International Conference
"ICT for Language Learning" 4th edition
PIXEL

Computer System for Learning Word Inflection in Slavic Languages. Case Study - Slovenian Nouns and Verbs

Jure Zupan
National Institute of Chemistry (Slovenia)
jure.zupan@ki.si

Abstract

In the paper an interactive computer system for learning Slovenian inflection is described and discussed. The system can be used in two ways: a) for finding all possible inflected forms of any noun, verb, or adjective; and b) to generate the proper lexeme from any given grammatically correct morphem that can appear in any text or spoken communication. The system is independent on the language data base. However, to operate properly in any specific language (written in Latin alphabet) the system must be provided with an adequate language dictionary of word 'roots' and the file of all possible inflection ending sets. In the present case the system has a list of about 100,000 'roots' and 240 sets of Slovenian inflection endings. For teaching the beginners or for the demonstration purpose an extensive dictionary of word 'roots' is not mandatory – a file of about one thousand words is sufficient. However, the corresponding inflection endings are necessary for the system to work properly. From the initial state the system can be easily upgraded by increasing the dictionary.

1.Introduction

Slavic languages are known for their inflection richness. In these languages the declension endings of nouns express gender (3), number (2-3), case (6-7), and animacy (yes/no). The conjugation endings of verbs express person (3), number (2-3), tense (2-4), aspect, mood, and voice. The adjectives are inflected for number, gender, and case according to the noun to which they are associated with. It is understandable that for a non-native speaker it is quite difficult to memorize and to use properly all possible variations that can appear in written or spoken sentences. In order to ease the learning and teaching of Slavic languages an interactive computer system for self-teaching and grammatical checking of declension and conjugation of nouns, verbs, and adjectives was designed. Even if the problem of the correct choice of endings for the spoken words is neglected, the problem of identification of the correct word (lemma) when writing or reading a foreign text still remains. Due to the fact that endings can be several characters long, a beginner may easily have problems to find the proper word which can be identified through the dictionary (the lemma). For example, the beginner is looking for the infinitive of the word *potujeva* (first person/dual/present tense of the verb *pot-ovati* (*Engl. to travel*). Even if the ending *-ujeva* is identified as such, there are 316 words in the Slovenian dictionary starting with the root *pot-.* If the beginner is more advanced and knows that the ending -ujeva is associated with verbs, still 112 verbs remain in the dictionary among the mentioned 316, to chose from. The root *pot-* itself has two meanings as an female noun *pot* (Engl. *way*) or male noun *pot* (*Engl. sweat*). The set of endings containing *–ujeva* is not a rare one. On the contrary, there are literary hundreds of verbs using this particular set of 21 endings (set No. 340, Figure 1, right). Taking into account that in Slovenian language there are 89 sets of verb endings, 122 sets of endings for masculine, feminine and neuter nouns, and 21 sets for adjectives [1,2] it became clear that an on-line help for learning inflection in Slavic languages is a very welcome tool to either students or teachers.

2. The system

The goal of the interactive program is to display all syntax possibilities of words or to parse different word syntaxes into proper lexical form (lemma). Input to the program is always one word which can be in any grammatically correct form. Immediately, after the query word is input via the pop-up window, the system displays another window with all possible inflections of the query. The resulting window contains the complete information about all possible syntaxes either for a noun, verb, or adjective, depending on the query. An examples of the output windows for a verb query is shown in Tables 1. In order to be understandable to English speaking readers, the comments in the output window associated with the endings of conjugation are translated from Slovenian, but either the root of the query word or the endings (shown in bold) are not. This means that Table 1 is not exactly the same as displayed by the computer, but the information given in the table is.

After the query word *delamo* (Engl. *we work*) is typed in the set of all 21 possible different forms of the verb's conjugation syntaxes is displayed. In lines 1 to 10 are 'root-endings' for formation of the present tense. The syntax of the third person in plural in the present tense (lines 9 and 10) can have two different forms (the verb *delati* has only one, line 9). The past and future tense endings (lines 11 to 15) are the same for both tenses, the difference is expressed by specific forms of the auxiliary verb (left column). In lines 17 to 20 are 'root-endings' for the formation of the imperative for different combinations of persons and numbers. In line 21 is ending of the passive participle. Asterisk marks the sought syntax form of the query word: 1st person, plural, present tense. The computer displays for nouns and adjectives are similar, giving always all possible syntax forms and corresponding endings.

