Creation and Validation of Large Lexica for Speech-to-Speech Translation Purposes

Hanne Fersøe[1], Elviira Hartikainen[2], Henk van den Heuvel[3], Giulio Maltese[4], Asuncion Moreno[5], Shaunie Shammass[6], Ute Ziegenhain[7]

[1]Center for Sprogteknologi (CST)
Njalsgade 80, Copenhagen, Denmark
hanne@cst.dk
[2]Nokia Mobile Phones
Itämerenkatu 11-13, 00180 Helsinki, Finland
elviira.hartikainen@nokia.com
[3]Speech Processing Expertise Center (SPEX)
Erasmusplein 1, 6525 HT Nijmegen, Netherlands
H.vandenHeuvel@let.kun.nl
[4]IBM Italy
Rome, Italy
giulio.maltese@it.ibm.com
[5]Universitat Politecnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain
asuncion@gps.tsc.upc.es
[6]NSC – Natural Speech Communication
33 Lazarov St., 75150 Rishon Le Zion, Israel
shaunie@nscspeech.com
[7]Siemens AG
Otto-Hahn-Ring 6, 81739 Munich, Germany
ute.ziegenhain@mchp.siemens.de

Abstract

This paper presents specifications and requirements for creation and validation of large lexica that are needed in automatic Speech Recognition (ASR), Text-to-Speech (TTS) and statistical Speech-to-Speech Translation (SST) systems. The prepared language resources are created and validated within the scope of the EU-project LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) during years 2002-2005. Large lexica consisting of phonetic, suprasegmental and morpho-syntactic content will be provided with well-documented specifications for 13 languages. A short summary of the LC-STAR project itself is presented. Overview about the specification for the corpora collection and word extraction as well as the specification and format of the lexica are presented. Particular attention is paid to the validation of the produced lexica and the lessons learnt during pre-validation. The created and validated language resources will be available via ELRA/ELDA.

1. Introduction

Language Resources (LR) are required for the development of automatic speech recognition (ASR), text-to-speech synthesis (TTS) and speech centered translation systems. Annotated speech databases needed for building speech recognition systems have been extensively developed in many languages and acoustic environments. However, there is a lack of written language resources specifically designed for voice driven applications.

The EU-project LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components) is oriented to producing such language resources. In this project, lexica for speech recognition and speech synthesis applications are developed containing phonetic, prosodic and morpho-syntactic information. In addition, statistical machine translation (SMT) resources consisting of bilingual and trilingual corpora in a specific domain that are aligned at the sentence level are developed for speech-based translation applications. The specifications, definitions of validation procedures and standards generated in the project can be seen as pioneering work in this field.

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). University of Maribor is an external partner to the consortium. SPEX (Speech Processing EXpertise) and CST (Center for Sprogteknologi) are responsible for validating the lexica.

For the time being the project covers 13 languages from all over the world: Catalan, Classical Arabic, Finnish, German, Greek, Hebrew, Italian, Mandarin, Russian, Slovenian, Spanish, Turkish, and US-English. The amount of languages may increase since the project is open for new external partners who will be given new languages and thus lexica to build.

In a consortium where each partner is responsible for the production of part of the LR, i.e. two languages per partner, it is important that each partner provides LR of equal quality ('E-quality'). Only E-quality in the final LR allows a fair exchange between partners at the end of the project as well as full benefit for the third parties for which the lexica are finally provided. Therefore, two independent validation centers were contacted by the consortium. These validation centers are currently both the official validation centers of the European Language Resources Association (ELRA): SPEX is responsible for validation of the formal and phonemic part of the lexica, and CST for the validation of the morphological and syntactic information. Both centers are independent in the sense that they have an academic background and do not produce any resources themselves in the project. Both validation centers were involved at an early stage of the lexicon production so that their validation activities could be of maximum benefit for the quality of the end products.

This paper focuses on the presentation of corpora, word list extraction, and lexica creation for ASR and TTS systems. There is special mention of the work carried out for specification of validation criteria, methodology and procedures.

