Considerations on using PLS for Slovenian Pronunciation Lexicon Construction

Jerneja Žganec Gros

Alpineon Research and Development
Alpineon d.o.o., Ulica Iga Grudna 15, SI-1000 Ljubljana, Slovenia
email: jerneja@alpineon.com, http://www.alpineon.com, http://www.voiceTRAN.com

1. Introduction

Consistent specification of word pronunciation is critical to the success of many speech technology applications. 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 (W3C PLS Version 1.0, 2006, Shamas & van den Heuvel, 2004; Fersøe et al., 2004).

Unlike as in other Slavic languages, 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 of crucial importance for the development of speech technology applications and linguistic research. Such a lexicon can be used either in its full-blown form or as 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 and 2004; 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); and none of them has used PLS.

Alpineon and the Fran Ramovš Institute of the Slovenian language from the Slovenian Academy of Arts and Sciences, have constructed 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 analysing contemporary slovenian text corpora.

2. The SI-PRON Lexicon

The procedure of the SI-PRON lexicon construction is described in detail in our forthcoming LREC paper (Žganec Gros et al., 2006b). Briefly, 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). The pronunciations of lexemes have been derived automatically for the *SSKJ* and *SSKJ* inflected word lists, 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.

The current version of SI-PRON contains over 1 million lexical entries. Along with Onomastica, SI-PRON presents a valuable language resource for development of speech technologies and Slovenian linguistic studies.

3. The SI-PRON Format

The pronunciation lexicon was constructed in form of a PLS document, which can be referenced from other markup languages, like SSML or SRGS.

The element **<lexeme>** represents a lexical entry and in SI-PRON only rarely includes information on multiple orthographies, but often on multiple pronunciations.

We are using the "x-sampa-SI-reduced" phonetic alphabet as the **alphabet** attribute of the **<phoneme>** element. This is a subset of the 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 acronym and abbreviation pronunciations.

3.1. Multiple Pronunciations

Providing multiple pronunciations for lexemes having the same meaning and orthography is important for speech recognition lexicons since they give information on variations of pronunciation in a given language.

Therefore, for many lexemes, words, and multi-word expressions, multiple standard pronunciations are specified in SI-PRON. Multiple pronunciations are represented by sequential **<phoneme>** elements within a single **<lexeme>** element.

Pronunciation preference – extensions needed ?

In TTS applications, typically only one pronunciation among the multiple pronunciation possibilities is required. Therefore, to indicate default pronunciation variation, the **prefer** attribute can be used in PLS. In SI-PRON, unless marked otherwise, the default pronunciation is the first pronunciation in *SSKJ*.

However, sometimes several pronunciation variations in SSKJ are (almost) equally preferred, whereas the actual

preferred pronunciation for the TTS engine may depend on the application. This is not to be confused with application-specific pronunciations, which can be handled in separate application-specific pronunciation lexica. What we have in mind is that there may exist several almost equally preferred pronunciations for a given grapheme, and the developers would like to have a mechanism that would enable them to systematically choose the preferred one.

Typically one of the two almost equally preferred pronunciations yields better rendering of input text if the application requires either overarticulated or fluent pronunciation. Therefore, we would welcome a new optional attribute to the **<phoneme>** element in PLS, the:

pron-style attribute indicating the preferred pronunciation variation of a lexeme with respect to the desired pronunciation style. The two attribute values, which would be useful for SI-PRON, are: "fluent", "overarticulated".

In addition, the **pron-style** optional attribute would need to be introduced into SSML, as a defined attribute for the **<voice>**, **<speak>**, **<p>**, and **<s>** elements.

For the same elements in SSML: **<voice>**, **<speak>**, **<p>**, and **<s>**, another optional attribute, **emotion**, would be useful (e.g. for computer games, where emotion changes occur frequently).

Example :

For Slovenian male nouns, ending with a vowel followed by "ilec", *SSKJ* often gives one of the following single or multiple pronunciations for the ultimate syllable "iUts"/"ilts", "ilts"/"iUts", "ilts", or "iUts"; examples would be Slovenian words "nosilec", "krotilec", "darovalec", etc. Many other cases of such pronunciation variations are known for Slovenian and are marked in *SSKJ*.

