

# Large Lexica for Speech-to-Speech Translation: From Specification to Creation

*Elviira Hartikainen<sup>1</sup>, Giulio Maltese<sup>2</sup>, Asuncion Moreno<sup>3</sup>, Shaunie Shammass<sup>4</sup>, Ute Ziegenhain<sup>5</sup>*

<sup>1</sup>Nokia Research Center, Finland (elviira.hartikainen@nokia.com)

<sup>2</sup>IBM Italy, Rome, Italy (giulio.maltese@it.ibm.com)

<sup>3</sup>Universitat Politècnica de Catalunya, Barcelona, Spain (asuncion@gps.tsc.upc.es)

<sup>4</sup>Natural Speech Communication (NSC), Israel (shaunie@nsc.co.il)

<sup>5</sup>Siemens AG, Munich, Germany (ute.ziegenhain@mchp.siemens.de)

## Abstract

This paper presents the corpora collection and lexica creation for the purposes of Automatic Speech Recognition (ASR) and Text-to-speech (TTS) that are needed in speech-to-speech translation (SST). These lexica will be specified, built and validated within the scope of the EU-project LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) during the years 2002-2005. Large lexica consisting of phonetic, prosodic and morpho-syntactic content will be provided with well-documented specifications for at least 12 languages [1]. This paper provides a short overview of the speech-to-speech translation lexica in general as well as a summary of the LC-STAR project itself. More detailed information about the specification for the corpora collection and word extraction as well as the specification and format of the lexica are presented in later chapters.

## 1. Introduction

Current approaches to the development of speech recognition, text to speech synthesis and speech centered translation necessitate the development of a wide range of Language Resources (LR).

Great advances have been made in the development of annotated speech databases needed for building speech recognition systems for many languages and for many applications in specific acoustical environments. Less attention has been given to the linguistic oriented language resources needed for the language transfer of SST components that include:

- Flexible vocabulary speech recognition
- High quality text-to-speech synthesis
- Speech centered translation

Such linguistic oriented resources include suitable text corpora for producing lexica that are enriched with phonetic, prosodic and morpho-syntactic information. These generic LRs are needed to build SST components covering a wide range of application domains in different languages. Currently, such large-scale LR are not publicly or commercially available and industrial standards are lacking.

Even the most important organizations that provide Language Resources in the field of Language Engineering have defects in their catalogues. Especially the data for lexical LR show a lack of coverage for many languages and applications. Specifically the following drawbacks are evident:

- Lack of coverage with respect to a wide range of application domains
- Lack of suitability either for speech synthesis or speech recognition
- Lack of quality control
- Lack of standards
- Lack of coverage in languages
- Mostly limited to research purposes

The main objective of the LC-STAR project is to make large lexica available for many languages that cover a wide range of domains along with the development of standards relating to content and quality. Pioneering work is being done for defining standards with respect to content and format issues.

For speech-centered translation, the project will focus on statistical approaches allowing an efficient transfer to other languages using suitable LR. The LR needed for this purpose are aligned bilingual text corpora and monolingual lexica with morpho-syntactic information.

This paper is organized as follows. Section 2 describes the LC-STAR project, including an overview of the partners and what languages are covered. Section 3 describes the specifications for corpora collection and word list extraction and Section 4 describes the specifications and formats for the lexica. Finally, the paper concludes with the current status of the project and quality assessment issues, which are discussed in Section 5.

## 2. Overview of the LC-STAR Project

The LC-STAR consortium consists of 4 industrial companies, 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). SPEX (Speech Processing EXpertise) and CST (Center for Sprogteknologi) are responsible for validating the lexica. Project partners have wide experience either in previous speech database projects (e.g. Speech-Dat family, Speecon etc.) or projects relating to machine translation (e.g. Verbmobil).

Currently, the project covers 12 languages from various parts of the world, whereby each partner is responsible for creating lexica for two languages. It is possible that more languages will eventually be covered (more lexica created) since the project is still open for new external partners. The list of all languages covered and responsible partners is presented in Table 1 below.

