



## LC-STAR Deliverable D0.2.6

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	2005-03-30
Actual date of delivery	2005-04-28
Deliverable number	D0.2.6
Deliverable name	Final Report
Type	Management Report
Status & version	Final Version 1.0
Number of pages	22
WP contributing to the deliverable	WP0
WP / Task / Deliverable responsible	WP0/D0.2. SIE
Other contributors	
Author(s)	Ute Ziegenhain (SIE)
EC Project Officer	Kimmo Rossi
Project Coordinator	Name: Harald Höge Company: Siemens AG, CT IC 5 Address: Otto-Hahn-Ring 6, 81739 München, Germany Phone: +49-89-636-53374 Fax: +49-89-636-49802 E-mail: <a href="mailto:harald.hoege@siemens.com">mailto:harald.hoege@siemens.com</a> Project web site: <a href="http://www.lc-star.com">http://www.lc-star.com</a>

### Document evolution:

Version	Date	Status	Notes
1.0	April 2005	Final	

## 0. Contents

1	Results of the Project.....	3
1.1	Overview and status of deliverables in chronological order .....	4
1.2	Specifications and development of language resources for ASR and TTS Systems (WP1/2/3/5).....	6
1.3	Specifications and development of language resources for Statistical Machine Translation Systems (SMT) (WP5) .....	8
1.3.1	Specifications and development of bilingual aligned corpora .....	8
1.3.2	Specifications and Development of 'Phrasal Lexica'.....	10
1.4	Specifications for validation of LR (WP6).....	11
1.4.1	Specification of validation of language resources for ASR and TTS .....	11
1.4.2	Specifications of validation criteria of language resources for SMT .....	12
1.4.3	Validation results (WP3/5/6).....	13
1.5	Results in SMT Research (WP4/7).....	13
1.6	Overview of Demonstrator (WP4/7) .....	16
1.6.1	Description of the Kernel .....	17
1.6.2	Acceptance testing of the demonstrator .....	18
2	Dissemination and Use (WP7) .....	19
3	References .....	22

# 1 Results of the Project

The objectives of the project were:

- Specifications and creation of large lexica specifically designed for voice driven applications in ASR and TTS systems covering a broad range of common domains and proper nouns. The lexica are produced for 12 (+1) languages from all over the world: Catalan, Finnish, German, Greek, Hebrew, Italian, Mandarin, Russian, Slovenian, Spanish, Standard Arabic and US-English. The Slovenian language resources are provided by our external partner. The tasks covered deliverables D1.1, D1.2, D3.0, D3.1/2 of the corresponding workpackages.
- Specifications and creation of lexica for ASR and TTS in a tourist domain specified within the project in 9 languages: Catalan, Finnish, German, Hebrew, Italian, Russian, Slovenian, Spanish and US-English. The tasks covered deliverables D5.1, D5.2, D5.4, D5.6.
- Specifications and creation of large bilingual aligned corpora in the tourist domain in the languages Catalan, Spanish and US-English. The tasks covered deliverables D5.3, D5.5. and D5.7
- Specifications and creation of bilingual aligned 'phrasal' lexica designed for speech-to-speech translation systems in 9 languages including: Catalan, Finnish, German, Hebrew, Italian, Russian, Slovenian, Spanish and US-English in the tourist domain specified within the project. The tasks covered deliverables D5.5, D5.6. and D5.8.
- Development of a statistically based translation baseline system and research in statistically based translation using different structured language resources. The tasks covered all deliverables in workpackage 4.
- Development and acceptance testing of a demonstrator platform demonstrating the system. The tasks covered deliverables D4.6, D7.1, and D7.2
- Validation of all language resources. The tasks covered all deliverables in workpackage 6.
- Dissemination und exploitation of the results. The tasks covered deliverables D7.3, D7.4, D7.5, D7.6

The consortium consisted of 6 (+1) partners: four industrial partners namely IBM, Nokia, NSC (Natural Speech Communication), and Siemens and 2 universities, RWTH-Aachen (Rheinisch-Westfälische Technische Hochschule Aachen) and UPC (Universitat Politècnica de Catalunya) and one external partner the University of Maribor, Slovenia. Three independent validation centres namely CST (Center for Sprogteknologi), Denmark, SPEX (Speech Processing EXpertise), The Netherlands and the University of Vigo, Spain are responsible for validating the language resources. Project partners have wide experience either in previous speech database projects (e.g. Speech-Dat family, Speecon etc.) and/or projects relating to machine translation (e.g. Verbmobil). The selected validation centres were highly qualified to perform the tasks.

Main results which are described in more detail in the following paragraphs are:

- Detailed specifications for corpora collection and wordlist creation, lexicon development, development bilingual aligned corpora and of bilingual phrasal lexica as well as definitions of validation criteria and procedures.

These specifications can be seen as pioneering work in this field and are publicly available from the project's web page for use in other projects.

- Creation and validation of high quality language resources.

- Research and experimental results in statistical machine translation using different structured language resources produced within the project.
- Development of a demonstrator platform 'Gaia' and acceptance testing of the platform.
- Dissemination and use of the language resources also via ELRA and other projects like TC-STAR<sup>1</sup> and ECESS<sup>2</sup>. Software for the demonstrator platform and DTD for large lexica will be made publicly available.

At the end of the project all deliverables were final except for one database in Workpackage 3 and exchange of the validated language resources between the partners has started. The project has quite successfully passed the final review in January.

The following pages are organized as follows: first an overview of all deliverables is given, a summary of the specifications of all language resources as well as the validation results are presented (Workpackges 1, 2, 3, 5, 6). The following chapters concentrate on the research results and the development and testing of the demonstrator (Workpackages 4 and 7). The final chapter present the dissemination and exploitation results (Workpackage 7).

## 1.1 Overview and status of deliverables in chronological order

As already mentioned except for one database in Workpackage 3 all deliverables have been finished during the lifetime of the project. Download information from the web page is given below.