Table 1. All possible conjugation forms of the verb *delati* (Engl. *to work*) as they are structured by the computer. The proper syntax of the query word *delamo* (Engl. *we work*; first person, plural, present tense) is marked with an asterisk.

Ending	Query word:	**delamo**	**delati,** verb	Ending set, No. 302
1	infinitive	**dela-ti**	to work	
2	**jaz**	**dela-m**	I work	
3	**ti**	**dela-š**	you work (singular)	
4	**on/ona/ono**	**dela-**	he/she/it works (singular)	
5	**midva**	**dela-va**	two of us work (dual)	
6	**vidva**	**dela-ta**	two of you work (dual)	
7	**mi**	**dela-mo**	* we work (plural)	
8	**vi**	**dela-te**	you work (plural)	
9	**oni**	**dela-jo**	they work (plural)	
10	**oni (druga oblika)**	**-**	they work (alternative form)	
11	**jaz /ti/on je/bo**	**dela-l**	past/future tense, masculine, singular	
12	**one so/one bodo**	**delal-e**	past/future tense, feminine, plural	
13	**ono je/ono bodo**	**delal-o**	past/future tense, neuter, singular	
14	**ona je/so /ona bo/bodo**	**delal-a**	past/future tense, several combinations possible	
15	**oni so / oni bodo**	**delal-i**	past/future tense, several combinations possible	
16	**midva/medve/midve**	**dela-jva**	imperative for us, dual	
17	**mi/me/me!**	**dela-jmo**	imperative for us, plural	
18	**ti!**	**dela-j**	imperative for you, singular	
19	**vidva/vedve/vidve!**	**dela-jta**	imperative for you, dual	
20	**vi/ve/ve!**	**dela-jte**	imperative for you, plural	
21	**trpni deležnik**	**dela-t**	passive participle	

3. Databases and methods

In general, most of the computerized linguistics work of parsing plain text is made with statistical methods [3], i.e., by comparing unknown phrases and words with already tagged corpuses of text. On the contrary to this approach, our algorithm is strictly grammatically driven. The program needs endings for all possible syntaxes and a data base of word 'roots'. The present version of the program is operating on endings and 'roots' of Slovenian language [1,2]. The 'root' is not a dictionary's entry nor the semantic origin of the word, but the part of the dictionary's entry (lemma) that is *not* changed by any inflected form among all possible grammatical syntaxes. Some 'roots' (masculine nouns and many adjectives) are identical to the lemmas, but even more can be up to six letters *shorter* than the dictionary's entry (feminine and neuter nouns and all verbs). The 'roots' are stored in the database in a random accessible way (Fig. 1, left).

Each 'root' can be associated with several lists of endings, not necessarily of the same type. For example the 'root' *kos-* is associated with six ending lists (Fig. 1). They are the endings for masculine (list endings No. 1 and 2) and feminine nouns (list No. 51), for adjectives (list No. 201), and for verb (list No. 362). Fig. 1. shows that the roots *kos-* and *pot-* are associated with 6 endings each. In the present (Slovenian) database some of the 'roots' are associated with as much as ten different ending lists. Lists of endings are exhaustive lists of all endings that can be added to nouns, verbs or adjectives to form all grammatically correct syntax forms. Each word type noun, verb, and adjective has different number of endings 18, 21, and 12, respectively. Specific ending be of the noun, verb, or adjective defining a specific syntax is always at the same position in any list of that type. For example, the ending for imperative for second person in dual is always on the 19th position of the 21-item list of verb endings (Table 1). The position of ending defines the proper grammatical attributes. Different list of endings can be very similar. For example, the lists No. 1 and No. 2. (Fig. 1, center) differ only in the fourth one (4th fall/singular/male nouns) signaling the animacy of the object (living – list No. 1 and nonliving – list No. 2). These two lists (No. 1 and No. 2) contain the declension endings of two identical lemmas *kos* (Engl. *blackbird*) and *kos* (Engl. *piece*) with two different meanings which are represented by the same 'root' (*kos-*).

