SI-PRON: a Pronunciation Lexicon for Slovenian

Jerneja Žganec Gros[1], Varja Cvetko-Orešnik[2], Primož Jakopin[2], Aleš Mihelič[1]

[1]Alpineon Research and Development
Alpineon d.o.o., Ulica Iga Grudna 15, SI-1000 Ljubljana, Slovenia
email: jerneja@alpineon.com, ales@alpineon.com
[2]Fran Ramovš Institute of the Slovenian Language
Scientific Research Centre of the Slovenian Academy of Sciences and Arts
Gosposka ulica 13, SI-1000 Ljubljana, Slovenia
e-mail: cvetko@zrc-sazu.si, primoz.jakopin@guest.arnes.si

Abstract

We present the efforts involved in designing SI-PRON, a comprehensive machine-readable pronunciation lexicon for Slovenian. It has been built from two sources and contains all the lemmas from the *Dictionary of Standard Slovenian* (*SSKJ*), the most frequent inflected word forms found in contemporary Slovenian texts, and a first pass of inflected word forms derived from *SSKJ* lemmas. The lexicon file contains the orthography, corresponding pronunciations, lemmas and morphosyntactic descriptors of lexical entries in a format based on requirements defined by the W3C Voice Browser Activity. The current version of the SI-PRON pronunciation lexicon contains over 1.4 million lexical entries.

The word list determination procedure, the generation and validation of phonetic transcriptions, and the lexicon format are described in the paper. Along with Onomastica, SI-PRON presents a valuable language resource for linguistic studies and research of speech technologies for Slovenian. The lexicon is already being used by the AlpSynth slovenian text-to-speech synthesis system and for generating audio samples of the *SSKJ* word list.

1. Introduction

Consistent specification of word pronunciation is critical to the success of many speech technology applications. Most state-of-the-art Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) systems rely on lexicons, which contain pronunciation information for many words. To provide for a maximum coverage of the words, multi-word expressions or even phrases, which commonly occur in a given application-domain, application-specific word or phrase pronunciations may be required, especially for application-specific proper nouns, such as personal names or location names.

Several guidelines have been reported to define the structure of a pronunciation lexicon, ranging from simple two-column ASCII lexicons providing the mapping between graphemic and phonemic transcriptions, to more general de-facto standards and new standardization attempts, which are also handling multiple orthographies and multiple pronunciations.

The ISO-TC37 initiative, which started at LREC 2002, initiated work on a family of ISO standards related to natural language processing (Romary et al., 2006). Currently these standards are available in working drafts of high-level specifications for word segmentation, feature structures, annotations, and also for lexicons. The high-level specifications build on lower-level specifications in form of language and country codes, data categories, code scripts, and Unicode. Lexicon specifications are covered by the “Lexical Markup Framework” under ISO 24613 (Romary et al., 2006). The same description structure in terms of morphology, syntax and semantics (and translation) applies to monolingual up to multilingual lexicons. Multi-word expressions are given special attention.

Another initiative, the W3C Voice Browser activity, has recently issued a last-call working draft of the Pronunciation Lexicon Specification (PLS) Version 1.0 (W3C PLS Version 1.0, 2006), which is expected to be soon submitted as a W3C candidate recommendation. The PLS document was designed to enable interoperable specification of pronunciation information for both ASR and TTS engines within voice browsing applications. The mark-up language allows one or more pronunciations for a word or phrase to be specified using a standard pronunciation alphabet or if necessary using vendor specific alphabets. Pronunciations are grouped together into the PLS document which may be referenced from other markup languages, such as the Speech Recognition Grammar Specification (SRGS) and the Speech Synthesis Markup Language (SSML).

The Pronunciation Lexicon Markup Language, based on PLS, is designed to allow open, portable specification of pronunciation information for speech recognition and speech synthesis engines. The language is intended to be easy to use by developers while supporting the accurate specification of pronunciation information for international use.

The LC-STAR project consortium published another set of recommendations for speech technology lexicons, with an emphasis on application in machine translation, speech recognition and speech synthesis (Shamas & van den Heuvel, 2004; Fersøe et al., 2004). A slovenian lexicon, produced by the University of Maribor, has been built in the scope of the project (Verdonik et al., 2004). Compared to the LC-STAR lexicon specifications, the current version of PLS lacks description specifications for more complex features, describing morphological, syntactic, and semantic features of lexical entries.

