A Computational Platform for development of Morphologic and phonetic lexica

Matej Rojc, Zdravko Kačič

Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova 17, 2000 Maribor
matej.rojc@uni-mb.si, kacic@uni-mb.si

Abstract:

Statistic approaches in speech technology, either based on statistical language models, trees, hidden Markov models or neural networks, represent the driving forces for the creation of language resources (LR), e.g. text corpora, pronunciation lexica and speech databases. This paper presents the system architecture for rapid construction of morphologic and phonetic lexica for Slovenian language. The integrated graphic user interface focuses in morphologic and phonetic aspects of the Slovenian language and allows the experts good performance in analysis time.

1. Introduction

During this decade an important effort can be identified in developing lexical knowledge databases, including different linguistic knowledge types as syntactic, morphological, phonological, semantic and others. Every natural language processing system needs to manage linguistic knowledge and high volume of lexical data, as well as incorporate efficient computational techniques to perform linguistic analysis. Practical NLP applications require large lexica resources, but as their construction is very time consuming and on the other hand since the development of natural language processing systems must be quick and efficient, the appropriate tools for rapid resources development is needed. In this paper the morphologic and grapheme-to-phoneme conversion tool with graphical interface that allows the construction of the two lexica as rapidly as possible, was created.

The tool allows experts to include, revise and validate the lexical knowledge, with explanation capabilities and linguistic knowledge editing, while achieving good performance in analysis time. The linguistic knowledge included in the graphical user interface focuses in morphologic and phonetic aspects of the Slovenian language. The emphasis of our tool was on making this process as productive as possible.

First the definition of needed text resources for lexica build-up will be given and data preparation step described in more details. Then the architecture of the tool including all modules will be presented and more detailed description of all modules given. The basic capabilities of the system will be discussed and used notation described. Also performance evaluation of the tool will be given and at the end conclusion will be drawn.

2. Text corpora

Nowadays a lot of text corpora resources are available in electronic form (e.g. Internet, CD-ROMs). In our work most of the needed text corpora to be processed for morphology and phonetic lexicon build-up were available on CD-ROMs with newspaper articles or were downloaded from the Internet in various formats (also texts from the literature were available). The obtained text corpus consisted of about 31 million words. After conversion into the text format, the tokenization of text into word tokens was performed to obtain the list of root items for build-up process of morphologic and phonetic lexica.

2.1. Tokenization and word selection process

Before using tool for building morphology and phonetic lexicon for Slovenian language, some text pre-processing on the obtained text corpus has to be done. According to the general, unrestricted nature of the text those algorithms have to be highly flexible and robust. The input text corpus (raw ASCII text) is fed to a tokenizer (finite-state machine) (Rojc,1999), which emits hypotheses about tokens and segments the input text into words.

The tokenizer engine is multilevel organised. At the lowest level the lexical scanner separate the input text into tokens. Some tokens may not be in dictionary form, appropriate for building up of morphology and phonetic lexicon. In this case the text normalisation processing level breaks such tokens into its constituent words.

All tokens like date, hour, cardinal and ordinal numbers are expanded into corresponding word forms during tokenization process at the 'expand text processing level'. The obtained words were sorted and word frequency was assigned to each one. Final list of items was defined using the 30.000 most frequent words in the input corpus (root forms).

3. System architecture for building up morphologic and phonetic lexica

In the figure 1 the system for lexica development is shown. From the figure we can see, that the system architecture consists of basically two levels. First the data preparation step is performed, followed by the lexica build-up level. The data-preparation step was already described above.

The lexica build-up level consists of three modules: rule-based linguistic module, automatic grapheme-to-

phoneme conversion module and graphic interface of the tool, which link all three modules together. Below the more detailed description of the modules will be given.

