



LC-STAR Deliverable D4.2

Project ref. no.	IST-2001-32216
Project acronym	LC-STAR
Project full title	Lexica and Corpora for Speech-to-Speech Translation Technologies
Security (distribution level)	Public
Contractual date of delivery	M08 = September 2002
Actual date of delivery	M20 = September 2003
Deliverable number	D4.2
Deliverable name	Description of language resources used for experiments
Type	Report
Number of pages	14
WP contributing to the deliverable	WP4
WP / Task / Deliverable responsible	WP4 - Task 4.2 - D4.2 UPC
Other contributors	RWT
Author(s)	Victoria Arranz (UPC) Núria Castell (UPC) Jesús Giménez (UPC) Hermann Ney (RWT) Nicola Ueffing (RWT)
EC Project Officer	Domenico Perrotta
Project Coordinator	Name: Harald Höge
Company:	Siemens AG, CT IC 5
Address:	Otto-Hahn-Ring 6, 81739 München, Germany
Phone:	+49-89-636-53374
Fax:	+49-89-636-49802
E-mail:	harald.h.hoege@mchp.siemens.de
Project web site:	http://www.lc-star.com
Keywords	Machine Translation, Speech Translation, Language Resources, Corpora
Abstract (for dissemination)	This document describes the LRs used in the first experiments as well as the experiments themselves. These experiments are described in detail, providing information on both the acquisition and expansion of already existing LRs.

Document evolution:

Version	Date	Security	Notes
V0.1	10.01.03	Project internal	First draft version - work document, to be discussed by WP partners
V0.2	06.05.03	Project internal	Pre-final version to distribute and discuss among WP partners
V1.0	15.09.03	Public	Final version

1 Introduction

The main goal of the LC-STAR project is to generate lexica and corpora for speech-to-speech translation systems. As reported on deliverable D4.1 [13], a number of approaches have been followed in order to perform speech to speech translation in the literature. One such approach that has proved to be very promising is that of statistical machine translation, which shows certain robustness against problems like speech recognition errors and is very efficient on the whole. Generally, such systems extract knowledge from aligned multilingual corpora. Thus, their performance is highly dependent on both the quantity and the quality of the data.

A crucial issue to deal with is that of figuring out what kind of information is required by those systems in order to produce good results, which, unfortunately, is not usually the case. As expected, this is not only a matter of determining the format of the data. There are a number of points yet to be argued, which involve aspects such as whether linguistic information (Part-of-Speech annotation, etc.) helps, and if so, what kind of information precisely, how to encode it, etc.

This deliverable will present the current state of progress for the different tasks and steps taken, as well as for the work still remaining. So far, some initial experiments have been carried out based on the resources already produced within the project. As it will be seen in section 4, details will be provided on the improvements achieved by using the new enriched language resources.

The deliverable is structured as follows: section 2 explains the existing lexical resources (LRs) that have been used and expanded in the project. Section 3 describes the different morphosyntactic annotations carried out for the three different languages under consideration (Catalan, English and Spanish). Section 4, as mentioned above, provides the preliminary translation experiments. Finally, the deliverable ends up with some conclusions drawn on the basis of the LR-development experience and the results obtained from the initial experiments.

2 LRs Used and Expanded

Before taking any decisions on which existing resources to use for our purposes, those available from LR repositories such as ELRA¹ and LDC² were considered. VerbMobil was chosen because it was manufactured from real recorded data and it focused on a semantically restricted domain, which fitted the purpose of the LC-STAR project very appropriately. A detailed description of the LR use and expansion can also be found in [2].

2.1 Verbmobil

Verbmobil was a long-term project of the German Federal Ministry of Education, Science, Research and Technology (BMBF, Projekträger DLR). One of

¹<http://www.elra.info>

²<http://www ldc.upenn.edu>

its outputs was a valuable set of spontaneous speech databases, which are now available to the public via ELRA. It also generated a speaker-independent and bidirectional SST system for spontaneous dialogues in mobile situations. As it can be seen, this project has generated very valuable output for the SST field, in what regards both machine translation technology and LRs for translation. Focusing on LC-STAR, our aim within the current workpackage is the creation of speech-to-speech centered and linguistically enriched trilingual LRs for Catalan, English and Spanish, which did not exist up to the date in those repositories mentioned in the previous section. It was then decided to take the VerbMobil corpus as starting point, given that it was based on real recorded conversations for a semantically restricted domain. This has allowed us to focus on the appointment scheduling domain. At a later stage, as reported on deliverable D5.3 [1], further LRs would be created from scratch focusing on some particular tourism-related subdomains, which still remained semantically related to the VerbMobil databases that have been used.