In Slovenian, lexical stress can be located on almost any syllable obeying hardly any rules. The stressed syllable in Slovenian may form the ultimate, the penultimate or the preantepenultimate syllable of a polysyllabic word. Speakers of Slovenian have to learn lexical stress positions along with learning the language. As a consequence, a pronunciation lexicon indicating lexical stress positions for as many Slovenian words as possible is crucial for the development of speech technology applications and linguistic research. Such a lexicon can be used either in its full-blown form or as a training material for machine learning techniques aimed at automatically predicting word pronunciations.

Several attempts towards pronunciation lexicon construction for Slovenian have been reported so far (Derlić & Kačič, 1997; Gros & Mihelič, 1999; Gros et al., 2001; Šef et al., 2002; Verdonik et al., 2002; Mihelič et al., 2003). However, none of them have used the full lemma set as given in the *Dictionary of Standard Slovenian* (*SSKJ*) (SSKJ, 1991).

The paper describes the construction of a comprehensive reference pronunciation lexicon for Slovenian based on two sources: the information from the *SSKJ* and another list of the most frequent inflected word forms, which has been derived by an analysis of contemporary slovenian text corpora.

2. The SI-PRON Pronunciation Lexicon

2.1. SI-PRON Word List

The work on designing a new pronunciation lexicon begins with the selection of words, multi-word expressions or phrases, which will be represented in the lexicon. Several word-list selection procedures are known (Ziegenheim, 2003).

The construction of the SI-PRON lexicon started with the complete lemma word list of 93,154 entries from the *SSKJ* provided by the Fran Ramovš Institute of the Slovenian Language, equipped with basic lexical stress information on the stressed vowels and pronunciation exceptions. The complete word pronunciations still had to be determined.

In order to further expand the SI-PRON word list, we are augmenting the *SSKJ* lemma descriptions with part-of-speech information and declension/conjugation categories (Toporišič, 1991), specifying the inflectional paradigms of the lemmas. Irregular inflected word forms are processed separately. Using automatic procedures, we are fully expanding the lemmas into inflected word forms. So far, over 1 million lexemes containing lexical stress information have been derived.

Since *SSKJ* contains many words derived from literary texts, and they are not so common in everyday situations, we decided to upgrade the SI-PRON pronunciation lexicon with a list of 50,000 most frequent inflected word forms whose lemmas are not covered by the *SSKJ* word list. This additional word list has been derived from a statistical analysis of a contemporary Slovenian text corpus. The corpus comprising over 3 million Slovenian words was composed mainly from fiction and mainstream Slovenian newspaper texts: *Delo*, *Večer*, and the former *Slovenec*. After tokenization and the elimination of numerals, named entities, acronyms, and abbreviations, the remaining text corpus included over 35 million tokens. Acronyms, abbreviations, and named entities were stored into separate word lists.

A statistical analysis performed on the text corpus showed that about 50.000 most frequent words accounted for approaching 95% of all words used in the text corpus (Gros & Mihelič, 1999). These words form the main additional word list. They were additionally equipped with part-of-speech tags indicating the part-of-speech function of the words in the text corpus.

2.2. Collocations and Multi-word Expressions

The identification of collocations, i.e. current combinations of words as they appear in context, can considerably increase the naturalness of synthetic speech. In human speech, collocations act as prosodic units and are subject to a higher degree of reduction and internal coarticulation than they would be had they been ordinary, separate words. We have chosen a lexical approach for handling collocations. The most common collocations or multi-word expressions, reflexive verbs included, are stored in a separate pronunciation lexicon.

3. Deriving SI-PRON Phonetic Transcriptions

We have developed a tool to automatically derive word pronunciations for the *SSKJ* inflected words, by looking-up their stem pronunciation and appending that of the correct inflection from inflectional paradigms and morphological rules of Slovenian (Toporišič, 1991).

Therefore, the pronunciations of lexemes have been derived automatically for the *SSKJ* and *SSKJ* inflected word lists (about 2,500 entries, mainly words of foreign origin and not obeying the general Slovenian pronunciation rules, have been manually transcribed), and semi-automatically for the rest of the word list. Automatic lexical stress assignment and automatic grapheme-to-phoneme conversion rules have been used to process the latter.