IBM	Italian
	Greek
Nokia	Finnish
	Mandarin
NSC	Hebrew
	US-English
RWTH Aachen	German
	Classical Arabic
Siemens	Turkish
	Russian
UPC	Spanish
	Catalan

Table 1: List of languages and responsible partners.

The languages that are covered include main languages of the world as well as ones that are less common. In addition, they represent a wide spectrum of language types, including languages constructed with minimal morphological information (e.g. Mandarin) as well as highly inflective languages (e.g. Hebrew). The range of language types raises interesting specification and standardization issues, which are addressed within the project.

### 3. Specifications for corpora collection and word list extraction

Each language-specific lexicon consists of three parts: 1) at least 50,000 inflected common word entries covering six major domains, 2) 45,000 proper names covering three major domains and 3) at least 5,000 entries for special voice-driven applications.

One major problem in the project was how to build the lexica for so many different languages that would adequately cover the most frequent words from the collected corpora and yet also reflect the linguistic aspects inherent in the language.

In the following section, the corpora collection and word extraction processes are described in more detail. Examples of special requirements or limitations that are based on language-dependent considerations are provided. In addition, examples of statistical coverage for some languages are shown [2].

#### 3.1. Domains for common words

For common words, corpora were collected in six major domains: sports and games, finance, news, culture, consumer information and personal communications. These domains were further divided into subdomains, (e.g. ‘local and international affairs’ and ‘editorials and opinions’ in the ‘News’ domain.) Each corpus was required to be at least 10 million tokens in total. A minimum of 1 million tokens was required in each domain with no upper limit, with the exception of the personal communications domain, where the minimum amount of data was limited to 500,000 tokens. This

exception was based on prior experiments that showed that this kind of data may contain considerable material needing time-consuming and costly manual cleaning.

Sources for the data collection included electronically available text corpora (if legally obtainable), e.g. newspapers, periodicals, books, manuals and Internet data (online magazines, newsletters, discussion groups, newsgroups etc.). Material from chat-rooms and other unedited sources were not allowed. The cut-off date for all corpora used was 1990. Furthermore, at least 50% of the newspapers and periodicals could not be older than five years. This approach was chosen to ensure that the word lists extracted from the corpora represent current usage from a wide range of domains and provide good lexical coverage (for further details see [3], [4]).

#### 3.1.1. Word list extraction for common words

To optimize the coverage criterion, the word lists were required 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 the most frequent 50,000 entries without singletons, abbreviations and proper names. The formal procedure included 1) language-dependent cleaning and tokenizing of the corpora, 2) removing proper names, typos (typing errors), and abbreviations, and 3) checking final coverage. The final wordlists were free of digits and punctuation marks. Capitalization was used, when available, to remove proper names; this was language-dependent, however, since not all languages have capitalization (cf. Hebrew, German, Turkish).

For counting the rank coverage for each domain, the following formula was used:

$$c(w_j) = \sum_{i=1}^j n(w_i) / N \quad (1)$$

where  $n(w)$  is the observed count of word  $w$  in the corpus and  $N$  is the total number of tokens (words).

For counting the self-coverage per domain, the coverage target  $t = 95\%$  was chosen and the word list was truncated at the point where  $c * 100$  exceeded the target  $t$ .

Remaining proper names and abbreviations were removed manually. Self-coverage was recomputed again and checked to see if it exceeded 95%. If this coverage was not reached, the target  $t$  was increased and the remaining proper names and abbreviations were removed once more.

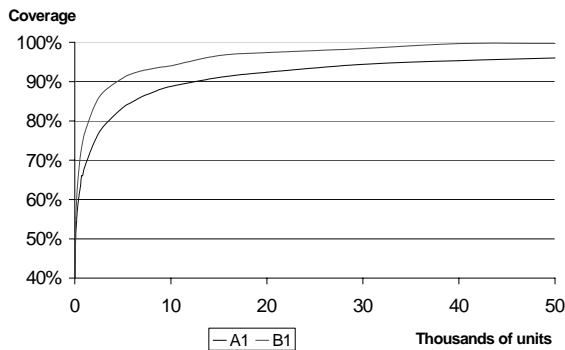
Finally, the word lists for all domains were merged. In case the merged word list was less than 50,000 entries, the coverage target  $t$  was increased and the whole coverage procedure was repeated. This iteration continued until the final word list exceeded 50,000 entries.