3.1. Automatic grapheme-to-phoneme conversion module

For speech recognition and text-to-speech synthesis the words in their grapheme representation have to be mapped onto their phonetic representation i.e. into their pronunciation. To be able to build grapheme-phoneme models, the availability of pronunciation lexica in their canonical form (a single phonemic representation for each word) is very important. In some languages this form can be easily derived from the grapheme form by a set of pronunciation rules, but for many languages (e.g. for Slovenian language) this relationship is too complex and is nowadays usually handled by manually made pronunciation lexica.

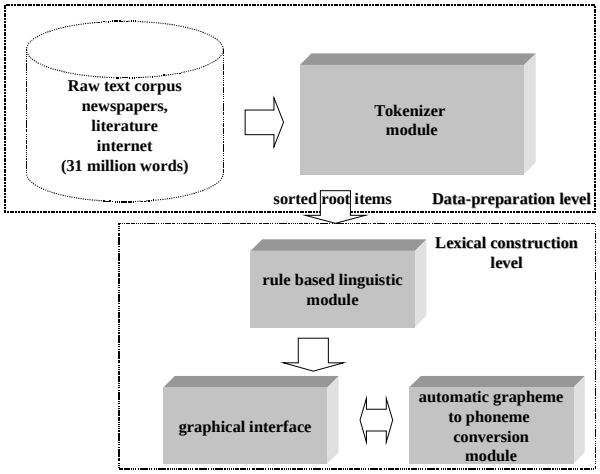

Figure 1. System architecture

To make it easier for linguistic expert developing the pronunciation lexicon, the automatic grapheme-to-phoneme conversion module was added. It consists of two parts. The first part uses rule-based approach and the second is based on data-driven approach, using neural networks.

The first part is intended for the use at the beginning of phonetic lexicon construction, when there is no material for learning neural network. First rule based stress assignment is done, followed by grapheme to phoneme conversion procedure. Then also rule-based syllabification on obtained phonetic transcriptions is performed. The result is verified and if needed corrected by the linguistic expert. After generated phonetic transcriptions are corrected, they are added to the phonetic lexicon.

As soon as we have enough data in the phonetic lexicon generated in the process described above, the data-driven approach, using neural network, can be used. The neural network which was taken for the basis of this part, is based on a method used and described in the Stuttgarter Neuronaler Netz Simulator SNNS (Zell,1994), which provides different training methods for a variety of applications. The data preparation, the generation of the training patterns and the training of neural network are done completely automatically. The transcription is performed in two steps. In the first, the graphemes are converted into phonemes and the syllable breaks inserted in the phoneme string. In second stress marks are inserted. The problem how to perform mapping between graphemes and phonemes for generation of training patterns for neural network, was solved automatically as proposed in (Hain,1999). The neural networks are learned off-line constantly during lexica development and then integrated into the grapheme-to-phoneme conversion module to increase its performance.

For both neural networks we used a multilayer perceptron (MLP) feedforward network with one hidden layer. Backpropagation algorithm was chosen as the learning algorithm for both networks.

The entry as it is written in the phonetic lexicon is:

/afrika
/a: – f r i – k a

af*er
a – f *E r

ag\ent
a – g \E n t

The pronunciation is derived from the IPA-Alphabet. In order to represent the IPA symbols in ASCII characters the SAMPA format is widely used. In the phonetic lexicon the phonetic transcription is written in SAMPA symbols for Slovenian language (SAMPA,1996). In the above entries '-' symbol marks syllable breaks and symbols '*', '/' and '\' are stress marks, since in Slovenian language three different type of stress exist (long and narrow, long and wide, short and wide).

3.2. Rule based linguistic representation component

Slovenian language is like other Slavic languages inflectional language and the linguistic representation for word depends on complex contextual factors. In this component most general linguistic rules were integrated. Since there are quite a lot of exceptions in Slovenian language, the linguistic expert verification has to be done after automatic generation of all forms for current item's linguistic category (POS – part of speech). All items that have to be processed are in their root form. As result all possible forms of the corresponding item, assigned with detailed linguistic information are defined. When the linguistic expert verified the obtained results, they all are added to the existing morphology lexicon in the prescribed format.