2.2 Selected Material

Out of all those databases available from VerbMobil, we selected 9 of them that contained recordings in English. We avoided those databases that had recordings in German and Japanese, as well as those containing mixed languages. The aim was to start from the US-English recordings and generate their counterparts in Catalan and Spanish by means of human translation. Working with mixed-language data might have been more confusing than helpful given that it would have required a considerable effort towards the selection of the English-language utterances. However, it should be mentioned that when selecting the data, we were relatively flexible with the speakers and allowed for both US-English and Denglish (English spoken by Germans), even if this may imply some non-real English word/expression to be uttered. In the interest of the size of the data to be used, we considered it appropriate.

The databases selected and purchased by UPC from ELRA were the following:

- VM CD 6.1 - VM61 (new edition)
- VM CD 8.1 - VM81 (new edition)
- VM CD 13.1 - VM13.1 (new edition)
- VM CD 23.1 - VM23.1 (BAS edition)
- VM CD 28.1 - VM28.1 (BAS edition)
- VM CD 31.1 - VM31.1 (BAS edition)
- VM CD 42.1 - VM42.1 (BAS edition)
- VM CD 43.1 - VM43.1 (BAS edition)
- VM CD 50.1 - VM50.1 (BAS edition)

A total amount of 287,655 tokens has been collected, resulting in a vocabulary size of 3,333.

2.3 Cleaning and Preservation Criteria

The VerbMobil³ corpora are annotated with various tags, like e.g. punctuation marks and pause markers. Some of those tags have been preserved because they were believed to be necessary for translation, in the sense that they could add some meaning to the discourse. This is the case for:

Punctuation marks: The insertion of standardized marks or signs in written documents to clarify the meaning and the separate structural units of sentences and utterances.

Filled pauses: The sound produced during spontaneous speech that represents a pause filled by a vocalization.

Foreign words: In transcription, the Foreign Word refers to words that appear in a transcript but do not belong to the same language as the primary language of the transcript.

Interjections: Expressions of surprise ('oh'), affirmation ('uh-huh', 'mhm'), negation ('mm', 'uh-uh') and discourse particles (such as 'well', 'anyway') are examples of interjections.

Proper names: The use of a specific marking convention to label certain proper nouns in transcriptions. These tags were initially just placed at the beginning of the proper name. They have been extended so as to indicate where the proper name ends as well.

Letter spelling: The act of spelling out a word or abbreviation.

Abbreviations: In order to differentiate them from letters, words that are acronyms and should be pronounced or spoken as one word - such as OPEC, ASCII, etc. - are not annotated.

Neologisms: *Neologism* is the term used to refer to a word that has been made-up by a speaker and that appears in a transcript of spontaneous speech dialogue. It can also be described as a word which neither appears in the dictionary of the primary spoken language nor is a foreign word.

Technical interruptions: A temporarily broken or missing section of the audio signal caused by technical disruptions, distortions, or disturbances in the recording equipment, or as a result of recording mistakes.

Turn breaks: If a speaker contribution is aborted immaturely, it is not possible to use '.' or '??' at the end.

Unidentifiable/Hard to identify The inability to understand what someone is saying.

³http://www.is.cs.cmu.edu/trl_conventions/projects/verbmobil.html

2.4 Translation

Those VerbMobil’s dialogues described as selected material (cf. section 2.2) focus on appointment scheduling and travel planning. They were considered such an appropriate source of information that it was decided to have them translated into Catalan and Spanish so as to later generate a trilingual aligned corpus.

Dialogues in VerbMobil look as follows:

e001ach2_000_ANV_230000: hi ~Mary, how are you.

e001ach1_001_SMA_230000: oh I am doing fine Mister ~Vandaloo,
how are you.

e001ach2_002_ANV_230000: pretty good, <uhm> I guess we need to
figure out a day, today is the first
of November, so within the next two
months when we can make it to, ~Hanover?

e001ach1_003_SMA_230000: that sounds like a good idea, let us see,
I have, <uhm> the, nineteenth twentieth
and twenty first I am available to travel.