3.1. Lexical Stress Assignment

The automatic lexical stress assignment algorithm for unseen words, which we applied is to a large extent determined by (un)stressable affixes, prefixes, and suffixes of morphs based upon observations by linguists (Toporišič, 1991).

For words that do not belong to these categories, the most probable stressed syllable is predicted using the results obtained by a statistical analysis of stress position depending on the number of syllables within a word (Gros & Mihelič, 1999).

3.2. Grapheme-to-Phoneme Rule Set for Slovenian

A collection of over 190 context-sensitive and context-free grapheme-to-allophone rules from the *AlpSynth standard words rule* set (Žganec Gros, 2006) translate each grapheme string into a series of allophones.

The rules are accessed sequentially until a rule that satisfies the current part of the input string is found. The transformation defined by that rule is then performed, and a pointer is incremented to point at the next unprocessed part of the input string, and so on until the whole string has been converted.

The context free rules are rare and they include a one-to-one correspondence, two-to-one correspondence and one-to-two correspondence.

The vast majority of the rules for grapheme-to-allophone transcription for Standard Slovene are context-sensitive. This means that a grapheme or a string of graphemes is transcribed differently according to its phonetic environment. Certainly all rules for determining which allophone of a certain phoneme is to be used in a phonetic sequence are context-dependent.

Each context-sensitive rule consists of four parts: the left context, the string to be transcribed, its right context and the phonetic transcription. A number of writing conventions has been adopted in order to keep the number of rules relatively small and readable. The left and the right context may contain wild characters describing larger phonetic sets, e.g.: '#' stands for vowels, '$' for consonants, '_' for white space.

The rules for consonants are rather straightforward, while those for vowels must handle vowel length and the variant realizations of the orthographic /e/ and the orthographic /o/ in stressed syllables.

A typical grapheme-to-allophone rule in the *AlpSynth standard words rule* set has the following structure:

left context	grapheme string	right context	allophone string
$	**/er/**	_	[@r]
=	/**n**/	k	[N]

The first rule says that the word final /er/ preceded by a consonant is transcribed as [@r] (e.g. /gaber/ -> [*ga:.b@r]). The second rule implies that any /n/ followed by /k/ is transcribed into [N] ([N] is the allophone of [n] when followed by /k/ or /g/, e.g. in /anka/ -> [*a:N.ka]).

Our initial rule set based on the one produced in 2001 (Gros et al., 2001) was based on various observations of expert linguists, e.g. (Toporišič, 1991) and other basic rule sets for Slovene grapheme-to-allophone transcription (Gros & Mihelič, 1999).

The initial set of rules has been undergoing continuous refinement ever since and resulted 194 rules of the *AlpSynth standard words rule* set (Žganec Gros, 2006). Rules for coarticulatory pronunciation corrections of words according to the words' left and to the right context are included.

In the recent years, telecommunication applications of ASR and TTS have increased in importance, e.g. automatic telephone directory inquiry systems. Names of locations (cities, streets, etc.) and other proper names cannot be mentally reconstructed from the context when listening to the messages, correct name pronunciation is required. The *AlpSynth standard word rules* developed for a standard Slovene vocabulary do not lead to satisfactory results when applied on names. Therefore, additional 'name-specific' rules were added to the final *AlpSynth standard words rule* set resulting in the *AlpSynth names rule* set.

3.3. Transcription Accuracy Experiment

In the previous subsection we describe, how a rule set for grapheme-to-allophone conversion of Slovene texts has been improved and evaluated. Another rule set has been developed for pronunciation of named entities. The phonemization errors were determined by comparing the automatic transcription outputs with manually verified pronunciation lexicon transcriptions.

A performance test applied on the SI-PRON *SSKJ*-based word list pronunciation lexicon showed error rates of about 25% in the stress assignment of unknown words and consequently in the phonetic transcription. If stress assignment and the transcriptions of graphemic /e/ and /o/ in stressed syllables was manually verified or known in advance, a transcription success rate of 99.01% was achieved for standard *SSKJ* words. A closer examination of the mismatches revealed that the majority of the errors could be attributed to inconsistencies in manual labeling when handcrafting the original *SSKJ*.