Due to the fact that the program allows to enter any grammatically correct word it may well happen that a given grammatically correct query can be derived from two or even more different lemmas and, consequently, have two or more completely different meanings. For example: the word *kosi*, can be derived from six (6) lexical entries (lemmas) (see Fig. 1):

a) or b) as a first fall of plural from two masculine nouns *kos* (Engl. *Blackbird*), ending list No. 1 or *kos* (Engl. *piece*), ending list No. 2
c) as a first fall of dual from the feminine noun *kosa* (Engl. *scythe*), ending list No. 51,
d) or e) as the third person of the present tense from two verbs *kosíti* (Engl. *to mow*), or *kósiti* (Engl. *to take lunch*), both with ending list No. 362, and finally,
f) from the adjective *kôs* (Engl. *oblique*), ending list No. 201. For this lemma the form *kôsi* can be used in seven (7) different syntax combinations of gender, number, and fall between the adjective *kôs* and any noun. These combinations are: masculine/plural/1st fall; feminine/singular/3rd or 5th fall; feminine/dual/1st or 4th fall; neuter/dual/1st or 4th fall).

One reason for such a variety of possibilities is the fact that in many languages the stress on the accentuated syllable is not marked. Although there are six different lemmas involved in the above example, there are only five different sets of endings used for producing the six results shown. The list of endings of the two verbs is the same in both cases. This is, of course, an extreme case, but by no means a very rare one. It shows the complexity of the Slavic language syntax possibilities clearly. At the same time it demonstrates why a program explaining all cases in such details, are very important for students and teachers of highly inflected languages.

The procedure that finds the correct grammatical syntax form is a two step process. In the first step which runs in cycles all possible 'roots' for a given query are found. In each cycle the root which is sought is one letter shorter than in a cycle before. In the first cycle the program searches for the 'root' identical to the query word. To explain this procedure more in detail, lets us take the six letter long query word *delamo* from Table 1 as an example. In this case, the 'roots' sought in the database are: *delamo-*, *delam-*, *dela-*, *del-*, *de-*, and *d-*. Three of them (*dela-*, *del-*, and *de-*) are associated with ending lists (Fig.1, left) and three are not (*delamo-*, *delam-*, and *de-*).

'Roots'	Identifications of ending lists					
...						
...						
...						
dela-	302					
...						
ljubez-	78					
...						
kos-	1	2	12	51	201	362
...						
del-	2	11	101	209	340	
...						
pot-	2	72	209	340	359	362
...						
de-	325					
...						
...						

No 1	No 2	No 12	No 51	No 78
-	-	-ec	-a	-en
-a	-a	-ca	-e	-ni
-u	-u	-cu	-i	-ni
-a	-	-ca	-o	-en
-u	-u	-cu	-i	-ni
-om	-om	-cem	-o	-nijo
-a	-a	-ca	-i	-ni
-ov	-ov	-cev	-	-ni
-oma	-oma	-cema	-ama	-nima
-a	-a	-ca	-i	-ni
-ih	-ih	-cih	-ah	-nih
-oma	-oma	-cema	-ama	-nima
-ih	-ih	-cih	-e	-ni
-ov	-ov	-cev	-	-ni
-om	-om	-cem	-ama	-nim
-e	-e	-ce	-e	-ni
-ih	-ih	-cih	-ah	-nih
-i	-i	-ci	-ami	-nimi

No 201	No 302	No 340	No 362
-	-ti	-ovati	-iti
-a	-m	-ujem	-im
-e	-š	-uješ	-iš
-i	-	-uje	-i
-o	-va	-ujeva	-iva
-em	-ta	-ujeta	-ita
-im	-mo	-ujemo	-imo
-ih	-te	-ujete	-ite
-ega	-jo	-ujejo	-ijo
-eu	-	-	-e
-ima	-l	-oval	-il
-imi	-la	-ovala	-ila
	-le	-ovalo	-ile
	-li	-ovale	-ili
	-lo	-ovali	-ilo
	-j	-uj	-i
	-jva	-ujva	-iva
	-jta	-ujta	-ita
	.jmo	-ujmo	-imo
	-jte	-ujte	-ite
	-t	-ovati	-it

Fig.1. Random access database (left) contains the 'roots' and the numbers of lists of endings. Nine lists of endings (right) are shown as example of the 240 existing ones. Lists for noun, adjective, and verb endings are numbered as follows: 1 – 200, 201 - 299, and 301 - 399, respectively.