(...)

As it can be observed in the example above, turn headers within the dialogues contain the following useful reference information:

- Dialogue identifier
- Turn number
- Speaker

Once the source-language dialogues are translated, two further turn headers are appended to every turn in every dialogue so as to index the utterances of the translations by means of an added language marker (CAT for Catalan, ENG for English and SPA for Spanish, see example below). This simplifies any search to be performed in the corpus at a later stage. The original turn headers are also modified accordingly as a matter of coherence, inserting the relevant language marker. Different turn headers with their language markers can be seen in the example below:

e001ach2_000_ANV_230000_ENG: hi ~Mary~, how are you.

e001ach2_000_ANV_230000_SPA: hola ~Mary~, cómo está.

e001ach2_000_ANV_230000_CAT: hola ~Mary~, com està.

e001ach1_001_SMA_230000_ENG: oh I am doing fine Mister ~Vandaloo~,
how are you.

e001ach1_001_SMA_230000_SPA: oh estoy bien señor ~Vandaloo~,
cómo está.

e001ach1_001_SMA_230000_CAT: oh estic bé senyor ~Vandaloo~,
com està.

We have spent over 1,000 hours in the human translation and over 200 hours in the automatic and human verification and correction of the VerbMobil selected material. As a result, we have obtained a Spanish corpus of 281,848 tokens (yielding a vocabulary size of 5,149 words), and a Catalan corpus of 277,955 running words (yielding a vocabulary size of 5,145 words).

3 Part-of-Speech Tagging

Once the VerbMobil corpus had been translated and validated, we moved onto the morphosyntactic annotation of the data. As already mentioned, good results have already been obtained in earlier attempts of statistical machine translation. However, all experts agree on the fact that a crucial issue here is the type of data used for the training and knowledge learning. Thus, bearing in mind the needs of the expert researchers and developers with whom we are working closely, LC-STAR is developing more sophisticated and linguistically enriched LRs. For that purpose, we have considered that providing part-of-speech (POS) information for the corpora developed was certainly the first step to take. Since we are working with three different languages, different tools have also had to be considered for their morphosyntactic labelling, in particular, in what refers to English. Our group at UPC (Barcelona) does already have a set of tools for the annotation and disambiguation of both Catalan and Spanish (cf. section below).

3.1 Morphosyntactic Annotation of Catalan and Spanish Data

The POS tagging of Catalan and Spanish data has been performed with our morphological analyser *MACO+* (*Morphological Analyzer Corpus Oriented*) [6], a robust and wide-coverage tool that accepts unrestricted text as input and provides all possible labels and lemmas for each word form, i.e., produces all possible morphological interpretations for each token. It is able to recognize and deal with numbers, proper nouns, punctuation, dates, abbreviations and multiwords.

The set of tags used to represent the morphological information is based on

those proposed by EAGLES⁴ for the morphosyntactic annotation of lexica and corpora for all European languages.

MACO+'s output is disambiguated by *RELAX* (*Relaxation Labelling Based Tagger*) [10]. This tool selects the POS and lemma for each word according to the given context. Currently, it produces an output with over 97% precision. The language model may be easily improved with the addition of new context constraints expressed in CG formalism, either hand-written or statistically acquired.

Although both these tools are also available for English language, their performance is not as accurate as for Catalan and Spanish. Thus, other tools specifically built for English were considered for the task.

3.2 Morphosyntactic Annotation of English Data

After considering a few tools available for English language, it was decided to use Eric Brill's POS tagger [4]. Its tags are based on the Penn Treebank project⁵. Thorsten Brants' TnT HMM-based tagger[3] has also been tried exhibiting a similar performance.