Del. No	Deliverable title	Nature	Status/download	Lead Participant
D1.1	Specification of corpora and word lists in 12 languages	R	finished	SIE
D7.3	Installation of Web-Platform	O	finished	NOK
D2.1*	Language-independent specification of contents of lexica	R	finished	IBM
D2.2*	Report on properties of each language relevant for determining effort	R	finished	IBM
D4.1	Overview on speech centered translation technology	R	finished	RWT
D2.3*	Definition of representation format	R	finished	IBM
D5.1	Creation of reference word list in US-English	P	finished	UPC
D1.2	Creation of word lists	P	finished	SIE
D2.4*	Language-dependent specification of the contents of the lexicon for each language	R	finished	IBM
D4.2	Description of LR used for experiments	R	finished	UPC
D4.3	Development of an experimental platform for speech-to-speech translation systems	O	finished	RWT
D5.2	Creation of word lists in 7 languages	P	finished	UPC
D5.3	Description of raw corpora	P	finished	UPC

<sup>1</sup> [www.tc-star.org](http://www.tc-star.org)

<sup>2</sup> [www.eccess.org](http://www.eccess.org)

D6.1	Specification of validation criteria for lexica for recognition and synthesis	R	finished	NSC
D7.6	Dissemination and Use Plan	R	finished / internal web	NOK
D4.4	First experimental results on baseline for speech-to-speech translation systems	R	finished	RWT
D5.4	Creation of lexica for speech recognition and speech synthesis	P	finished	UPC
D3.0 (new)	Specifications of lexicon interchange format	R	finished / internal web	UPC
D3.1 *	Creation of lexica for recognition	P	finished (except one language)	UPC
D3.2 *	Creation of lexica for synthesis	P	finished (except one language)	UPC
D5.5	Language independent specification of LR for translation	R	finished	UPC
D5.6	Language specific specification of LR for translation	R	finished	UPC
D4.5	Results on different structured LR for speech-to-speech translation systems	R	finished / internal web	RWT/UPC
D6.2	Validation of lexica for recognition and synthesis	R	finished / internal web	NSC
D6.3	Specification of validation criteria for LR for speech centered translation	R	finished	NSC
D4.6	Development of Demonstrator	D	finished	UPC
D5.7	Creation of aligned text corpora (3 language pairs)	P	finished	UPC
D5.8	Creation of lexica for speech centered translation (8 languages)	P	finished	UPC
D7.5	Distribution schemes for LR	R		NOK
D6.4	Validation of LR for speech centered translation	R	finished / internal web	NSC
D7.1	Demonstration of language transfer	O	finished	UPC
D7.2	Acceptance Testing of the Demonstrator	R	finished	UPC
D7.4	Awareness actions	O	finished	NOK
D7.7	Technological Implementation Plan	R	finished	NOK

\* now one document

For downloading documents/information please use the following links/keywords.

Address of webpage:

<http://www.lc-star.com/>

All deliverables of type report can be can be downloaded from the webpage following keyword: 'public documents' in not otherwise stated:

<http://www.lc-star.com/archive.htm>

Other deliverables may be downloaded from the internal web pages on request (please contact: ute.ziegenhain@siemens.com).

In addition three public management reports are available: D0.3.1. Annual Public Report, D0.3.2. Annual Public Report , D0.3.3. Annual Public Report

## 1.2 Specifications and development of language resources for ASR and TTS Systems (WP1/2/3/5)

The specifications for a) corpora collection and wordlist creation, b) lexicon development and c) definitions of the validation procedure and standards generated can be seen as pioneering work in this field.

For corpora collection and wordlist creation (Ziegenhain, et al., 2003) a broad range of common domains and domains for proper names have been chosen to be collected from electronically available resources.

The lexica consisted of : at least 50,000 entries in the common domains, at least 45,000 entries in proper names and special application entries covering typical voice driven applications.

For proper names 3 major domains were covered: person names (including first and last names), place names (including street names, etc.) and organizations. Each of these three major domains for proper names had to cover a minimum of 10% and a maximum of 50% of all 45,000 entries due to specific geographic and cultural peculiarities in the countries (e.g. few geographic names in Finland, few last names in China)

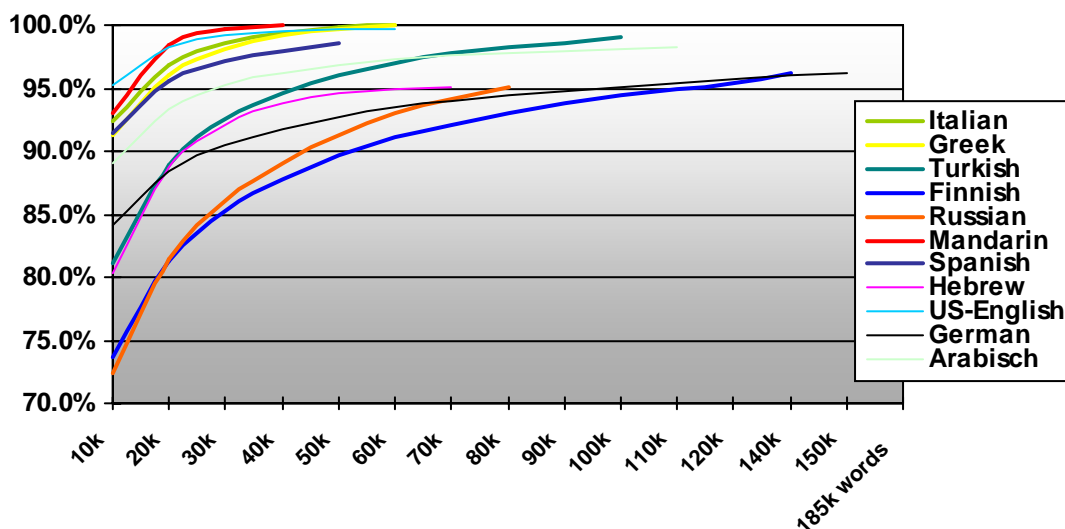
Special applications domains consisted of common domains relevant for all applications like for example: all numbers, digits and dates, entries for measures, greetings, abbreviations, spelling, etc.. Additionally special entries for information services (e.g. banking, weather forecast), information retrieval (sports, news) and telecommunications (web, computer and networking devices) have been collected.. A reference wordlist of 5.700 entries in US-English was build and then translated in all target languages.

For common words 6 major semantic domains were covered: consumer information, culture (including travelling), finance, news, personal communications, sports and games. The required minimum size of the corpus was 10 million tokens of 'cleaned' text (e.g. digits, punctuation marks, typos etc. removed). Due to the coverage requirements described below the actual collected amounts were much higher (cf. overview of coverage).