As a consequence, we argue that, in order to semi-automatically derive phonetic transcriptions for Slovenian words not covered by the lexicon with a 0.01% error rate, only manual validation of the stress position and its type have to be carried out, starting from automatically predicted stress positions. The rest can be performed automatically by applying our upgraded grapheme-to-phoneme conversion rule set.

4. SI-PRON Format

The SI-PRON lexicon format complies with the Pronunciation Lexicon Specification (PLS) Version 1.0, a W3C Voice Browser Activity working draft of syntax specification for pronunciation lexicons (W3C PLS Version 1.0, 2006). This lexicon specification has been recommended for use by speech recognition and speech synthesis engines in voice browser applications.

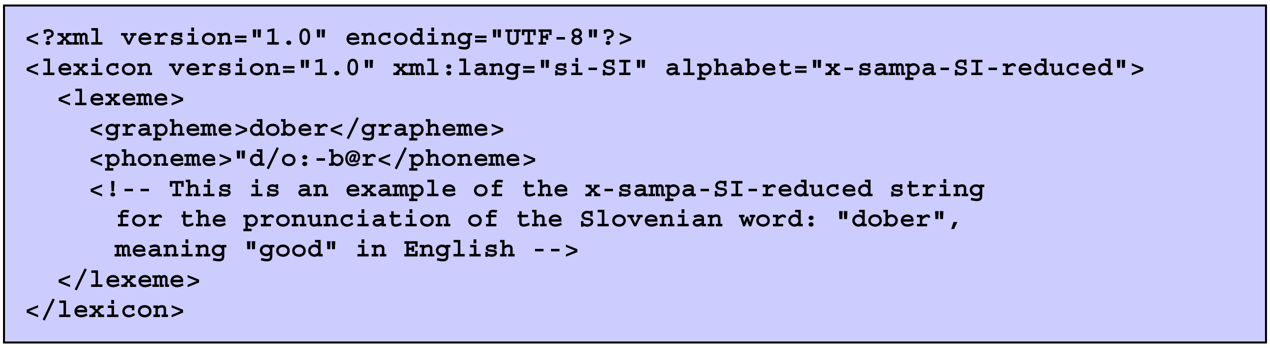
The element `<lexeme>` represents a lexical entry and may include multiple orthographies and multiple pronunciation information. An example of a simple lexicon file with a single lexeme within SI-PRON would be as shown in Fig. 1.

In the Pronunciation Lexicon Specification, the pronunciation alphabet is specified by the `alphabet` attribute of the `<phoneme>` element. We are using the "x-sampa-SI-reduced" phonetic alphabet, a subset of the X-SAMPA set as defined for Slovenian (Zemljak et al., 2002), augmented with additional markers for slovenian lexical stress accents (acute, circumflex, and grave) and tonemic accents (tonemic acute and tonemic circumflex). Both primary and secondary stress positions are marked.

The `<alias>` element is used to provide the pronunciation of an acronym or an abbreviation in terms of an expanded orthographic representation, as shown in Fig. 2.

4.1. Homographs

Homographs or words with the same spelling but different pronunciations can be treated in two ways. If we do not want to distinguish between the two words then we can represent them as alternate pronunciations within the same `<lexeme>` element. In the opposite case, two different `<lexeme>` elements need to be used. In both cases the application using the lexicon will not be able to distinguish when to apply the first or the second transcription unless additional information, such as context-specific attributes or part-of-speech information is provided.



```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xml:lang="si-SI" alphabet="x-sampa-SI-reduced">
  <lexeme>
    <grapheme>dober</grapheme>
    <phoneme>"d/o:-b@r</phoneme>
    <!-- This is an example of the x-sampa-SI-reduced string
      for the pronunciation of the Slovenian word: "dober",
      meaning "good" in English -->
  </lexeme>
</lexicon>
```

Fig. 1. An example of a simple lexicon file with a single lexeme within SI-PRON.

4.2. Multiple Pronunciations for the Same Orthography

Providing multiple pronunciations for items sharing the same orthography and meaning is important for speech recognition lexicons because they provide information on variations of pronunciation within a language. Therefore, for many lexemes, words, and multi-word expressions, multiple standard pronunciations are specified, including those taking into account possible coarticulation effects at word boundaries. Multiple pronunciations are indicated by subsequent `<phoneme>` elements within one `<lexeme>` element.