Eric Brill's transformation-based error-driven learning tool was chosen for the following reasons:

- **It is highly portable:** while other successful tools have been designed for a particular sublanguage, Brill's tagger's trainable nature offers an inexpensive porting capability, even if the new domains (and new languages) present completely different grammar, structural specification or even lexical entries.
- **It can be trained on tagged corpora:** although we did not intend to train it at first, it was interesting to consider this possibility as a plus for our choice. At later stages of our LR development, we may consider it relevant to train the tool on some particular data or further language.
- **Small size of training data:** Brill's tagger can be used with small training corpora. This system can start from scratch by making use of no-language or domain-specific knowledge and then be trained to obtain this knowledge from relatively small corpora.
- **It is a simpler tool to use:** It is a relatively simple tool to use and get acquainted with, both for tagging based on default values or for training on some specific data.

Before mentioning some of the problems encountered during our POS tagging process, it should just be added that as a first approach, Brill's tagger has been used as-is, so as to see what was capable of. As it will be seen in the following section, though, a few problems popped up that had been feared but not fully anticipated, and that had to be dealt with.

⁴<http://www.ilc.cnr.it/EAGLES96/home.html>

⁵<http://www.cis.upenn.edu/~treebank/home.html>

3.3 Problems Encountered during the POS Tagging Process

As mentioned above, conversations may contain several ungrammaticalities such as false starts, corrections, repetitions, filled pauses, etc. These phenomena cause the POS-tagger performance to significantly decrease. Obviously, the training data differ significantly from the data to be tagged. That is a major drawback in our research that must be overcome. If morphosyntactic information is to be used it should be taken into account that the error in tagging is surely going to propagate into the rest of the system.

The next section describes how parts of the created trilingual corpora have been used for setting up an experimental SST system and what results have been obtained.

4 Translation Experiments with Expanded Corpora

This section describes the translation experiments conducted using the corpora which are being built within the LC-STAR project and which have been described in section 2. These experiments take transcriptions of recorded dialogs as input, i.e. the training and testing material consists of written text that contains the characteristics of spontaneous speech – such as hesitations and syntactic errors – but without speech recognition errors. Section 4.1 describes the goals of these experiments as well as ideas for the incorporation of the language resources into the translation process. The trilingual corpus that is successively being built within the project is described in Section 4.2. Baseline results for two different versions of this corpus that have so far been accomplished will be presented in Section 4.3. Further details on these experiments can also be found in [12].

4.1 Goals

The goal of WP4 is to investigate the impact of the amount and type of training data on speech translation quality. Since the acquisition of language resources (LRs) is expensive, the relation between data acquisition costs and translation quality is to be analysed.

Different methods for the incorporation of the Part of Speech (POS) information and the baseforms of words will be explored. In a statistical machine translation (SMT) system, two points are possible for the integration of the additional information into the translation process: the *translation* model and the *language* model. In WP4, an analysis of the effect of the integration into both of the models is planned.

In the translation model, the expanded corpora can be used along various directions:

- Methods to overcome data sparseness for the highly inflected languages Catalan and Spanish, e.g. with hierarchical lexicon models as proposed in [7].

- The use of the additional information as features in maximum entropy lexicon models.
- The improvement of alignment quality by taking the additional information into account.
- Approaches that enrich the English part with information to help translation into the morphologically richer languages.

The language model (LM) can be extended as follows:

- POS-based LMs (e.g. n -grams with longer history than the standard trigram).
- LMs based on lemmata rather than fullform words.
- Differentiation between different sentence types (e.g. interrogative clause, clause of statement, etc.).
- Interpolation of different LMs.

The integration of the improved models into the translation process can be performed in two different ways: they can either be incorporated directly into the search process, replacing the baseline models. Or a first pass of the search can be carried out (using the baseline models) to create an N -best list or a word graph as described in [14]. This can then be rescored in a second search pass. This method is recommended for more complex models which would otherwise yield large modifications of the search process.

A further possibility for the use of the LRs is to first translate from English into the Catalan or Spanish baseforms, because we found in prestudies that this reduces the word error rate for about 10%. A second step would then be needed that produces the flexion of the word. Nevertheless, this requires a target-language grammar that can determine the correct fullform.

4.2 Data Description

Within the project, a trilingual sentence-aligned corpus is successively being built. It comprises the Catalan, English and Spanish languages. At the time of this report, two versions of the corpus had been accomplished, which from now on will be referred to as *Corpus 1* and *Corpus 2*. For both corpora, data from VerbMobil were selected, translated and tagged. *Corpus 1* consists of about 4,6k sentences; which has been extended to about 13k in *Corpus 2*. The statistics are given in Tables 1 and 2.

