LC-STAR II: starring more lexica

Ute Ziegenhain, Hanne Fersøe, Henk van den Heuvel, Asuncion Moreno
Siemens AG,
Otto-Hahn-Ring 6, 81739 Munich, Germany
Center for Sprokteknologi (CST)
Njalsgade 80, Copenhagen, Denmark
Speech Processing Expertise Center (SPEX)
Erasmusplein 1, 6525 HT Nijmwegen, The Netherlands
Universtitat Politecnica de Catalunya (UPC)
Jordi Girona 1-3, 08034 Barcelona, Spain
email: ute.ziegenhain@siemens.com, hanne@cst.dk, H.vandenHeuvel@let.kun.nl, asuncion@gps.tsc.upc.es

Abstract

LC-STAR II is a follow-up project of the EU funded project LC-STAR (Lexica and Corpora for Speech-to-Speech Translation Components, IST-2001-32216). LC-STAR II develops large lexica containing information for speech processing in ten languages targeting especially automatic speech recognition and text to speech synthesis but also other applications like speech-to-speech translation and tagging. The project follows by large the specifications developed within the scope of LC-STAR covering thirteen languages: Catalan, Finnish, German, Greek, Hebrew, Italian, Mandarin Chinese, Russian, Turkish, Slovenian, Spanish, Standard Arabic and US-English. The ten new LC-STAR II languages are: Brazilian-Portuguese, Cantonese, Czech, English-UK, French, Hindi, Polish, Portuguese, Slovak, and Urdu. The project started in 2006 with a lifetime of two years. The project is funded by a consortium, which includes Microsoft (USA), Nokia (Finland), NSC (Israel), Siemens (Germany) and Harmann/Becker (Germany). The project is coordinated by UPC (Spain) and validation is performed by SPEX (The Netherlands), and CST (Denmark). The developed language resources will be shared among partners.This paper presents a summary of the creation of word lists and lexica and an overview of adaptations of the specifications and conceptual representation model from LC-STAR to the new languages. The validation procedure will be presented too.

1. Introduction

Continuing the success of the EU funded project LC-STAR (IST-2001-32216) a non funded follow-up project LC-STAR II has been started in 2006. The objective of LC-STAR II is the creation of large lexica with phonetic and linguistic information targeting especially automatic speech recognition and text to speech synthesis but also other applications like speech-to-speech translation and tagging. The project follows by large the specifications developed within the scope of LC-STAR covering thirteen languages: Catalan, Finnish, German, Greek, Hebrew, Italian, Mandarin Chinese, Russian, Turkish, Slovenian, Spanish, Standard Arabic and US-English. The ten new LC-STAR II languages are: Brazilian-Portuguese, Cantonese, Czech, English-UK, French, Hindi, Polish, Portuguese, Slovak, and Urdu. The project started in 2006 with a lifetime of two years. Producing partners in the current project are: Harmann/Becker (Germany), Microsoft (USA), Nokia (Finland/China), NSC (Israel), Siemens AG (Germany). The project is coordinated by UPC (Spain). Validation is performed by SPEX (The Netherlands) and CST (Denmark). The lexical databases will be exchanged by the producing partners with guaranteed rights of use. Some of the resources will also be made available via the ELRA/ELDA channel (France).

The paper is organized as follows: we present a short description on the requirements to create large word lists from large electronic corpora (Ziegenhain et al. 2003) and the changes in requirements on coverage. The format and the linguistic content (Maltese et al. 2005) as well as a few examples are provided too. Additional features which were necessary to cover the new languages will be detailed. An overview of the validation procedure will be presented.

2. Corpora Domains And Requirements

2.1. Semantic domains for word list creation

Following LC-STAR specifications large text corpora for common words (Ziegenhain et al., 2003) were collected for six major semantic domains: news, finance, sports and games, culture and entertainment, consumer information and personal communications (newsgroups, editors' notes, etc.). The semantic domains for proper names include three major domains: person names (including first and last names and other [1]), place names (major cities, geographical names, addresses, etc.) and organisations (e.g. companies, non-profit organisations, brand names). In addition to the list of common entries and proper names a list of entries from closed set category has to be created manually for each given language.