Whenever there are two pronunciation variations in *SSKJ*, they typically account for an overarticulated (e.g. "ilts") or a more fluent (e.g. "iUts") pronunciation variation. The pronunciation order as indicated in *SSKJ* indicates a slight pronunciation preference in standard usage and should still be indicated by the **prefer** attribute. In order to enable high-quality TTS such pronunciation differentiations should be captured in the text rendering process.

This would avoid the confusion of having a multitude of TTS pronunciation lexicons with different variations of the default pronunciation as given by the **prefer** attribute. The multiple lexicons are impossible to edit synchronously, and the proposed approach would allow us to use one master pronunciation lexicon.

How to denote dialects ?

The current PLS is monolingual/monodialectal. It would be beneficial to have the possibility to denote dialects/(and even sociolects?). If we really want bring the applications towards the users, and have collected their user profiles, etc., it would be useful to have an increasing number of sociolect-dependent pronunciations stored in a master pronunciation lexicon. Another optional attribute in SSML would be needed for this purpose for the **<voice>**, **<speak>**, **<p>**, and **<s>** elements. This would allow for a much higher degree of personalization and provide a powerful tool for application developers. Remark: rfc3066-like identifiers could be used for indicating dialects.

How to denote a pronunciation source/creator ?

At the present version of the PLS, only the **<metadata>** element allows do describe the creator of the lexicon. It would be useful to know the source of multiple pronunciations, esp. if the PLS document was obtained by merging several PLS documents. Often some pronunciation sources/creators are more reliable than others, not to mention that often ASR pronunciations are derived automatically.

Therefore, an optional **pron-source** attribute indicating the pronunciation source would be welcome in the PLS **<phoneme>** element.

3.2. Part-of-Speech Tags

A common attribute of Slavic languages is that they have extensive and complex inflectional paradigms. Slovenian is no exception. For example, the declension – the inflectional paradigm for nouns (along with adjectives and some pronouns), has apart from the six cases, three genders, and two language-universal numbers (singular and plural), another dimension in the number system, the dual (like ancient Greek!). Dual is used in full in Slovenian, not only as a remnant, and it is one of the most distinctive features of the language.

A derivational scheme/paradigm for providing prefix and suffix morphological rules (with indications of lexical stress position shifts) would enable construction of more compact lexicons for Slavic languages.

The most recent specification of the PLS focuses on the major features described in the current PLS requirements document, Ver. 1.0. Additional features, such as those providing morphological, syntactic and semantic information associated with pronunciations, would be welcome in future revisions of the PLS specification, as POS information is often crucial for assigning proper pronunciation to an input token. These aspects may be tackled in the next version of the PLS specification.

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

4. Conclusion

Due to arbitrary lexical stress position, pronunciation lexica are of crucial importance for development of speech technology applications and linguistic research for Slovenian. They are not only used for providing application-specific pronunciations or pronunciations of names, but are indispensable in any TTS or ASR system.

The task of constructing a master pronunciation lexicon is very tedious and time-consuming and should not be often repeated. Therefore, a master-lexicon approach is best suited for Slovenian TTS, in which many speaking-style pronunciation nuances are captured. Therefore we propose refined extensions to both PLS and SSML, which are described in sections 3.1. and 3.2. and mainly deal with multiple pronunciations, descriptions of pronunciation creators and morphosyntactic descriptions.

Reference list

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, 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, 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.

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, HU, 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", *Int. 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). 2nd 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., (2006a). Text-to-speech synthesis for embedded peech user interfaces, In *WSEAS Transactions on Communications,* No. 4, Vol. 5, pp. 543-548.

Žganec Gros, J., Cvetko-Orešnik, V., Jakopin, P., Mihelič, A., (2006b). SI-PRON: a Pronunciation Lexicon for Slovenian, *Proceedings of the Fifth International Conference on Language Resources and Evaluation LREC'06*, Genova, Italy.