The corpus consists of transcriptions of spontaneously spoken dialogues. Thus, the sentences often lack correct syntactic structure and contain hesitations. The domain of this task is appointment scheduling and travel arrangements.

For Catalan and Spanish, the baseforms and the Part of Speech labels of the words are given as additional knowledge sources. The lemmatization (see section 3) was produced using the morphosyntactic processing tools from UPC

Table 1: Training and test conditions on *Corpus 1* (*number of words without punctuation).

		English	Spanish	Catalan
Training:	Sentences	4 643		
	Words	37 728	36 486	36 895
	Words*	29 817	28 585	28 989
Vocabulary	Size	1 227	2 156	2 070
	Singletons	419 (34%)	1 031 (48%)	967 (47%)
Develop:	Sentences	265		
	Words	2 171	2 069	2 089
	Unk. words	18 (0.8%)	41 (2.0%)	43 (2.1%)
Test:	Sentences	250		
	Words	2 097	2 010	2 026
	Unk. words	36 (1.7%)	44 (2.2%)	37 (1.8%)

Table 2: Training and test conditions on *Corpus 2* (*number of words without punctuation).

		English	Spanish	Catalan
Training	Sentences	13 352		
	Words	123 454	118 534	118 137
	Words*	101 738	96 997	96 503
Vocab.	Size	2 154	3 933	3 572
	Singletons	790 (37%)	1 844 (47%)	1 658 (47%)
Develop	Sentences	272		
	Words	2 267	2 217	2 211
	Unk. Words	21 (0.9%)	34 (1.5%)	34 (1.5%)
Test	Sentences	262		
	Words	2 626	2 451	2 470
	Unk. Words	17 (0.6%)	30 (1.2%)	35 (1.4%)

Barcelona: MACO+ and RELAX (see section 3.1).

The English part of the corpus was annotated with Part of Speech (POS) information. This was produced using Eric Brill’s tagger (cf section 3.2), which can be downloaded from <http://www.research.microsoft.com/users/brill/>.

4.3 Experiments and Results

We performed baseline experiments on the two settings *Corpus 1* and *Corpus 2* that are described in Section 4.2. In these experiments, only the bilingual texts containing the fullform words were taken for training and testing.

We set up two SMT systems for each of the language pairs: A single word based system which implements the so-called Model IBM-4 as introduced in [5]. For details about this system, the reader is referred to [9, 11]. The other system

is based on groups of words, so-called Alignment Templates, rather than single words. It is described in [9, 8].

As evaluation metrics, we used the word error rate (WER) and the BLEU score as described in [13]. Since BLEU measures quality, we transform it into an error measure by determining 100-BLEU.

Tables 3 and 4 contain an assessment of translation quality of the two systems on the two setups. We performed experiments on all six language pairs.

Table 3: Translation Error Rates [%] on *Corpus 1* for two SMT systems.

		Develop		Test	
Translation	System	WER	100-BLEU	WER	100-BLEU
Spanish-English	AT	30.1	52.4	28.2	48.0
	SWB	32.2	53.3	30.3	49.1
Catalan-English	AT	31.1	51.6	30.8	50.7
	SWB	33.4	51.8	33.4	52.7
English-Catalan	AT	33.5	54.6	33.8	55.5
	SWB	38.8	59.2	37.2	56.9
English-Spanish	AT	35.5	57.0	34.0	54.6
	SWB	38.0	57.8	38.8	58.6
Spanish-Catalan	AT	12.7	25.4	13.8	26.5
	SWB	15.3	28.0	16.4	28.8
Catalan-Spanish	AT	17.0	32.1	16.6	29.6
	SWB	18.9	35.7	19.1	31.8

As the results show, the translation quality increases with the amount of bilingual training data that is available.

The results can be grouped into three different levels of difficulty: obviously, the translation between Catalan and Spanish is the easiest task. This is not surprising, since this is the most similar language pair with respect to word order and degree of language inflection.

Translations from Spanish into English and from Catalan into English are of similar severity. The structural differences between English and either of the other two languages are the same: the word order varies as well as the richness of morphology. The vocabulary sizes of Spanish and Catalan are in the same range, whereas the English part of the corpus has a much smaller vocabulary and a lower rate of singletons.