Furthermore a special application word (SAP) list which has been collected in LC-STAR I for the purpose of voice driven applications is included in the lexica. The SAP list which contains approximately 5700 entries is a manual collection of entries in seven semantic domains which are partially different from the ones described above. The SAP list also includes abbreviations (e.g. web domains, most commonly used abbreviations, ISO language abbreviations, etc.), letters and all natural numbers of a given language. The basic idea of creating the SAP list

[1] e.g. for Russian patronymic names have been included

was to include entries which were likely not to occur frequently in the large corpora or which are deleted from the word lists during the tokenization process (e.g. numbers when presented as digits are discarded from the final word lists). Except for letters and numbers all entries are provided in English (embedded in example sentences) and translated into the target languages. Entries from the SAP list have a special ID to mark them as such.

2.2. Requirements

In the following a short overview of the requirements on size and corpora coverage are presented. A detailed description can be found in in Ziegenhain et al. (2003). Requirements on size for the six major domains are: collect corpora of at least 10 Mio of tokens over all six semantic domains and at least 1 Mio of tokens in each domain. Electronic text material should be preferred and the texts should be no older than five years.

The word lists for common words were then 'cleaned' from all entries with frequency one, from all digits and special characters to reach a target of 95% self-coverage on the self collected corpus. The final common word lists for all languages contain at minimum 50.000 common word entries (inflected forms) to meet the requirements with no upper limit depending on the language and the complexity of its morphological structure. One result of the first project was that the richer the morphological structure the higher the number of entries.

The word list for the proper names in each language contains a minimum of 45.000 entries and the special application list at least 5.500 translated entries plus numbers and letters again depending on the language.

2.2. Changes in requirements

During the first project for Mandarin some exceptions to the specifications were required. In LCSTAR II it turned out that for Hindi and Urdu the electronically available material was not sufficient to meet the lower limit of 50.000 entries. An increase in corpus size was therefore the first step.

For Hindi the size was increased from 10 Mio to more than 14 Mio resulting in a word list with 45. 506 words and a coverage on 100% of the corpus. Adding new material did not yield any better results. The size of the word list was then increased by adding spelling variants of compound words occuring equally frequent in the corpora as different entries. E.g. ख़रीदनेवाला or ख़रीदने वाला 'buyer'. In a second step entries with low frequency (between two and four) which had been discarded in the beginning were also included to reach the final size of 51.000 entries.

In Urdu the situation was even worse: about 31. 000 common words were extracted from a corpus of 10 million already reaching the target of 100 coverage. An increase in corpus size to 12 million only yielded about 2000 new words most of which were singletons (which normally would be discarded from the final list). It seems that the online corpora for Urdu have only a very limited vocabulary up till now [2]. It was therefore agreed to use other resources (literature, dictionaries) related to the defined domains to meet the requirements on final size.

3. Format And Linguistic Content

3.1. Format

In the next paragraphs the format and linguistic content of the lexica as well as the changes will be presented. For a detailed description of the format and content also see (Maltese et al., 2005) and (Hartikainen et al., 2003). The lexica are coded using XML mark-up language.

The main reason for using an XML/DTD format were well-known advantages like:

- widely known technique
- many tools supporting it are available
- supports Unicode, so also languages with writing systems that are not based on the Latin alphabet can be represented adequately
- allows easy and concise representation of one-to-many relations (one word having multiple pronunciations, one word having multiple POS codes, one abbreviation having multiple expansions, etc.)
- easily definable and flexible syntax,
- easy well-formedness tests are possible using publicly available tools.

At the beginning of the first LC-STAR project no generally accepted standard XML/DTD structure for lexicons that was suited to the purposes of our project already existed although the formal specifications of PAROLE and other projects lexica were taken into account (e.g. Ruimy et al. (1998). For this reason, an XML format and a DTD specific to the LC-STAR project has been developed.

The experiences within the first project showed that the format is very flexible and easily extendible to new languages. So for exchange and consistency purposes with LCSTAR lexica the same format has been used with slight modifications of the generic DTD. The language specific DTD's which have been developed during the project are subsets of the generic DTD. The generic DTD now covers twenty-three languages from all over the world and has been made publicly available on the web-side (http://www.lc-star.org/).

