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

Hanne Fersøe, Elviira Hartikainen, Henk van den Heuvel, Giulio Maltese, Asuncion Moreno, Shaunie Shammass, Ute Ziegenhain

1Center for Sprogteknologi (CST)  
Njalsgade 80, Copenhagen, Denmark  
hanne@cst.dk

2Nokia Mobile Phones  
Itämerenkatu 11-13, 00180 Helsinki, Finland  
elviira.hartikainen@nokia.com

3Speech Processing Expertise Center (SPEX)  
Erasmusplein 1, 6525 HT Nijmegen, Netherlands  
H.vandenHeuvel@let.kun.nl

4IBM Italy  
Rome, Italy  
giulio.maltese@it.ibm.com

5Universitat Politecnica de Catalunya  
Jordi Girona, 1-3, 08034 Barcelona, Spain  
asuncion@gps.tsc.upc.es

6NSC – Natural Speech Communication  
33 Lazarov St., 75150 Rishon Le Zion, Israel  
shaunie@nsc.co.il

7Siemens 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 (Asuncion, Henk)

- Brief description of the project (scope, participants, languages) (Asuncion)
- Overview of validation in general and short reasoning e.g. why, how and when it has to be done (Henk)

In a consortium where each partner is responsible for the production of part of the LR (i.e. in LC-STAR 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. 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): In LC-STAR, 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 resource 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.

- Goal of the paper: to present the corpora collection and lexica creation for TTS purposes as well as specify criteria for validation and describe the actual validation process (Asuncion)
- Structure of the paper (Asuncion)

2. Overview of the corpora and word lists (Ute)

- Domains for common words/proper names/SAP words
- Coverage issues
3. Specifications and requirements for building the lexica (Giulio)

- Requirements for a common DTD for all languages
- Which information present in lexica (e.g. phonetic, prosodic, morpho-syntactic) and how
- Problems, lessons learned

4. Validation criteria and processes (Shaunie, Hanne, Henk)

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), the other for validating morphological and syntactic information.

Two types of validation are done: automatic and manual. Automatic tests are done on formal aspects that can be tested with software. 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 orthographic and 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.

Software was developed within the scope of the project specifically for testing all formal aspects of the lexica. Since a generic DTD was written to capture all formal features of the lexica, a lot 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, such as for sufficient coverage of various domains, missing POS tags etc. special software had to be written, which was done in Perl.

The manual checks test for 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 future practical use.

Validation criteria were developed that were stringent enough to ensure a quality lexicon, but that were realistic for lexica producers to accomplish. 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 a 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 full production of the lexica, full validation is done, similar to the validation of the “mini-lexicon”, but 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 individual parts (e.g. sufficient special application words, 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.

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

1. problems related to completeness checking of closed word classes. Word classes can be defined according to different theories (grammatical, linguistic) and methods (application oriented criteria), but the word list specifications (Ziegenhain, 2003) never intended to go into such detailed linguistic or pragmatic definitions. Therefore well known problem areas such as e.g. types of pronouns and the criteria for classifying them were unspecified, leaving the validators with the job of determining the basic criteria in order for them to be able to validate the word lists.
2. problems related to conflicting interpretation of linguistic POS tags/attributes and consequent ramifications for validation
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. early detection of formal problems, such as invalid symbols, etc.
5. insufficient language-specific documentation, regarding:
   a. the objective of the closed word classes
   b. the instructions for POS-tagging
   c. the use of stress marks in phonemic transcriptions

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

5. Conclusions (Elviira with contribution of others)

- time schedule for final production/validation
- availability/distribution of the lexica

6. References
List here all the references you used

Shaunie Shammass and Henk van den Heuvel, “Specification of validation criteria for lexicons for recognition and synthesis”, LC-Star Deliverable 6.1.

Ziegenhain, U. (2003). Specification of corpora and word lists in 12 languages. LC-STAR Deliverable D1.1.