus 04] www-i6.informatik.rwth-aachen.de/∼tcstar/corpusEsEn.23nov2004.gz, 2004.

[RWTH Tools 04] www-i6.informatik.rwth-aachen.de/∼tcstar/corpusTools.24nov2004.tgz, 2004.

[SVMTool 05] SVMTool: a general POS tagger generator based on Support Vector Machines, developed at the TALP research center, UPC. This tool is available on-line at http://www.lsi.upc.edu/ nlp/SVMTool/, 2005.

[Stolcke 02] A. Stolcke. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pp.901-904, Denver, CO, September 2002. (This toolkit is available on line at http://www.speech.sri.com/projects/srilm/)

[Tillmann 03] C. Tillmann. A projection extension algorithm for statistical machine translation. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp 1–8, Sapporo, Japan, July 2003.

[Vogel 03] Stephan Vogel. SMT Decoder Disected: Word Reordering In *Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, 2003.

[Vogel & Hewavitharana^{+} 04] Stephan Vogel, Sanjika Hewavitharana, Muntsin Kolss, Alex Waibel. The ISL Statistical Machine Translation System for Spoken Language Translation. In *Int. Workshop on Spoken Language Translation*, Kyoto, Japan, 2004.

[Vogel & Ney^{+} 1996] S. Vogel, H. Ney and C. Tillmann: HMM-based word alignment in statistical machine translation. In *Proceedings of COLING-96*, pp. 836–841. 1996.

[Xu & Zens^{+} 05] J. Xu, R. Zens, H. Ney. Sentence Segmentation Using IBM Word Alignment Model 1. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, May 2005. To appear.

[Zens & Ney 04] R. Zens, H. Ney. Improvements in Phrase-Based Statistical Machine Translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pp. 257–264, Boston, MA, May 2004.

[Zens & Och^{+} 02] R. Zens, F.J. Och, H. Ney. Phrase-Based Statistical Machine Translation. In M. Jarke, J. Koehler, G. Lakemeyer, editors, *25th German Conf. on Artificial Intelligence (KI2002)*, Vol. 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, Springer Verlag, pp. 18–32, Aachen, Germany, September 2002.

[Zhang & Vogel^{+} 03] Ying Zhang, Stephan Vogel and Alex Waibel. Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation. In *Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, 2003.

[Zhang & Kikui^{+} 04] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, W. K. Lo. A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation. In *Proc. of The 20th International Conference on Computational Linguistics (COLING'04)*, Geneve, Switzerland, 2004.