3.2. Linguistic information

All LCSTAR lexica consist of so called entrygroup (ENTRYGROUP) elements. An entrygroup is itself defined as a unit that is identified by it's canonical orthography (alternative spellings allowed). An entrygroup consists of one ore more entry elements (e.g. multiple part of speech elements, multiple pronunciations, compound entries, etc.). One advantage of the approach is that new languages especially non European languages can easily be integrated without changing the format. In

[2] The same occurs in other languages like e.g Hindi, Telugu, or others where online resources are scarce.

each entrygroup it is mandatory to provide lemma and the POS or word class represented by the orthographic string as well as phonetic and prosodic information. For the phoentic content we use the SAMPA notation (http://www.phon.ucl.ac.uk/home/sampa/index.html).
For languages where no standard symbol set exists (e.g. Hindi, Urdu, Brazilien_Portuguese) we use as a basis either sets developed in other speech data collection projects (e.g. for Hindi from the LILA project) or the set will be developed by native phonetic experts mapping the IPA transcription to the corresponding SAMPA notation..
The tag set and linguistic attributes were originally based on the EAGLES tag set (EAGLES 1996) developed for European languages. However it was quite obvious that the EAGLES recommendations had to be adapted to also cover non European languages. The list of POS tags and attributes had therefore been already been enlarged in LCSTAR especially for Chinese, Arabic and Turkish and again in LCSTAR II for Hindi and Urdu. The changes necessary for LCSTAR II are described in section 3.3.
The twenty-three POS tags itself have internal morphological and semantic attributes which are either used by all languages (e.g. number), by groups of languages (e.g. gender for European languages) or are language specific (e.g. 'harp' for Indian languages, semantic attributes like 'as_if' for Turkish adverbs etc.)
In the following paragraph a few examples of lexicon entries[3] and their structure are provided.

3.2.1 'Normal' entries

Normal entries occur in the lexicon in their normalized orthography (ENTRYGROUP) and consists of one or more entry elements (ENTRY). For example in English the orthographic form 'I' could either present the letter (LET) or a personal pronoun (PRO):

```
<ENTRYGROUP orthography="I">
        <ENTRY>
            <LET/>
            <LEMMA>I</LEMMA>
            <PHONETIC>" aI</PHONETIC>
        </ENTRY>
        <ENTRY>
            <PRO number="singular" person="1"
type="personal"/>
            <LEMMA>I</LEMMA>
            <PHONETIC>" aI</PHONETIC>
        </ENTRY>
    </ENTRYGROUP>
```

or an example from Russian *лабиринт* (labyrinth)

```
<ENTRYGROUP orthography="лабиринт">
        <ENTRY>
            <NOM class="common" number="singular"
gender="masculine" case="accusative" type="not_animated"/>
            <LEMMA>лабиринт</LEMMA>
             <PHONETIC>l a - b' i - " r' i n t</PHONETIC>
        </ENTRY>
        <ENTRY>
            <NOM class="common" number="singular"
gender="masculine" case="nominative" type="not_animated"/>
            <LEMMA>лабиринт</LEMMA>
            <PHONETIC>l a - b' i - " r' i n t</PHONETIC>
        </ENTRY>
</ENTRYGROUP>
```

The common noun (NOM) лабиринт (labyrinth) can either be analyzed as the accusative singular or nominative singular form. In case of ambigous word forms the entry is simly doubled which leads to an increase in lexicon size. However the format is more clearly arranged and more easily processed..

3.2.2. Complex entries

Some complex entries enter the lexicon as 'normal' entries for better performance in speech recognition, synthesis and especially speech-to-speech translation. Often they are from Latin origin (xml:lang="la") but the essential criterium to treat them as a one token entry is that the original reading is often substituted. Formally the blanks between these tokens are replaced by underscore: for example *nota bene* -> *nota_bene*. For these entries the syntactic category of the language in which they enter the lexicon is provided omitting that in the native language the category might be a different one. In the example below the latin phrase *nota bene* is marked as an adverb and the lemma consists of the complex form:

```
<ENTRYGROUP orthography="nota_bene" xml:lang="la">
        <ENTRY>
            <ADV/>
            <LEMMA>nota_bene</LEMMA>
            <PHONETIC>n " @U - t @ # b " e - n
eI</PHONETIC>
        </ENTRY>
    </ENTRYGROUP>
```