3.3. Graphic interface

Figure 3 shows the graphic interface of the system. The interface visualises all the information generated with the rule based linguistic module and automatic grapheme-to-phoneme conversion module. Through it the expert also performs editing, correction, and verification actions.

4. The system implementation

The architecture of the system is very modular and is multilingual oriented. The needed changes for the use in other languages are minimal - the appropriate system modules must be replaced. The tokenization module is already multilingual as such (Rojc,1999), and also statistical part of grapheme-to-phoneme conversion module is multilingual, since it uses data-driven approach based on neural networks. The graphic user interface currently supports only Slovenian language, but in the future development of multilingual architecture is possible.

All modules are written in C++ program language, except graphic interface, that is written in java language using Visual J++. The whole system runs on all Windows platforms –Windows 95/98 and Windows NT.

5. Working with the system

The morphology analysis represents the main part of the system. The graphic interface consists of three panels. The first panel is fixed and is intended as a starting point of the analysis, that the expert has to perform.

The linguistic expert has to load the already created phonetic and morphology lexica or he can create new one. As input he must load also list of all items, to which phonetic and morphology information has to be assigned. Then the expert manually chooses the item from the list. If the item is not in the root form, he must correct it. In our case all items in the list were already in root form and alphabetically ordered.

Next the expert verifies the type and position of stress and marks suffix in the item if exists. He further chooses the syntactic category for corresponding item (part-of-speech): noun, verb, adjective, number, pronoun, adverb, conjunction, interjection, article, and predicative. This action opens the appropriate second and third panel for corresponding part-of-speech.

According to this selection, the rule-based linguistic representation component generates and fills all attributes and values into their fields. The linguistic expert than verifies all values and correct them if needed. Below are defined some basic actions, that has to be performed for specific part-of-speech category:

- adjective: the comparison is automatically performed and conjugation/declension panel has to be activated,
- nouns: appropriate declension must be chosen and conjugation/ declension panel is activated,
- verbs: many panels for building various verb forms has to be activated and verified after automatic generation (infinitive, supine, participle, verb conjugation etc.),
- number: the expert determine gender, number etc. for root form and activates conjugation/declension panel,
- pronoun: for root form its type, gender, person, number and case is determined, and also conjugation/ declension panel is activated,
- adverb: comparison is automatically performed and type is chosen,
- conjunction: the type is determined,
- interjection: the type is determined,
- article: the type is determined,
- predicative: the type and case are determined.

When the expert verifies the content of the second panel, he has to save the information into the memory pushing 'insert' button. Then he moves to the third panel, where automatically generated conjugation/declension forms of the root item are performed (adjective, nouns, verbs, and numbers).

In Slovenian language there are quite a lot of exceptions, which can not be always correctly interpreted by the rule-based linguistic representation component. Sometimes also happens that the stress change the position and type in the word during conjugation/declension process. Since this is very hard to predict using rules, manual correction of the expert is needed.