For optimizing the criterion, the common word lists had to achieve a self-coverage of at least 95% in each domain and at least 95% over all domains. Furthermore, the final wordlist had to contain at least the most frequent 50,000 entries without singletons, abbreviations and proper names. The formal procedure consisted of language-dependent cleaning and tokenizing of the corpora, verifying the requirements as well as removing proper names, typos, abbreviations etc. manually if necessary. Different methods were provided to remove the proper names automatically but this approach did not apply to all languages (especially for those with no capitalization (e.g. Hebrew, German, Mandarin). Self-coverage was re-computed and re-checked so that it was at least 95%. In the last step, the word lists for all domains were merged to form a single word list. In case the merged word list contained less than 50,000 entries, the coverage target was increased and the whole procedure was iterated beginning from counting the number of occurrences of all distinct tokens in the corpus. The iteration continued until the final word list reached at least 50,000 entries or the coverage of 100% was attained.

The common wordlist sizes ranged from 38,564 entries covering 100 % of a corpus of over 20 million tokens for Mandarin to 140,000 entries for Finnish covering 95% of 17 million of tokens. These results reflect the morphological diversity of the languages: Mandarin without any morphological information while Finnish is well known as highly a inflected language. Another example is German which has a rich morphological system and additionally a high compounding system: 100,000 entries had to be collected to meet

the requirements due to word compounding although de-compounding methods have



been applied.

#### Overview coverage

In addition a list of so called closed set (function) word classes (e.g. prepositions, pronouns, conjunctions, etc.) has been included in the final word list (lexicon) for all languages. Function words are frequently used in various domains however not all will be covered by corpora collection. Those which are not contained in the corpora were added manually.

It is crucial that lexica contain enough grammatical, morphological and phonetic information required by the ASR/TTS components in SST applications. One of the initial tasks in the project was to specify the grammatical information needed for each language (Maltese et al., 2003). The resulting information was merged into a unique list of part-of speech (POS) tags. Most of the POS tags have an internal structure with attributes that may be either common to several languages (e.g. number) or specific only to a subset of languages (e.g. case) or concern only specific individual languages (e.g. polarity for Turkish verbs). The advantage of the chosen approach is that a single description of grammatical features can cover the whole set of 13 languages. For each word in a given lexicon, lemma and phonetic information are specified along with the POS information as described above. In agglutinative languages, such as Turkish and Hebrew, some morphological boundary information is also marked. Phonetic transcriptions use the SAMPA symbols, available in each language, and include information concerning primary stress, syllable boundaries and word boundaries for multi-word entries where a pause may occur in the utterance. Multiple POS, lemmas and phonetic transcriptions can be specified for a given word. Information included in the lexica is coded with an XML-based mark-up language that represents the linguistic information in a formal and unambiguous manner. Representation with XML is both easy to read and to process. The content information described so far is included in a Document Type Definition (DTD), a formally specified grammar that covers all languages in the project and is easily extendible to other languages as well.

A lexicon consists of a collection of entry group elements. An entry group refers to a given word, whereby its spelling is the key to the entry group and is therefore obviously

mandatory. An entry group consists of one or more entries and can also contain compound entry elements, which are used in agglutinative languages (see below). An entry refers to a specific grammatical/morphological role of a vocabulary entry, e.g. ‘can’ (verb) and ‘can’ (noun). For each entry POS, lemma and phonetic transcriptions must be specified. For abbreviations, multiple expansions can be specified. Assimilation or agglutination phenomena occur in some languages (e.g. Catalan, Hebrew, Italian, Spanish, Turkish) and are tackled via so called compound entries. Besides its spelling, its phonetic transcription and its (optional) lemma, a compound entry consists of two or more compound entry elements (i.e. a subset of a compound entry) which are simply links to other entries. Each compound entry element must have an orthography and a full grammatical tagging (i.e. POS and all of its attributes). The union of the two unambiguously identifies the compound entry element. In the lexicon, only the information relevant to the target language has to be specified; each attribute therefore has as default value NS (=Not Specified), thereby avoiding the need to specify non-existing or non-relevant features of the language (e.g. case in Italian or gender in Finnish, which do not exist).

### **1.3 Specifications and development of language resources for Statistical Machine Translation Systems (SMT) (WP5)**

One of the main research aim in the project was to study the usefulness of different structured language resources for statistical based translation systems especially with respect to the scarce data problem which is a main drawback of all statistical approaches: large aligned databases are rarely available and are very costly to produce. Therefore different structured resources have been created and used for research. The language resources are described in more detail below.

#### **1.3.1 Specifications and development of bilingual aligned corpora**

One of the most powerful resources for statistical machine translation are multilingual aligned corpora (speech, text). The syntax of spoken language is different from written language. Spontaneous speech contains colloquial forms, conversational forms like e.g. hesitations, etc. and ungrammatical features such as e.g. false starts, corrections, repetitions, incomplete syntactical structures, etcetera. Three bilingual aligned corpora from transcriptions of spoken dialogues have been created: Spanish/US-English, Catalan/US-English, and Spanish/Catalan, thus providing a source of learning typical problems of spontaneous speech (Moreno et al, 2004; Arranz et al. 2004; Ueffing et al., 2004). Data have been selected from dialogues in a tourist domain. For statistical machine translation purposes at least a segment alignment is desirable, where a segment consists of as few sentences as possible. The dialogues chosen are aligned at turn level and sentence level and consist of publicly available dialogues from the Verbmobil project in US-English and selected dialogues from the TALP speech database in Spanish and Catalan covering a tourist domain.

The selected corpora are summarized in the table below:

	Verbmobil	TALP-tourism	
	English	Spanish	Catalan
Dialogues	510	156	122
Turns	4728	8701	7161
Phrases	11063	19246	14822
Tokens	110992	291757	219480
Words	110279	208749	156730

Starting from the Verbmobil and TALP-tourism project's documentation the transcription of the databases has been adapted to the specific characteristics of the bilingual aligned corpora.

The transcription of spoken dialogues is orthographic: it is a case sensitive transcription preserving also some punctuation marks and a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files (e.g. truncated or missing words, etc.). The extra marks contained in the transcription are helpful to interpret the sentence type and structure of the utterance.