3.2.3. Compound entries and contractions

In contrary to complex entries compound entries and contractions (ENTRY_COMP) are entries which are composed out of two parts where the second part is glued to the first part (e.g. pronouns in Italian) or a shortened form (clitic). These forms are frequent in many languages and were not separated for better performance. For example the English form *academy's* can be analyzed as either the genitive singular form of the noun *academy* (as such it enters the lexicon as a single entry). On the other hand it is composed out of the noun and the clitic form of the auxiliary *is* (or *has* - but this reading is rather rare in written text). For the latter reading the entry is treated as a cmpound entry. For each part a list of the categories is provided which are links to other entries:

```
<ENTRYGROUP orthography="academy's">
        <ENTRY>
            <NOM class="common" number="singular"
type="possessive"/>
            <LEMMA>academy</LEMMA>
            <PHONETIC>@ - k " { - d @ - m I z</PHONETIC>
```

[3] the examples will be provided in either UK English, Polish, Hindi, Turkish or Russian from lexica owned by Siemens AG

```
        </ENTRY>
        <ENTRY_COMP>
            <PHONETIC>@ - k " { - d @ - m I z</PHONETIC>
            <ENTRY_EL orthography="academy">
                <NOM class="common" number="singular"/>
            </ENTRY_EL>
            <ENTRY_EL orthography="is">
                <AUX number="singular" person="3"
tense="present" type="finite"/>
            </ENTRY_EL>
        </ENTRY_COMP>
```

3.2.4. Abbreviations

Entries that need to be expanded to be pronounced properly are labeled as abbreviations. For these words the lemma and detailed POS information will be provided and in addition a phonetic representation for the expanded form:

```
<ENTRYGROUP orthography="Mr">
        <ABB>
            <EXP expansion="Mister">
                <ENTRY>
                    <NOM class="common" number="singular"/>
                    <LEMMA>Mister</LEMMA>
                    <PHONETIC>m " I - s t @</PHONETIC>
                </ENTRY>
            </EXP>
        </ABB>
    </ENTRYGROUP>
```

Acronyms (e.g. IBM, NASDAQ, AIDS, etc.) are treated as 'normal' entries. No expansions are therefore provided. A special tag allows to mark them as spelled which is for example of interest for text-to-phoneme training.

3.2.5. Numerals

For English numerals are classified as either type 'cardinal' or 'ordinal' respectively 'ratio'. Other types like 'reified' and 'collective' to account for the complex Polish numeral system are available too.
Examples are:

```
<ENTRYGROUP orthography="eighty">
        <ENTRY>
            <NUM number="singular" type="cardinal"/>
            <LEMMA>eighty</LEMMA>
            <PHONETIC>" eI - t I</PHONETIC>
        </ENTRY>
    </ENTRYGROUP>

<ENTRYGROUP orthography="eleventh">
        <ENTRY>
            <NUM number="singular" type="ordinal"/>
            <LEMMA>eleven</LEMMA>
            <PHONETIC>I - l " e - v @ n T</PHONETIC>
        </ENTRY>
        <ENTRY>
            <NUM number="singular" type="ratio"/>
            <LEMMA>eleven</LEMMA>
            <PHONETIC>I - l " e - v @ n T</PHONETIC>
        </ENTRY>
    </ENTRYGROUP>
```

```
<ENTRYGROUP orthography="dwójka">
        <ENTRY>
            <NUM number="singular" gender="feminine"
case="nominative" type="reified" />
            <LEMMA>dwójka</LEMMA>
            <PHONETIC>" d v u j - k a</PHONETIC>
            </ENTRY>
</ENTRYGROUP>
```

Polish numbers are much more complex than the simple regular adjectives. Certain numbers act more like nouns than adjectives in specific case forms, while they act more like adjectives in others. E.g. reified numerals, which are feminine nouns ending in *-ka*, are used to refer to items by numerical designation: 1 *jedynka*, 2 *dwójka*, 3 *trójka*, etc. For example, *dziesiątka* could be used to refer to room number 10; a 10-millimeter wrench; a bus number 10, and so on. Reified numerals may be used colloquially in place of collective numerals: *dwójka dzieci* a couple of kids.

3.2.6. Special symbols (punctuation marks and keyboard symbols)

In general punctuation marks are not spoken. But in certain contexts like for example email or web addresses or in applications like dictation they are pronounced. A typical entry has the following structure:

```
<ENTRYGROUP orthography="!">
        <ENTRY>
            <PUN/>
            <LEMMA>!</LEMMA>
            <PHONETIC>" E k s k l @ m eI Z n m A
k</PHONETIC>
            <APP>
                <SBD type="1.3." entries="2"/>
            </APP>
        </ENTRY>
    </ENTRYGROUP>
```

where the tag APP points to the SAP list entry which are indentified by type and number.