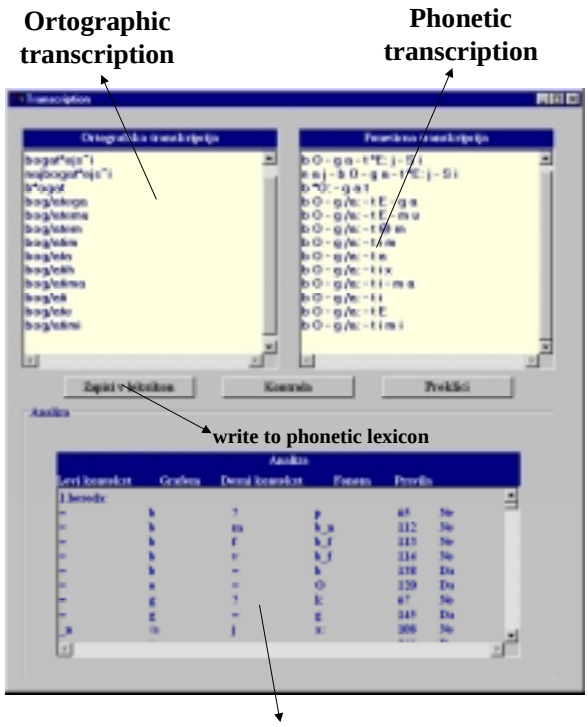

Figure 2. Grapheme-to-phoneme conversion window

The graphic interface tries to make this correction as easier as possible. The exceptions, that exist only in particular cases of corresponding declension, can be resolved by choosing between different available templates, which are defined in the linguistic representation component. Automatic generation process generates the most probable one. The verified information is then saved into the memory using again 'insert' button on the corresponding panel.

After that, the expert returns to the first panel, where he is able to inspect the saved information in the main

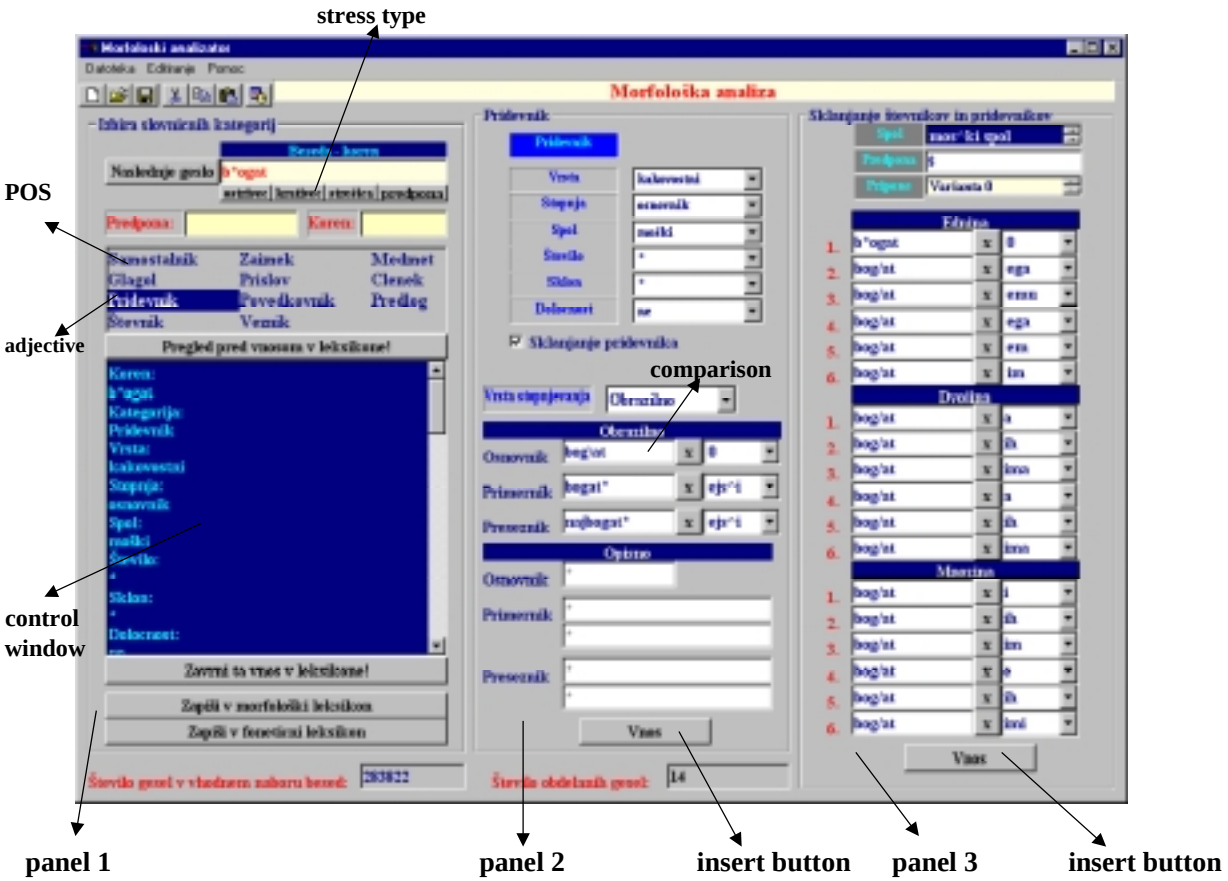

Figure 3. Graphic interface of the system

window. If some mistakes still exist, they can be corrected here, before the information is written into the morphology lexicon.