The original transcriptions have been translated in the corresponding target languages: Catalan, Spanish and US-English resulting in 6 language pairs: Catalan/Spanish, Catalan/US-English, Spanish/Catalan, Spanish/US-English, US-English/Catalan and US-English/Spanish.

A translation methodology was chosen to produce a lower level of perplexity in the system while still maintaining the competitiveness of statistical machine translation techniques which do not use syntactic information explicitly. The general rules for translation from one language to all target languages are summarized below.

Keep the translations as literal as possible to the source text, while preserving their correctness (keeping the whole meaning) and naturalness (acceptability with respect to a competent target language speaker). When several choices are possible, the desired target text should preserve the source syntax and word order as much as possible. However, this does not apply to those cases where such literal translations may either change the semantics or add ambiguity. In case of ambiguities the translator should take into account the domain (tourist environment) and select the most plausible translation.

Proper names are only translated when they have a translation in the target language; otherwise they are kept and marked as foreign language. Abbreviations, numbers and digits are normally expanded.

As mentioned above marks for speech events are included in the transliterations. As a general rule, these marks are not need to be translated. However word fragments and unintelligible words etc. may affect the translation of the sentence or a set of words in the sentence and should therefore be marked appropriately in the target language.

The resulting corpora consist of a set of transcriptions of spontaneous spoken dialogues in a source language and their corresponding translation in a target language. The translations are aligned at turn and sentence level and have the following structure:

A bilingual Corpora is defined by a set of documents spoken in one or more source languages and translated into one or more target languages:

- Description
- Bilingual lexica domain or scenario
- Date
- Documents

Each document is a set of dialogues spoken in a source language and translated into a target language. It's specified by:

- Source language
- Target language
- Dialogues

A dialogue is a set of turns between two speakers. It's specified by:

- Dialogue number
- Dialogue scenario or subdomain
- Turns

A turn is a sentence or a set of sentences spoken by one speaker without other speaker's interruption, its translation and the alignment between the source and the target sentences. It's specified by

- Turn number
- Speaker
- Orthographic transcription
- Source language segments
- Target language segments
- Alignment

An XML-based mark-up language was chosen to represent the linguistic information in a formal, unambiguous manner and easy to parse. Any XML version 1.0 compliant parser able to deal with UTF-16 can be used to parse the data.

### 1.3.2 Specifications and Development of 'Phrasal Lexica'

State of the art statistical machine translation (SMT) systems use large corpora of bilingual aligned sentences for training. However large resources most often are not available for given languages and the creation needs a lot of time and manual work. Recently, experiments using bilingual phrases (i.e. very short sentences or syntactic phrases) show that short bilingual aligned segments are a useful resource for SMT systems (Vogel et al., 2003, Koehn et al., 2003). Bilingual phrases are phrases typically used in a restricted semantic domain, e.g. the tourist domain covered by this project. Short phrases like '*I need*', '*where is*', '*could you help me*' are commonly used in tourist domains. The advantages of such a phrasal based lexicon (phrasal lexicon, D5.8) over a bilingual aligned corpus consisting of whole sentences are the following:

- The word alignment can be learned much more accurately, because the number of possible alignments (and thus the number of possible misalignments) between words is reduced significantly for short segments (Venugopal et al., 2003, Imamura, 2002).
- The phrasal lexicon contains many variants of the same phrase, e.g. different conjugations of verbs, such as *I will go there, you will go there, ...* The alignment and lexicon model could learn such dependencies more accurately.
- In (S)MT, disambiguation is performed based on the context of the word. Thus, the system could learn the correct use of a word from the phrase in which it is embedded. In addition the use of POS and base form (lemma) information in the phrasal lexicon was helpful to improve the results. A more detailed overview on the results is given in (Ueffing& Ney, 2005) and in chapter 1.5 of this document.

Specifications for the creation of phrasal lexica have been developed (Moreno et al., 2004). In a first step reference phrases typical for a certain semantic domain which then are translated in the target languages have to be created.

The procedure to obtain the reference phrases can roughly be summarized as follows:

1. Create large text corpora in a reference language (US-English) in a given domain (tourist)
2. Select the most frequent content words (i.e. nouns, verbs, adjectives, etc.) in such domain to create a representative word list in a given domain (tourist). Please note that a general, domain independent, word list was already created in the project (D1.2). To get a maximized representative domain-specific word list, only content words have therefore been selected assuming that function words were covered by the large list. Otherwise it is recommended to chose all frequent words.
3. For each word in the word list, provide the context in which the words are embedded. The result is a list of either very short sentences or phrases, called the reference corpus. Manually add a set of typical phrasal expressions like e.g. those used in tourists' guides.
4. Translate the reference corpus in all target languages.
5. For each sentence, tag each token according to POS, without attributes and provide the lemma.
6. Create a phrasal lexicon containing aligned phrases /sentences, disambiguated POS and lemma using a DTD specifically developed for this purpose.

The final reference corpus consisted of 10K phrases and short sentences in US-English which have been translated in the target languages: Catalan, Finnish, German, Hebrew, Italian, Russian, Slovenian and Spanish following the above specifications.

Additionally the resulting phrasal lexica were the basis for the development of the wordlists and lexica for ASR and TTS covering only the tourist domain: for each entry in the phrasal lexicon phonetic, morphological and syntactic information as specified in D2 has been provided (D5.2, D5.4).

## 1.4 Specifications for validation of LR (WP6)

The project has been the first one in which validation of language resources has been addressed in such a complete and detailed way. Specifications for validation of lexica and aligned corpora have been developed and are publicly available (Shammas et al., 2003, 2004).

All aspects of the language resources are validated, including orthography, phonetic transcription, supra segmental aspects as well as morphological and syntactic information. Tests for checking alignments and translation quality have been developed. Two types of validation have been applied to all language resources: automatic and manual. Automatic tests are done on formal aspects that can be tested with software whereas manual checks are those that require sophisticated knowledge of the language. Validation reports document the results and are distributed with the final LR for exchange. As already mentioned three independent validation centres were involved: CST (Center for Sprooktechnologie), Denmark) SPEX (Speech Processing EXperise), The Netherlands and the University of Vigo (Spain). SPEX and CST have validated the wordlists and large lexica for ASR and TTS (D1.2, D31/2), CST the bilingual aligned corpora (D5.8) and the University of Vigo the bilingual aligned corpus for SMT (D5.7).