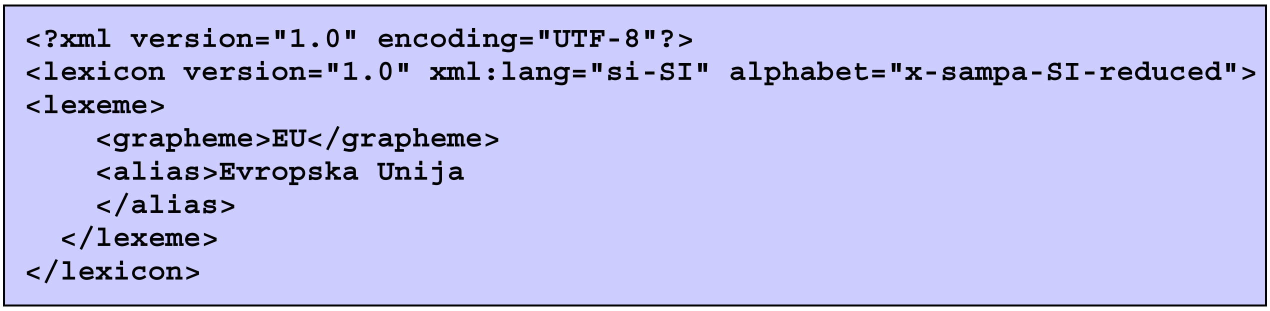
In text-to-speech synthesis applications, typically only one pronunciation among the multiple pronunciation possibilities is required. Therefore, to indicate preferred pronunciation variation, the `prefer` attribute is used in the latest version of the lexicon.

4.3. Homophones

Homophones, words having the same pronunciation, but with different meanings and orthographies, are not so frequent in Slovenian. Pronunciations are explicitly bound to one or more orthographies within a `<lexeme>` element so homophones are easy to handle.

4.4. Multiple Orthographies

Sometimes multiple orthographies of a word share the same meaning and pronunciation. They are presented with subsequent `<grapheme>` elements within a single `<lexeme>` element.



```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xml:lang="si-SI" alphabet="x-sampa-SI-reduced">
<lexeme>
    <grapheme>EU</grapheme>
    <alias>Evropska Unija
    </alias>
  </lexeme>
</lexicon>
```

Fig. 2. The `<alias>` element is used to provide the pronunciation of an acronym or an abbreviation in terms of an expanded orthographic representation.

4.5. Part-of-Speech Tags

The most recent specification of the PLS focuses on the major features described in the PLS requirements document. Many more complex features, such as those providing morphological, syntactic and semantic information associated with pronunciations, are expected to be introduced in a future revision of the PLS specification.

Therefore, proprietary `<lemma>` and `<morphsynt>` elements have been additionally defined for SI-PRON. Multext-East morphosyntactic descriptors for the slovenian language, as described in (Erjavec, 2004), were used to provide the part-of-speech information of the lexemes, along with the lemmas.

5. SI-PRON Validation

Finally, the SI-PRON lexicon has been subjected to an automatic validation as a way to ensure that the structure of the document is well-formed and conforms with the chosen Document Type Definition (DTD).

Additionally, manual validation of both phonemic transcriptions and morphosyntactic descriptions was performed on a subset of the lexicon comprising 5.000 lexical entries. A subset from the LC-STAR lexicon specifications for lexicon validation criteria was used (Shamas and den Heuvel, 2002).

A lexicon editing tool with a user-friendly interface has been designed to allow inspecting, editing, browsing and automatic validation of the pronunciation lexicon.

6. Conclusion

The design and structure of SI-PRON, a comprehensive machine-readable pronunciation lexicon for Slovenian, has been described.

The word list determination procedure, the generation and validation of phonetic transcriptions, and the lexicon format were presented. SI-PRON has been built from two sources and contains all the lemmas from the *Dictionary of Standard Slovenian* (*SSKJ*), the most frequent inflected word forms found in contemporary Slovenian texts, and a first pass of the inflected word forms derived from the SSKJ lemmas.