To verify automatically generated phonetic transcription, the expert has to open window, which visualises generated phonetic transcription, done with automatic grapheme to phoneme conversion module. The window is shown on Figure 2. The user is actually able to switch between using rule-based or data-driven approach (using neural network). Since a lot of inflectional forms for corresponding item are the same, only different orthographic words are send to the grapheme-to-phoneme conversion module.

The grapheme-to-phoneme transcription module gets as input only non-duplicated words and returns the corresponding phonetic transcriptions with syllable break marks '-' and stress marks.

The linguistic expert verifies the results and writes everything into the phonetic lexicon. Below there is also control window for inspection, which rules were used to identify the problem in case of errors in the obtained phonetic transcriptions. This is possible if rule-based approach is used.

6. Notation and attribute/value tables

The notation format of the output of the system for building morphology and phonetic lexica, consists of linear strings of characters representing the morphosyntactic information to be associated with word forms. We constructed the string following the philosophy of the Intermediate Format proposed in the EAGLES Corpus proposal (Leech and Wilson, 1994). It consists of agreed symbols in predefined and fixed positions. The categories used with the relevant attributes and values are based on EAGLES documents (MULTEXT,1996).

7. Performance evaluation of the system

Currently six linguistic experts work with the system, in order to build Slovenian morphology and phonetic lexica. After more than half a year of intensive work, the system was evaluated as a very efficient help for the expert. The linguistic expert finds it very easy for use, accurate enough in automatic generation of linguistic descriptions for items and also in grapheme to phoneme transcriptions. In case of errors, they can be corrected fast and easy without extensive typing, but mostly using mouse. They are able to verify approximately 100 root items in 15 hours. Since the Slovenian language is very inflectional language, in average 30 inflectional forms (during analysis of verbs most of the inflectional forms are generated) per root item are generated. That means that we get in average 3000 inflectional forms for 100 root items. Since a lot of duplicated inflectional forms are obtained during conjugation/declension, the phonetic

lexicon is in average ten times smaller than the morphologic lexicon.

8. Conclusion

Currently six linguistic experts work very intensively with the developed system. The morphologic lexicon currently contains more than 400.000 items (root items plus corresponding inflectional forms) and about ten times smaller phonetic lexicon (49.000). The alphabetic coverage will be achieved in the next few months, since every expert works on root items from before determined alphabetic letter. It is planned to use the Slovenian phonetic lexicon in research work in the field of automatic continuous speech recognition of Slovenian language. Both phonetic and morphology lexicon will be used for Slovenian text-to-speech synthesis.

9. References

Matej Rojc, Janez Stergar, Ralph Wilhelm, Horst-Udo Hain, Martin Holzapfel, Bogomir Horvar, (1999) A Multilingual text processing Engine for Text-To-Speech Synthesis system, Proceedings EUROSPEECH 1999, Budapest, pp. 2107-2110

Zell, A.(1994). Simulation Neuronaler Netze. Bonn, Paris; Reading, Mass., Addison-Wesley

Horst-Udo Hain, (1999) Automation of the training procedure for neural networks performing multi-lingual grapheme to phoneme conversion, Proceedings EUROSPEECH 1999, Budapest, pp. 2087-2090

SAMPA for Slovenian, (1998)
http://www.phon.ucl.ac.uk/home/sampa/sloven-uni.html

MULTEXT project lexical specifications, (1996)
http://www.lpl.univaix.fr/projects/multext/LEX/LEX.Specifications.html