### 1.4.1 Specification of validation of language resources for ASR and TTS

Validation is performed both automatic and manual.

The automatic checks test aspects such as :

- Correct numbers of entries per domain (names/words).

- Only valid phonetic symbols, POS tags and attributes per POS are used according to the language-dependent specifications.
- Proper XML format. Since a generic and a language-specific DTD was written to capture all formal features of the lexica, a great number of formal criteria could be automatically tested by checking it against the DTD by an off-the shelf parser such as XMLSpy. For other checks, e.g., checking for sufficient coverage of various domains, missing POS tags etc. as well as other formal aspects of the lexica, special software was written in Perl.

Manually the following checks were performed on 1000 entries/POS tags:

- The correctness of spelling, phonetic transcriptions and supra segmental aspects.
- POS tags and their corresponding attributes.
- Documentation

## 1.4.2 Specifications of validation criteria of language resources for SMT

Validation again is performed both automatic and manual.

### 1.4.2.1 Validation of Phrasal Lexica

The translations of the US-English reference corpus in each of the target languages have been provided along with each token of the translated entry tagged for POS and lemma (without attributes) in XML format with the compliant DTD. The correctness of the format was checked automatically.

The following manual checks on 600 items chosen from the reference corpus to ensure that all languages were checked on the same set were performed:

- Is it a possible translation of the English reference phrase, one that a native speaker would utter?
- Is the syntax possible?
- Is it semantically correct and does it convey the same general meaning as the English reference?

Furthermore the orthography of each translated phrase has been checked.

The validation checks of the phrasal lexica also included manual checking of POS and lemma tagging.

For all manual validation checks validators had to provide corrections for any reported translation, POS and lemma mistakes and were instructed to report any suspicion of systematic errors.

### 1.4.2.2 Validation of Bilingual Aligned Corpora

The aligned corpora in English (EN), Spanish (ES) and Catalan (CA) and two translation pairs: EN2ES, EN2CA; CA2ES, CA2EN; ES2CA, ES2EN were checked automatically for correctness of DTD etc.

For each language pair, three dialogues were chosen for manual validation: 1 short dialogue, 1 medium and 1 long dialogue amounting to a total of 18 dialogues to check. The validation set was composed of 3000 word-pairs chosen randomly from the 18 dialogues but with some constraints. No isolated words or isolated sentences or turns are chosen. Words must be chosen as consecutive parts of the dialogue. The same manual checks as those applied to the phrasal lexica have been performed.

### 1.4.3 Validation results (WP3/5/6)

All language resources created in Workpackage 3 and 5 have been validated by end of the project except for one database which is currently undergoing validation and one missing database which will be submitted within the next month.

Overview of all language resources and responsible partners:

Partners	Wordlists in 12(+1), (D1.1), 8 (+1) languages (D5.1, D5.2)	large lexica for ASR and TTS (D3.1/2), lexica for tourist domain (D5.4)	D5.7	D5.8
SIE	Russian / Turkish	Russian / Turkish		Russian
IBM	Greek / Italian	Greek / Italian		Italian
UPC	Catalan / Spanish & Catalan / Spanish / US-English for D5.1/2	Catalan /Spanish & Catalan / Spanish / US-English for D5.4	Catalan / Spanish / US-English	Catalan, Spanish, US-English
RWTH	German / Standard Arabic	German / Standard Arabic		German
NSC	Hebrew / US-English	Hebrew / US-English		Hebrew
NOK	Finnish / Mandarin Chinese	Finnish / Mandarin Chinese		Finnish
UMB	Slovenian	Slovenian		Slovenian

The quality of the produced LR was very high. In all databases the error limits which were kept very low were far below the limits (Shammas, 2003; 2004). Detailed validation reports have been produced and will be delivered with the final databases. All partners agreed that the validation process is a necessary means to ensure high quality data. However there were also some difficulties encountered which were mainly due to the fact that the project was the first one to validate large written language resources in such a variety and to such a great extend (Shammas et al, 2005).

## 1.5 Results in SMT Research (WP4/7)

A statistically based translation baseline system has been developed (Arranz et al., 2002) which was the basis for further research. The language pairs to translate within were: English -> Catalan, English -> Spanish, Catalan -> English, Catalan -> Spanish, Spanish -> English, Spanish -> Catalan.

One of the research aims was to study the effects of different structured language resources on speech-to-speech translation systems (Ueffing&Ney, 2005). The following language resources developed in the project were used:

- Bilingual aligned corpus in English, Spanish and Catalan in a tourist domain.
- Recordings of spontaneous dialogues and their translations (500k running words) annotated with POS-tags and base forms (the latter only for Spanish and Catalan).
- Phrasal lexica (10k) containing short phrases and sentences automatically annotated with POS-tags and base forms (for Spanish and Catalan).

- Verb forms consisting of a list of full forms for each base form extracted with the UPC tagger for Spanish and Catalan<sup>3</sup>.

The baseline statistical machine translation<sup>4</sup> system makes use of a so-called Alignment Template System. These are pairs of source and target language phrases with an alignment within the phrases. The alignment templates are build at the level of word classes. This improves the generalization capability of the model. The system takes a number of different sub-models into account which are combined log-linearly. Those sub-models are:

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

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

The baseline results for all six translation directions on the development and test corpora using the alignment template system is described in (Ueffing & Ney, 2003). The baseline system has been improved in the duration of the project, such that the translation quality of this system lies beyond that of the system described in the above mentioned title.

Results of the baseline system can be roughly summarized as follows:

- the translation from Spanish into Catalan and vice versa is by far the easiest one: The WER lies around 11-12%. Considering the fact that there are only single references available for the develop and test corpora, it is assumed that the actual error rate lies even below this value.
- the translation from Spanish into English and from Catalan into English are approximately of the same complexity and much more difficult due to differences in morphological and syntactic structures. This fact is reflected in the error rates which are very similar (around 40% WER).
- the translation from English into Spanish and Catalan ahs been the hardest task, the baseline system achieved a translation quality of 41% to 43% in WER.