Table 2 shows examples of total amounts of words in collected corpora for Russian, Spanish and Mandarin.

Language	Total amount of words	Amount of different words	Coverage
Russian	11,086,030	187,049	96.33%
Spanish	36,648,565	254,439	98.48%
Mandarin	30,874,101	49,870	96.68%

Table 2: Examples of the amounts of words in corpora for Russian, Spanish and Mandarin.

Figure 1 shows the results of the corpus coverage on an independent test corpus of 2,000,000 words in Spanish. The independent test corpus was collected from newspapers, Internet texts and books. The coverage with 50,000 words is 98%.



**Figure 1:** Coverage in % on a Spanish text corpus in terms of the number of units chosen in the reference 41 million words corpora. a) Dark line shows the coverage in terms of words. b) Light line shows the coverage in terms of lemmas.

### 3.2. Domains for proper names and special applications

For proper names, three major domains were chosen: personal names (first and last names), place names and organizations. Place and organizational names were further divided (e.g., non-profit and for-profit organizations, local and international companies, brand names for “organizational” names). If possible, electronically available sources were used; otherwise, keyboarding and scanning were alternative methods used in the collection process.

At least 45,000 proper names were collected. The entries had to cover each of the three domains within the following range: a minimum of 10% and a maximum of 50%. The actual distribution was free; however, all the three domains together had to add up to 100%. This approach was chosen since the possible distribution of entries in domains varies between different languages and/or countries (e.g. few city names in Finland and Israel, few last names in Mandarin, few first names in Russian).

The special application word list consists of numbers, letters, abbreviations and seven major semantic domains related to voice-driven applications. For voice-driven application words, a reference word list of 5,700 entries in US-English was collected and translated into the other languages. This entailed providing example sentences for translation, since there were language-specific considerations involved in the translation process (e.g. case in Russian, morphological considerations in Turkish, etc.)

## 4. Specification and format of the lexica

In order to fulfill the needs of ASR/TTS components of Speech-to-Speech translation applications, the lexica has to contain detailed grammatical, morphological, and phonetic information for each language. The grammatical information needed per language within the scope of the lexica was specified at the beginning of the project [5]. The available information in all languages was merged into a unique list of

POS tags (part-of-speech tags). Most of the POS have an internal structure with attributes common to several languages (e.g. number or gender), but some POS relate only to a subset of languages (e.g. case), or are relevant only for specific individual languages (e.g. polarity for Turkish verbs). The advantage of such an approach is that a single description of grammatical features can cover the set of twelve languages.

In addition to grammatical features, morphological (including lemma) and phonetic information is specified for each word in a given lexicon. In agglutinative languages, such as Turkish, some morphological boundary information is also provided. Phonetic transcriptions use SAMPA symbols, which are specified for each lexica; foreign words are phonetized according to the SAMPA set. Syllable boundaries and primary stress are also provided in the phonetic transcriptions. For each word, it is possible to specify more than one POS, and/or more than one lemma, and/or more than one phonetic transcription.

### 4.1. Formal representation of the lexica

The formal representation of lexica is implemented via an XML-based mark-up language that meets the requirements of representing the linguistic information in a formal, unambiguous manner. Such a representation is both easy to read and easily processed by generic applications. A formally specified grammar (Document Type Definition or DTD) containing all the described linguistic information allows for automatic validation of the XML-based lexica.

#### 4.1.1. Entries and entry groups

Basically, a lexicon consists of a set of entry group elements. An entry group refers to a generic entry in a vocabulary, e.g. masculine singular for an adjective, singular for a noun and infinitive for a verb. The spelling of an entry is the key to the entry group. For each entry group, it is therefore mandatory to specify an orthography (with allowed alternative spellings) plus one or more entry or compound entry elements.

An entry refers to one specific grammatical/morphological meaning of a vocabulary entry, like *can* (verb) and *can* (noun). For each entry, one or more POS, one or more lemmas and one or more phonetic transcriptions must be specified. Special tags are used to mark the words that are included in the special application wordlist. For abbreviations, multiple expansions can be specified.