In the second step the 'roots' associated with the lists of endings are combined with *all* endings of *all* lists they are associated with. The 'root' *dela-* is combined with all 21 endings of the list No. 302, the 'root' *del-* with all endings of five lists No: 2, 11, 101, 209, and 340, and finally, the ending *de-* with all endings of list No. 325. If the combination of the 'root-ending' is found to be identical to the query word the correct syntax is established. In the discussed case, only the root *dela-* together with the seventh ending in the list No. 302 (*–mo*) yields the correct query word *dela-mo*. From the list number and the position of the ending in the list the complete grammatical information is deduced. The seventh position of the ending *–mo*, signals the 1st person of the plural in the present tense, while the sequential number of the endings list (No. 302) reveals the combination 'root-ending-No.302' is a verb.

4. Conclusion

The main feature of the described system is its independence from the user's knowledge about the word's type that he or she wants to investigate. Additionally, the system is not based on a tagged corpus of text, which are often copyrighted. The system's 'knowledge' depends of what the user (teacher) is prepared to input into it: the number and type of the words to be thought and the extension of the grammar (i.e., the number of different endings lists) the user is prepared to collect and input. Of course, the system can be gradually expanded. Besides being an educational tool to help students and teachers to master the complexity of declension of some Slavic languages, it has an additional ambition to become a part of general program for parsing the text on-line through grammatical rules [4].

In principle the system is designed for Slavic languages. In order to ensure the general functioning of the same program shell for displaying the described word analysis for another language two different databases should be formed. Because of the possession of the source code of the program it can be easily reprogrammed to comply with any specific needs to other languages (written in Latin alphabet). Unfortunately, to collect all inflection endings in highly inflected language is very specific and not an easy task. If a spell-checking program for that particular language already exist and is available, its computer-readable dictionary can be implemented for such a purpose, or alternatively, the endings can be extracted. There are several such possibilities for Slovenian Language [5,6]. To compose a dictionary of roots is not so difficult, but is rather time consuming. Usually, a common dictionary contains between 100,000 to 200,000 words [7], what is comparably large amount for use as a source of 'roots', hence, smaller dictionaries are better suited for the initialization of the program. Of course, any size random accessible root-database formatted as it is shown in Fig. 1, can be implemented and used in our system – once it exists. To be applicable on an individual basis, the described system is designed in such a way that the users (mostly the teachers) can update its dictionary slowly, step by step, according to the needs and extensiveness of teaching. Due to the fact that in most cases teaching of a new language starts with simplest and most general inflection forms it is not necessary that the applied database of 'roots' contains the entire dictionary of the word roots.

References

1. J. Zupan, *Problemi in nekaj rešitev računalniških obdelav slovenskih besedil*, (in Slovenian, English Abstract), Slavistična revije, 1999, 47(3), 277-296,
2. J. Zupan, The Application Of Artificial Neural Networks In Linguistics,Published in Clark, J. W., Lindenau, T., Ristig, M.L. (Editors). Scientific Applications Of Neural Nets, Lecture notes in physics, 522. Berlin, Springer, 1999, 224-242
3. J. Carroll, Statistical parsing, in R. Dale, H. Moisl, and H. Somers, Eds., Handbook of Natural Language Processing, Marcel Dekker, 2000, p. 525–543.

4. J. Zupan, New Concept for a Network Dictionary of Meanings in Slovenian Language, Eds. M-E. Ćavar, D. Ćavar, Book of Abstracts of the 4th Annual Meeting of the Slavic Linguistic Society, Zadar 2009, p. 112-113.
5. BesAna, v 2.03, Spell-checking Program for Slovenian Texts, Instruction Manual, Amebis, d.o.o., Kamnik, Slovenia 1993, and J. Zupan, SLONCEK Spell-check program for Slovenian Language (in Slovenian), Informatica 15, 1993, 3, 21-32,
6. J. Zupan, Extended version of the Root-Dictionary of Slovenian Language (mentioned in the Zupan's article in the newspaper DELO, 5-th February 1998, p.16.
7. Dictionary of Slovenian Literary Language, Slovenian Academy of Science and Art and Državna Založba Slovenije, Ljubljana, 1995 (available on CD-rom).