In the following sections, methods making use of additional language resources that improve the performance of the translation system are described. The following refinements were done:

- use the phrasal lexicon as additional resource,
- use verb expansions for Spanish and Catalan,
- use sentence class based language models (LMs).
- Word alignment consists of lexicon smoothing with base forms and the use of hierarchical models using base forms and POS.

The experimental setup consisted of:

<sup>3</sup> these verbal forms have not been developed within the scope of the project but were delivered as an additional resource

<sup>4</sup> a more detailed overview of the current system can be found in Bender & Zens, 2004

- Phrasal lexica as additional language resource in training
- Investigation of effects for the whole training corpus (40k), the reduced training corpus (1k) or no training corpus (0k). In addition
- Verb expansions for Spanish and Catalan verbs using verb full form list produced by the UPC tagger are used
- adding additional training data (2k new entries) to the language model.

#### Results:

Translation error rates [%] on the development and test set for Spanish -> English. Results are presented for different training corpora; 40k and 1k sentences and phrasal lexica only. Effect of adding the phrasal lexica to the training material, verb expansions and a full language model.

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

Summary: As can be seen the best results are obtained by using the full training corpus with the additional phrasal lexicon. However, for the reduced bilingual corpus, the importance of the phrasal lexicon is significant. The degradation in terms of WER and PER by using only 2.5% of the original corpus is not higher than 25% relative if the additional phrasal lexicon and language resources containing morphological information are available. This can be further improved by up to 1.9% absolute in WER if monolingual in-domain data is available and a better language model can be trained. The advantage of using such a small corpus is that its acquisition should not require any particular effort since in the matter of fact 10k parallel sentences in two or three languages can be also produced manually keeping efforts in limits. Using only the phrasal lexicon and additional resources, the obtained error rates are similar to those for the small training corpus.

In addition sentence dependent language models have been used. Within this approach, a standard n-gram language model is interpolated with a special sentence dependent language model. Each sentence is filtered according to various regular expressions. In this setup, the triggers are punctuation marks such as '?', '!' and '.', which separate the sentences into three classes: questions, exclamations and declarative sentences. Another possibility is to add language dependent triggers that usually introduce subordinate clauses, e.g. *because* in English and *porque* in Spanish, or to base the triggers on POS-tag information of the verbs.

Several different regular expression triggers based on punctuation marks and POS verb tags for additional language models which are interpolated with a base language model using five-grams and modified Kneser-Ney discounting. The interpolation lambdas are optimized on the development set for each class. The perplexities can be reduced around 10% relative. Rescoring experiments were performed on N-best lists using the sentence class language models and IBM-1 translation model.

For all four translation directions the rescoring experiments on N-best lists slightly improved the translation performance.

Another experiment using different structured language resource consisted in lexicon smoothing with base forms. The standard lexicon model is based on full form words only. For inflected languages such as Spanish and Catalan this might cause problems, because many full form words occur only a few times in the training corpus. Compared to English, the token/type ratio for Spanish and Catalan is usually much lower (e.g. English 99.4, Catalan 66.6, Spanish 63.2). The information that multiple full form words share the same base form is not used in the lexicon model. To take this information into account, the lexicon model has been smoothed with a backing-off lexicon that is based on word base forms.

The effect of this smoothing method on the quality of the word alignment was evaluated on a the Verbmobil German-English translation task, because there were manually annotated data available. It results in an improvement of the alignment error rate for both translation directions.

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

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

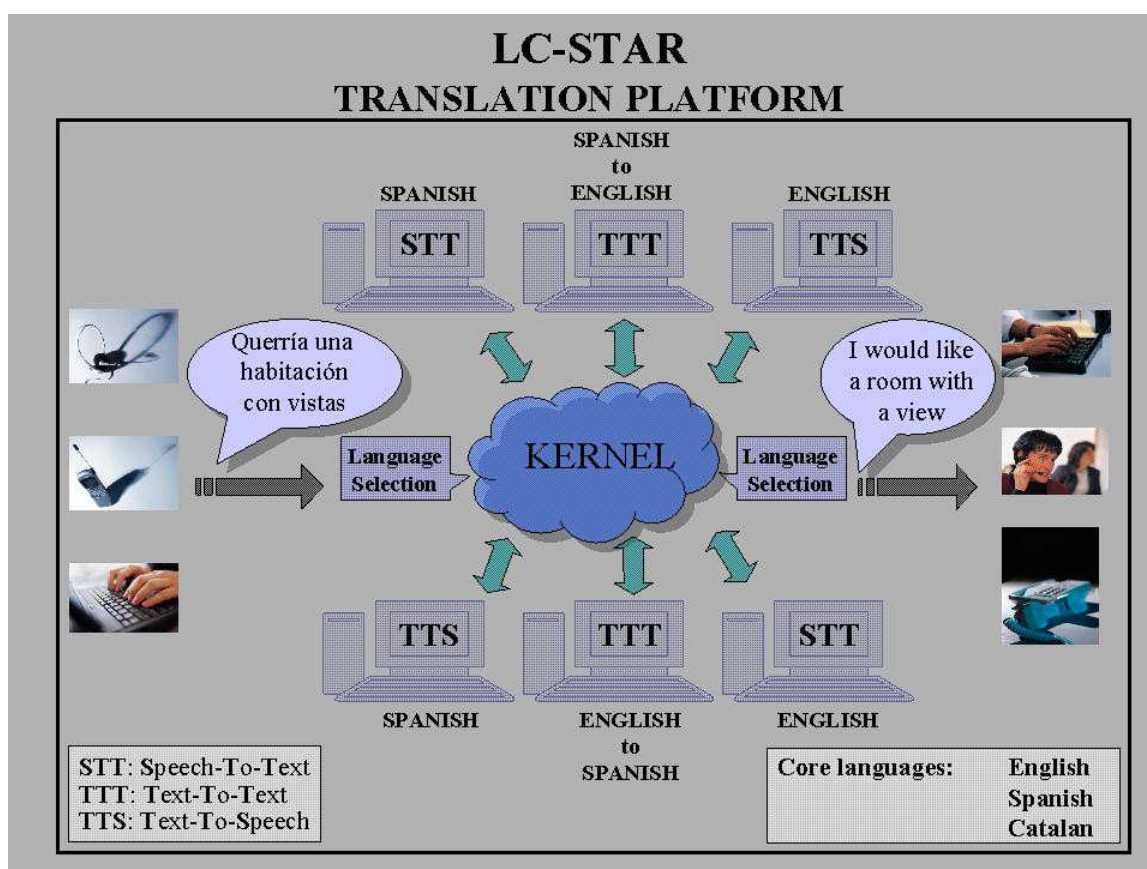
Several other methods have been applied to come to a tighter coupling between ASR and SMT. Also the use of WordNet has been studied which did not improve the results.