Moreover, the number of running words in the corpus differs between English on the one hand and Catalan and Spanish on the other one. The English text comprises more words than the Catalan and Spanish parts which is also due to structural differences between the languages.

The most difficult task is the translation from English into Catalan or into Spanish. This is caused by the high level of inflection of the two languages compared to English. Moreover, the English word forms often do not contain

Table 4: Translation Error Rates [%] on *Corpus 2* for two SMT systems.

		Develop		Test	
Translation	System	WER	100-BLEU	WER	100-BLEU
Spanish–English	AT	28.8	50.9	27.2	44.8
	SWB	30.6	51.2	27.7	44.8
Catalan–English	AT	30.7	53.8	27.7	45.7
	SWB	32.2	53.7	28.7	45.0
English–Catalan	AT	34.1	54.8	28.0	45.0
	SWB	37.6	58.2	33.0	49.2
English–Spanish	AT	34.7	54.2	30.4	47.6
	SWB	35.4	57.6	32.1	48.9
Spanish–Catalan	AT	14.6	26.5	15.0	27.2
	SWB	17.1	30.1	18.2	31.9
Catalan–Spanish	AT	17.2	31.7	17.9	30.8
	SWB	16.4	31.8	18.4	31.0

enough information in order to choose the correct inflection in Catalan and Spanish.

The problems described above will be tackled in the future by the extensions to the models proposed in Section 4.1. The methods presented there should make it easier to handle the structural differences between the language pairs that are dealt with in the LC-STAR project.

5 Conclusions

This document has described the initial languages resources that we have used in the LC-STAR project as a starting point both to create new LRs, and to make initial translation experiments.

The selected material in English has been translated into Catalan and Spanish. Further, the corpora have been expanded by adding morphosyntactic information. These expanded corpora have been used to start with translation experiments, which have proved that linguistic information can improve statistical machine translation results.

The next step in the project is to study translation improvements by working with a bigger trilingual corpus, as well as adding more information, and even by using more resources such as, for instance, monolingual lexica.

References

- [1] V. Arranz, N. Castell, and J. Giménez. Description of raw corpora. Technical Report Deliverable D5.3, LC-STAR project by the European Community (IST project ref. no. 2001-32216), 2003.

- [2] V. Arranz, N. Castell, and J. Giménez. Development of language resources for speech-to-speech translation. In *RANLP-03*, Borovets, Bulgaria, September 2003.
- [3] T. Brants. Tnt – a statistical part-of-speech tagger. In *ANLP-00*, Seattle, WA, USA, April-May 2000.
- [4] E. Brill. *A Corpus-Based Approach to Language Learning*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, Pittsburgh, PA, USA, 1993.
- [5] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- [6] J. Carmona, S. Cervell, L. Màrquez, M. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An environment for morphosyntactic processing of unrestricted spanish text. In *LREC-98*, Granada, Spain, September 1998.
- [7] S. Nießen and H. Ney. Toward hierarchical models for statistical machine translation of inflected languages. In *ACL01-DDMT*, pages 47–54, Toulouse, France, July 2001.
- [8] F. J. Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, Computer Science Department, RWTH Aachen – University of Technology, Germany, 2002.
- [9] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, USA, June 1999.
- [10] L. Padró. *A Hybrid Environment for Syntax-Semantic Tagging*. PhD thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain, 1997.
- [11] C. Tillmann. *Word Re-Ordering and Dynamic Programming based Search Algorithm for Statistical Machine Translation*. PhD thesis, Computer Science Department, RWTH Aachen – University of Technology, Germany, 2000.
- [12] N. Ueffing and H. Ney. First experimental results on baseline speech-to-speech translation systems. Technical Report Deliverable D4.4, LC-STAR project by the European Community (IST project ref. no. 2001-32216), 2003.
- [13] N. Ueffing, H. Ney, V. Arranz, and N. Castell. Overview of speech centered translation. Technical Report Deliverable D4.1, LC-STAR project by the European Community (IST project ref. no. 2001-32216), 2002.

- [14] N. Ueffing, F. J. Och, and H. Ney. Generation of word graphs in statistical machine translation. In *EMNLP-02*, pages 156–163, Philadelphia, PA, USA, July 2002.