The paper is organized as follows: Section 2 describes the specifications for corpora collection and word list extraction, section 3 describes the specifications and formats for the lexica, and section 4 describes the validation criteria and procedures. The paper concludes with the current status of the project and plans for availability and distribution of the data.

2. Overview of the Corpora and Word Lists

Large amounts of corpora have been collected from electronically available sources and used for the wordlist extraction of common words. The required minimum size for the corpus size was 10 million tokens of 'cleaned' text (e.g. with digits, punctuation marks and typos removed). However due to coverage requirements (cf. 2.2) the actual collected amounts were much higher.

2.1. Domains for Common Words, Proper Names and SAP Words

For common words, 6 major semantic domains were covered: sports and games, finance, news, culture (including travelling), consumer information and personal communication. For proper names, 3 major domains were covered: person names (including first and last names), place names and organizations, whereby the last two categories were further subdivided into subdomains. 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. A detailed description can be found in Hartikainen et al, 2003.

A special application word list consisted of numbers, letters, abbreviations and seven major semantic domains selected exclusively for voice driven applications. For the latter domains, a reference word list of 5,700 entries in US-English was collected and translated into all other languages covered by the project.

2.2. Word List Coverage and Specifications

For optimizing the coverage 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 contained language-dependent cleaning and tokenizing of the corpora, verifying the size requirements as well as removing the proper names and abbreviations. The final wordlists were free of digits, punctuation marks and most common typos. 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).

Remaining proper names and abbreviations were removed manually. 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 word list sizes ranged from 38,564 entries covering 100% of a corpus of over 20 million tokens for Mandarin Chinese to 140,000 entries for Finnish covering 95% of 17 million tokens. These results reflect the morphological diversity of the languages: Mandarin on the one hand has no inflection at all while Finnish and Russian, for example, are well known as highly-inflected languages. Another example is German where 100,000 entries had to be collected to meet the requirements due to word compounding. Although de-compounding methods were used in word list creation, German word list sizes were more comparable to those of Finnish and Russian.

3. Specifications and Requirements for Building the Lexica

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 & Montecchio, 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.

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, one or more POS, one or more lemmas and one or more phonetic transcriptions must be specified. Special tags are used to mark words included in the special application word list. 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 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 has the default value *NS* (=*Not Specified*) which is always implied, 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).

4. Validation Criteria and Processes

The LC-STAR project is the first project in which validation of the lexica has been addressed in such a complete and detailed way. All aspects of the lexica are validated, including orthography, phonetic transcription, suprasegmental aspects such as stress, syllabification and tones (in tone languages), as well as morphological and syntactic information.

Owing to the wide range of topics that are dealt with in the validation process, it was necessary to involve two independent validation centers: one for validating the formal and phonetic aspects (SPEX) and the other for validating morphological and syntactic information (CST).

4.1. Automatic vs. Manual Validation

Two types of validation are done: automatic and manual. Automatic tests are done on formal aspects that can be tested with software whereas manual checks are those that require sophisticated linguistic knowledge of the language.

The automatic checks test aspects such as:

1.	Correct numbers of entries per domain (names/words) are present according to the specifications
2.	Only valid phonetic symbols are used according to the documentation provided
3.	Only valid POS tags and attributes per POS are used according to the language-dependent specifications
4.	Proper XML format is used

Since a generic 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.

The manual checks test the correctness of spelling, phonetic transcriptions, suprasegmental aspects, POS tags and their corresponding attributes. In addition, the documentation is manually checked to ensure that those unfamiliar with the language in question will be able to fully understand the content of the lexicon for the future use.

4.2. Different Stages of Validation

Validation criteria were developed to be stringent enough to ensure a quality lexicon, but also to be realistic for lexica producers to accomplish. All this was ensured by a two-stage validation procedure, whereby a pre-validation check ensured that the lexica producers were “on the right track” and that no outstanding problems were envisioned before costly and time-consuming production of the full lexica. The pre-validation stage in itself consisted of two parts: one that checked the lists of entries, and another that checked a small subset of the lexicon (a “mini-lexicon”) that contained all aspects of the final full lexicon (including phonemic transcriptions and POS-tags).