Further details are presented in (Ueffing& Ney, 2005)

## 1.6 Overview of Demonstrator (WP4/7)

The demonstrator platform, Gaia, is a telephone server which can offer translation in the three research languages Catalan, Spanish and US-English in the tourist domain (Bonafonte, 2004). The platform structure follows a modular design, with different parts

running in separate machines. The speech processing tasks (recognition, translation and synthesis) are implemented in independent servers which communicate with the platform kernel via standard network sockets. Distribution is mandatory since each task can be computationally demanding. Moreover, it allowed the different partners to implement their particular solutions in different platforms if desired, achieving a high grade of flexibility. A client/server approach has been followed: each part is conceived as an entity in itself, with the ability or requirement to run independently of the platform. Hence, the different parts of the platform are implemented as independent programs, using standard sockets and a predefined network protocol to communicate with the different modules of the system. This is the approach followed in the Galaxy project at MIT or in the concluded national Spanish project BASURDE among others. Our interest was to integrate existing solutions into a common framework. Figure 1 shows the different modules of the platform and its distribution across a network environment.



Overview Gaia

### 1.6.1 Description of the Kernel

The platform logic is implemented in the kernel. It handles the communication between the platform clients, has knowledge of the different servers (ASR, SMT and TTS), and establishes appropriate connections based on the languages and configurations chosen by the user.

The kernel of the platform communicates with different types of servers:

Terminal servers: they collect the input of the user and provide the output. Three dual terminal servers have been developed: i) telephone terminal, to interact using the

telephone through Dialogic cards; ii) speech console terminal, to interact using speech through and IP connection and iii) text console terminal which is mainly used to test the translation engine. In LC-STAR, the demonstration is based on the telephone terminal: two users can communicate using different languages, through a translation service provided by *Gaia* through the telephone.

Speech Technology servers. *Gaia* is prepared to compare different technologies. Therefore, more than one server for a given technology can exist. During the project, the following technology servers have being integrated:

- Speech recognition: provided by UPC.
- Speech synthesis: provided by UPC and from the Festival project.
- Spoken translation: with the technology provided by RWTH and from UPC. The technology provided by RWTH includes the advances achieved during the project.

Visualization server: allows monitoring the result of the technology in all the components for each dialog.

Configuration and debugging servers.

Acoustic models for ASR for Spanish and Catalan have been trained using either the TALP-tourism corpus or a combination with SpeechDat databases. Both, the acoustic databases and the software to train the acoustic models (RAMSES) are provided by UPC.

For English the MACROPHONE corpus has been used. However, this corpus was not adapted to the task. Furthermore, for this task there was no corpus in English available for development and testing. For this reason, it was decided to avoid the use of English at the input and test the portability of the translation in 4 directions: Spanish-English; Catalan-English; Spanish-Catalan; Catalan-Spanish.

The language models for speech recognition are trained from the TALP-tourism corpus, using both the source sentences and the translated sentences. First, several classes are defined (hotels, names, cities of the world, etc.). Then, class n-grams are inferred using variable-length n-grams (x-gram), linear discounting and back-off smoothing. The toolkit to estimate the n-grams is part of RAMSES, the UPC Continuous Speech Recognition System. Several trials were done to add the Verbmobil corpus or tourist web pages to the training material but there was no significant decrease on the perplexity. The LC-STAR lexicons for Spanish and Catalan (UPC) and English (NSC) have been used for the speech recognition and speech synthesis engines.

For a more detailed description of *Gaia*, refer to [Pérez & Bonafonte, 2003].

### 1.6.2 Acceptance testing of the demonstrator

The final acceptance testing of the demonstrator using included (Bonafonte, 2005):

- qualitative end-to-end evaluation at the utterance level,
- task oriented evaluation which aims to evaluate if users can achieve a given task with the state of the technology. Additionally, the subjects are asked to fill a qualitative questionnaire about some aspects of the system.

Detailed results are provided in Bonafonte, 2005. The qualitative evaluation of the acceptance of the platform revealed that although users did not always give a high score

to the system they still prefer to use the speech modality for both input and output in most of the proposed situations.

The language transfer within the different languages has been successfully demonstrated at the final review meeting.

## 2 Dissemination and Use (WP7)

In a first step a web page has been created which will be kept up to date until June 2006. In the beginning the project has been promoted in the LangTech2002 conference and FP6 Launch Conference.

Recently an article has been published on the IST Results webpage

<http://istresults.cordis.lu/index.cfm/section/news/tpl/article/BrowsingType/Features/ID/74181>

to promote the findings to a broad audience.

In short term all developed language resources will be used to optimize existing ASR, TTS, text-to-phoneme and/or SMT systems and on long-term to quickly adapt to new applications and customers' demands.

The created language resources are distributed via ELRA no later than 18 months after the official end of the project. The official end of the project is 31<sup>st</sup> of January 2005, which sets the deadline for the LR distribution for 30<sup>th</sup> of June 2006.

The software of demonstrator platform and the DTD for the large lexica for ASR and TTS are publicly available for the community as well as all specifications for creation of language resources and validation criteria.

A cooperation with ELRA on a new project called 'Unified Lexicon Approach' has been started. The idea is to study if it is possible to join different levels of information from various lexica (pronunciation, morphology, syntax, semantic) already available (eg. Parole / Simple lexica) with the purpose to create ever larger lexical databases. The LC-STAR lexica are envisaged to be a basic resource kit for these databases.

First steps to establish a new non-funded consortium (LC-STAR II) to collect even more language resources have been undertaken.

Results will also be used in the EU-funded project TC-STAR and ECESS a non-funded project with partners all over the world concentrating on speech synthesis.