In some languages (e.g. Catalan, Hebrew, Italian, Spanish, Turkish) there are assimilation or agglutination phenomena that can be dealt with by using compound entries. Besides its spelling and its phonetic transcription, a compound entry is composed of a series of two or more entry elements (a subset of an entry) which are simply links to other entries. Each entry element must be characterized by an orthography and must contain one POS tag, together with all of its attributes (it is assumed here that orthography and fully specified POS tag - i.e. POS + all of its attributes - unambiguously identifies an entry).

### 4.2. Specifications for the lexica

In a given lexicon, only the information relevant to the language under examination has to be specified; each attribute has the default value *NS* (=Not Specified), which is always implied, thereby avoiding the need to specify non-existing or

non-relevant features in a given language (e.g. case in Italian or gender in Finnish).

Figure 2 shows an example of entry groups in Spanish and Turkish in XML-based coding.

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE LEXICA SYSTEM "NewLexica6.dtd" >
<LEXICA xml:lang="es">
  <ENTRYGROUP orthography="blanco">
    <ENTRY>
      <NOM gender="masculine"
          number="singular" />
      <LEMMA>blanco</LEMMA>
      <PHONETIC>" b l a N - k o</PHONETIC>
    </ENTRY>
    <ENTRY>
      <ADJ gender="masculine"
          number="singular"
          degree="positive" />
      <LEMMA>blanco</LEMMA>
      <PHONETIC>" b l a N - k o</PHONETIC>
    </ENTRY>
  </ENTRYGROUP>
  <ENTRYGROUP orthography="acıkmı•t1"
              xml:lang="tr">
    <ENTRY>
      <VER polarity="positive"
          tense="narrative_past"
          number="singular"
          person="3" />
      <LEMMA>acıkmı</LEMMA>
      <PHONETIC>a - dZ lk - "m1 S -
t1</PHONETIC>
    </ENTRY>
    <ENTRY>
      <VER polarity="positive"
          tense="past"
          number="singular"
          person="3" />
      <LEMMA>acıkmı</LEMMA>
      <PHONETIC>a - dZ lk-"m1 S -
t1</PHONETIC>
    </ENTRY>
  </ENTRYGROUP>
```

Figure 2: Example of XML-based coding of lexica in Spanish and Turkish.

## 5. Conclusions

The specifications and corpora collection for all languages have been completed. The main difficulty encountered was to find a common approach for word list extraction and linguistic description suitable for the wide variety of languages covered in the project. The partners are now building the final lexica and the prevalidation process will soon begin. New partners are free to join the project to build lexica in complimentary languages within the existing time limit.

The first phase of the project has as its goal to build large lexica suitable for ASR and TTS purposes. Later phases of the project have a goal to create special speech-to-speech translation lexica. At RWTH, speech-to-speech translation experiments using different methods are being carried out. The main purpose is to find out if machine translation can be improved if more linguistic features are added to translation lexica. Based on the results of these experiments, lexica for translating into seven languages (Catalan, Finnish, German, Hebrew, Italian, Russian and Spanish) will be specified and created. The reference word lists in US-English for these 'translation lexica' will be created from aligned corpora covering a tourist domain. A demonstrator showing the language transfer within 3 language pairs (Catalan, Spanish, US-English) will be built.

## 6. References

- [1] Project homepage: <http://www.lc-star.com>
- [2] Ziegenhain, U. et al. Specification of corpora and word lists in 12 languages. Public Project Deliverable D1.1. 2003.
- [3] Adda-Decker, M. & Lamel, L. The use of lexica in automatic speech recognition. in: van Eynde, F. & Gibbon, D. Lexicon development for speech and language processing. p.235-266. 2000.
- [4] Hu, R. Zong, C. Iso-Sipilä, J. & Xu, B. Investigation and Analysis on Designing Chinese Balance Corpus, in: Proceedings of ISCSLP conference in Taiwan. p.335-338. 2002.
- [5] Maltese, G. Montecchio, C. et al. General and language-specific specification of contents of lexica. Public Project Deliverables D2. 2003.