3.3. Changes

For LCSTAR II remarkably few changes in the linguistic content were necessary to cover the linguistic properties of the ten new languages:
A new feature for the treatment of liasion phenomena and schwa elision in French has been introduced. The approach is similar to the one desribed in Ferrane et al. 1992. In addition two more POS tags had to be added for treatment of special morphological and case marking features in Urdu (Madiha Ijaz & Sarmad Hussain 2007, Butt Miriam 2005): Harf (HAR) and case carker (CM). 'Harf' is a labelling for an unbound morpheme (clitic) which has no meaning unless it is combined with other words.
Following the theory outlined in Butt & King 2005 case markers which are separated by blank were introduced as a special POS feature in Urdu (and other languages) rather than a morphological feature. For more detailed information cf. (Butt 2005).
As already mentioned in previous paragraphs some additional feature types had to be added to account e.g. for the complex numeral and morphological system in Polish,

Czech and Slovac. All necessary changes have been included in the language specific descriptions and DTD's but also in the generic DTD.

4. Validation

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

The automatic checks address aspects such as:

- the minimum numbers of entries per domain (names/words) are provided
- coverage of the resulting entries on the original corpora is sufficiently high
- only valid orthographic and phonetic symbols are used
- only valid POS tags and attributes per POS are used according to the language-dependent specifications
- 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. 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 software has been distributed to all partners.

The manual checks deal with the correctness of spelling, phonetic transcriptions (including stress and syllabification), the correct assignment of POS tags and their corresponding attributes, and the correctness and completeness of lists of words and word forms belonging to the closed word classes. 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 warrant a high quality lexicon, but that were realistic for lexica producers to accomplish. The validation centres implemented a two-stage validation procedure: pre-validation and full validation. First, a *pre-validation* check ensured that the lexica producers were "on the right track" and that no outstanding problems were to be expected before the costly and time-consuming production of the full lexica. This pre-validation stage in itself consisted of two parts: one that checked the lists of envisaged lexicon 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). Second, after full production of the lexica, a *full validation* is carried out, similar to the validation of the "mini-lexicon", but with some final added checks to ensure the total quality of the final lexicon. 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).

At the time of editing this paper, all lexica have been pre-validated and five out of ten have been validated

5. Conclusion

The LC-STAR approach to create lexica for speech applications has been successfully applied to twenty-three languages up till now. The consortium is therefore planning to set up a third project to enlarge the portfolio of languages. New partners are welcome to join the project. Information will be provided in the LC-STAR II webpage.

6. References

Butt, Miriam and Tracy Holloway King (2005). The Status of Case. In Veneeta Dayal and Anoop Mahajan (Eds.). *Clause Structure in South Asian Languages*, Berlin: Springer Verlag, pp. 153–198.

Butt, Miriam (2005b) The Dative/Ergative Connection. http://ling.uni-konstanz.de/pages/home/butt/ (March 27th, 2008)

EAGLES Guidelines for Computational Lexica. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages (1996). EAGLES: http://www.ilc.cnr.it/EAGLES/home.html (March, 27th, 2008).

Ferrane, I. et al. (1998). Besoins lexicaux a la lumiere de l'analyse statistique du corpus de textes du projet "BREF": le lexique "BDLEX" du francais ecrit et oral. In *Proceedings of COLING*, pp. 1203 - 1208.

Hartikainen, E. et al. (2003). Large Lexica for Speech-to-Speech Translation: From Specification to Creation In *Proceedings of Eurospeech*, pp. 1529-1532.

Maltese, G. et al. (2004). General and language-specific specification of contents of lexica in 13 languages. Deliverables D2 of the LC-STAR project. Public Documents: http://www.lc-star.org/ (March 27th, 2008)

Madiha Ijaz, Sarmad Hussain (2007). Corpus Based Urdu Lexicon Development. In *Proceedings of Languge and Technology, Peshawar, Pakistan* ,pp. 1-12.

Nilda Ruimy et al. (1998). LE-PAROLE project: The Italian Syntactic Lexicon. In *EURALEX'98*, Université de Liège, Proceedings Volume I p. 259.

Ziegenhain et al. (2003). Specification of corpora and word lists in 12 languages. Deliverable D1.2. of the LCSTAR project. Public Documents: http://www.lc-star.org/ (March 27th, 2008).