Scientific results have been presented in the following conferences:

Author(s):	Title:
N. Ueffing, H. Ney	Training Corpus Size and Statistical Machine Translation Quality Informatiktage 2002 of the Gesellschaft für Informatik e.V., Bad Schussenried, Germany, November 2002.
N. Ueffing, H. Ney	Using POS Information for Statistical Machine Translation into Morphologically Rich Languages EACL2003, Budapest, Hungary, April 2003.
E. Hartikainen, G. Maltese, A. Moreno, S. Shammass, U. Ziegenhain	Large lexica for Speech-to-Speech Translation: From Specification to Creation Eurospeech 2003, Geneva, Switzerland, September

	2003.
D. Conejero, J. Giménez, V. Arranz, A. Bonafonte, N. Pascual, N. Castell, A. Moreno	Lexica and Corpora for Speech-to-Speech Translation: A Trilingual Approach Eurospeech 2003, Geneva, Switzerland, September 2003.
A. Moreno	LC-STAR Project Presentation SEPLN2003, Madrid, Spain, September 2003.
G. Leusch, N. Ueffing, H. Ney	A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation MT Summit IX, New Orleans, LA, September 2003. Proceedings p. 240-247.
N. Ueffing, K. Macherey, H. Ney	Confidence Measures for Statistical Machine Translation MT Summit IX, New Orleans, LA, September 2003. Proceedings p. 394-401.
V. Arranz, N. Castell, J. Giménez	Development of Language Resources for Speech-to-Speech Translation RANLP2003, Borovets, Bulgaria, September 2003.
H. Fersøe, E. Hartikainen, H. van den Heuvel, G. Maltese, A. Moreno, S. Shammass, U. Ziegenhain	Creation and Validation of Large Lexica for Speech-to-Speech Translation Purposes LREC2004, Lisbon, Portugal, May 2004.
M. Popovic, H. Ney	Towards the Use of Word Stems and Suffixes for Statistical Machine Translation LREC2004, Lisbon, Portugal, May 2004.
D. Verdonik, M. Rojc, Z. Kacic	Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components LREC2004, Lisbon, Portugal, May 2004.
V. Arranz, N. Castell, J.M. Crego, J. Giménez, A. de Gispert, P. Lambert	Bilingual Connections for Trilingual Corpora: An XML Approach LREC2004, Lisbon, Portugal, May 2004.
U. Ziegenhain, A. Moreno, N. Castell	Creation of lexica for statistical based speech-to-speech translation AST 2004, Maribor, Slovenia, July 2004.
D. Verdonik	Slovenian lexica and corpora within the LC-STAR project

	AST 2004, Maribor, Slovenia, July 2004.
R. Zens, E. Matusov, H. Ney	Improved Word Alignment Using a Symmetric Lexicon Model Coling2004, Geneva, Switzerland, August 2004.
M. Popovic, H. Ney	Improving Word Alignment Quality Using Morpho-Syntactic Information Coling2004, Geneva, Switzerland, August 2004.
X. Perez, A. Bonafonte	GAIA: A Software Platform for the Integration of Speech Translation Technologies In: Proceedings of ICSLP 2004.

### 3 References

- Koehn, P., Knight, K. "Feature-Rich Statistical Translation of Noun Phrases". 41st Annual Meeting of the ACL, Sapporo, Japan, July-2003.
- Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., Waibel, A. "The CMU Statistical Machine Translation System". MT-Summit, New Orleans, USA, September-2003.
- Venugopal, A., Vogel, S., Waibel, A. "Effective Phrase Translation Extraction from Alignment Models" 41st Annual Meeting of the ACL, Sapporo, Japan, July-2003.
- Imamura, K. "Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT". 9th Int. Conf. on Theoretical and Methodological Issues in MT. Keihanna, Japan, March-2002.
- S. Nießen, H. Ney. "Statistical Machine Translation with Scarce Resources using Morpho-Syntactic Information", In: Computational Linguistics 2004.
- O. Bender, R. Zens, E. Matusov, H. Ney: Alignment Templates: the RWTH SMT System. In Proc. of the Int. Workshop on Spoken Language Translation (IWSLT), Kyoto, Japan, September 2004.
- X. Perez, A. Bonafonte : GAIA: A Software Platform for the Integration of Speech Translation Technologies. In: Proceedings of ICSLP 2004.
- Deliverables:
- U. Ziegenhain et al. (2002): Specification of corpora and word lists in 12 languages. LC-STAR Project IST-2001-32216 Deliverable D1.1.
- G. Maltese et al. (2003): General and language specific specification of contents of lexica. LC-STAR Project IST-2001-32216 Deliverables D2.1, D2.2, D2.3, D2.4 in one document, 2003.
- V. Arranz, N. Castell, J. Jiménez, H. Ney, N. Ueffing (2003): Description of language resources used for experiment. LC-STAR Project IST-2001-32216 Deliverable D4.2
- V. Arranz, N. Castell, J. Jiménez, A. Moreno (2004): Description of raw corpora. LC-STAR project IST-2001- 32216 Deliverable D5.3
- N. Ueffing & H. Ney (2005): Results on different structured language resources. LC-STAR Project IST-2001-32216 Deliverable D4.5
- A. Moreno et al. (2004): Language independent specification of language resources for translation. LC-STAR Project IST-2001-32216 Deliverable D5.5
- N. Castell et al.(2004): Language dependent specifications of language resources for translation. LC-STAR Project IST-2001-32216 Deliverable D5.6

S. Shammass et al. (2003): Specifications of validation criteria for lexica for speech recognition and synthesis. LC-STAR Project IST-2001-32216 Deliverable D6.1

S. Shammass et al. (2004): Specifications of validation criteria for lexica for language resources for speech centered translation. LC-STAR Project IST-2001-32216 Deliverable D6.3

S. Shammass et al. (2005): Validation of for lexica for speech recognition and synthesis. LC-STAR Project IST-2001-32216 Deliverable D6.2

S. Shammass et al. (2005): Validation of language resources for speech centred translation. LC-STAR Project IST-2001-32216 Deliverable D6.4

A. Bonafonte et al.(2005): Acceptance testing of the demonstrator. LC-STAR Project IST-2001-32216 Deliverable D7.2