After the full production of the lexica, full validation is done, similar to the validation of the “mini-lexicon”, though with some final added checks to ensure the total quality of the final output. These additional checks include adherence to minimal sizes of the full lexicon and its component parts according to the specifications (e.g. sufficient special application words and sufficient names of each category).

If outstanding problems are found at the full validation stage, it may be necessary for the producers to rework some parts of the lexicon, in which case a re-validation of the defective part becomes necessary.

4.3. Observations and Common Problems

The implemented validation procedure will ensure that top-quality lexica are produced while streamlining the effort for production. The pre-validation phase, which was completed at the beginning of 2004, was essential for this effort so that problems could be addressed at early stages of the production and typical problems could be shared among all producing partners. Common problems found at the pre-validation stage included:

1. Problems related to completeness checking of closed word classes (such as pronouns, indefinite/definite articles, conjunctions etc.). Word classes can be defined according to different theories (e.g. grammatical, linguistic) and differing methods (e.g. application oriented criteria). However, the word list specifications (Ziegenhain, 2003) never intended to go into such detailed linguistic or pragmatic definitions. Instead, the attention was paid to make the classes as relevant as possible for use in statistical Speech-to-Speech Translation (SST) applications. Therefore, well-known problem areas such as e.g. types of pronouns and the criteria for classifying them were unspecified, which resulted in problems during the pre-validation stage and uncertainty of which criteria to use these closed word classes.
2. Problems related to conflicting interpretation of linguistic POS tags/attributes and consequent ramifications for validation. Like previously mentioned, the purpose of the producers of the lexica was to make the lexica as useful for SST applications and thus the POS tags were defined according to these criteria. The validators' point of view however sometimes reflected that it was dealt with different linguistic backgrounds and these two standpoints were occasionally in contradiction.
3. Problems related to judging what constitutes an acceptable or possible pronunciation of an entry in any given language and the correct use of stress marks in multiword entries.
4. Insufficient language-specific documentation in the following cases:
   a. the main objective of including the closed word classes
   b. the instructions for POS-tagging
   c. the use of stress marks in phonemic transcriptions.
5. Varying quality of validations. Although the documentation was meant to contain the same information for all covered languages, the quality of the validation results varied among the languages. A number of different native experts were used for the manual validation, (one expert per language) with varying backgrounds and validation methods used. The insufficient documentation that was provided also had a contributing influence here.
6. The lack of information provided to the validators. For example, in validating the special application words, the native experts received only part of the needed information and thus they were not able to validate the entries appropriately. This deficiency was however rectified and the native experts were later provided with sufficient information.

The ensuing discussions between the consortium and the validators showed the need for improving the documentation so that validators would have a better understanding of the framework and criteria for which the lexica was created.

5. Conclusions and Remarks for the Future

Since there exist many approaches for interpreting and implementing linguistic and pragmatic data (like the lexica producers' vs. the validators' different approaches and points of views), it is important to provide exhaustive documentation in order to avoid any misinterpretations. The pre-validation played a very important role especially in clarifying approaches, point of views and different aspects requiring documentation. A lot of such problems were handled and solved and during this process, information in the lexica was therby unified and clarified.

At the beginning of 2004, directives for the documentation were finally completed and finalized as well as most of the final lexica and the final, full validation can be carried out during the spring-summer 2004. The final validation phase will take one or two months depending upon whether all the lexica are delivered at the same time or in sequence. After the project is finished in the beginning of 2005, all lexica will be available for public use via ELRA.

6. References

Shammass, S. & van den Heuvel, H., (2004). Specification of validation criteria for lexicons for recognition and synthesis", LC-STAR Deliverable D6.1. available from www.lc-star.com.

Ziegenhain, U., (2003). Specification of corpora and word lists in 12 languages. LC-STAR Deliverable D1.1. available from www.lc-star.com.

Hartikainen, E., Maltese, G., Moreno, A., Shammass, S., Ziegenhain, U., (2003). Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In Proceedings of Eurospeech 2003. Geneva, Switzerland.

Maltese, G. & Montecchio, C., (2003). General and language-specific specification of contents of lexica. LC-STAR Deliverable D2. available from www.lc-star.com.