The lexicon file contains the orthography, corresponding pronunciations, lemmas and morphosyntactic descriptors of lexical entries in a format based on requirements defined by the W3C Voice Browser Activity. The current version of the SI-PRON pronunciation lexicon contains over 1.4 million lexical entries.

Along with Onomastica, SI-PRON presents a valuable language resource for linguistic studies and research and development of speech technologies for Slovenian. The lexicon is already being used by the AlpSynth Slovenian text-to-speech synthesis system (Žganec Gros, 2006) and for generating audio samples of the *SSKJ* word list, which are available at the very end of the *SSKJ* lexical entry descriptions (SSKJ audio, 2005).

7. Acknowledgements

This work has been funded in part by the Slovenian Research Agency and the Slovenian Ministry of Defense, under contracts no. L6-5405, no. V2-0896, and no. M2-0019.

8. References

Derlić, R., Kačič, Z., (1996). Definition of pronunciation dictionary of names and letter-to-sound rules for Slovene language - project Onomastica. In *Proceedings of the 2nd International Workshop on Speech dialog man-machine*, Maribor, Slovenia, June 26-27, pp. 153-158.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'04*, Lisbon, Portugal, pp. 1535-1538.

Fersøe, H., Hartikainen, E., van den Heuvel, H., Maltese G., Moreno A., Shammass S., Ziegenhain U. (2004). Creation and Validation of Large Lexica for Speech-to-Speech Translation Purposes. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'04*, Lisbon, Portugal.

Gros, J., Mihelič, F., (1999). Acquisition of an extensive rule set for Slovene grapheme-to-allophone transcription. In *Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH'99*, Budapest, Hungary, pp. 2075-2078.

Gros, J., Mihelič, F., Pavešić, N., Žganec, M., Mihelič, A., Knez, M., Merčun, A., Škerl, D., (2001). The phonectic SMS reader. In *Proceedings of the Text, speech and dialogue 4th international conference*, Železná Ruda, Czech Republic, Lecture notes in artificial intelligence, 2166. Berlin: Springer, pp. 334-340

Mihelič, F., Žganec Gros, J., Dobrišek, S., Žibert, J. and Pavešić, N., (2003). "Spoken language resources at LUKS of the University of Ljubljana", *International Journal on Speech Technologies*, Vol. 6, No. 3, pp. 221-232.

PLS-W3C, (2006). Pronunciation Lexicon Specification (PLS) Version 1.0, W3C Working Draft 31 January 2006. available from http://www.w3.org/TR/pronunciation-lexicon/S4.7.

Romary, L., Francopoulo, G., Monachini, M. and Salmon-Alt, S. (2006). Lexical Markup Framework: working to reach a consensual ISO standard on lexicons. To be presented at LREC'06 as a tutorial. Genoa, Italy.

SSKJ audio (2006). available from http://bos.zrc-sazu.si/sskj.html.

Verdonik, D., Rojc, M., Kačič, Z., Horvat, B., (2002). Zasnova in izgradnja oblikoslovnega in glasovnega slovarja za slovenski knjižni jezik. In *Zbornik konference Jezikovne tehnologije'02*. Editors: Tomaž Erjavec, Jerneja Gros, Ljubljana, Slovenia, pp. 44-48.

Verdonik, D., Rojc, M. and Kačič, Z., (2004). Creating Slovenian language resources for development of speech-to-speech translation components, In *Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC'04*. Lisbon, Portugal, pp. 1399-1402.

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.

SSKJ (1997). *Slovar slovenskega knjižnega jezika* (The Dictionary of Standard Slovenian). 2^{nd} edition, Ljubljana: DZS.

Šef, T., Gams, M., Škrjanc, M., (2002). Automatic lexical stress assignment of unknown words for highly inflected Slovenian language. In *Zbornik 11. mednarodne Elektrotehniške in računalniške konference ERK 2002*. Portorož, Slovenija., pp. 247-250. in Slovenian.

Toporišič, J. (1991). *Slovenska Slovenica* (Slovene Grammar). Založba Obzorja Maribor, (in Slovene)

Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P., (2002). Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija*, Vol. 50, No. 2, pp. 159-169.

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

Žganec Gros, J., (2006). Text-to-speech synthesis for embedded peech user interfaces, In *WSEAS Transactions on Communications,* No. 4, Vol. 5, pp. 543-548.