Original research article
UDK 811.163.41`366:811.163.41`282
DOI 10.21618/fil1919036d
COBISS.RS-ID 8214040

Strahinja B. Dimitrijević[1]
University of Banja Luka
Faculty of Philosophy
Laboratory of Experimental Psychology

THE ROLE OF PHONOTACTIC INFORMATION IN THE PROCESSING AND PRODUCTION OF INFLECTIONAL MORPHOLOGY [2]

Abstract: *In two separate studies, it was examined to what extent in languages with rich inflectional morphology, such as Serbian, in the tasks of automatic perception and word production we can rely on phonotactic information, i.e. permitted combinations of phonemes/graphemes. In the first study, with the help of support vector machines, word classes discrimination was carried out based on bigrams and trigrams generated at the morphosyntactic category level. In the second study, the production of inflected forms was explored, with the help of memory-based learning, relying on phonotactic information from the last four syllables of lemmas. Maximum discrimination accuracy of inflected word classes was obtained on the basis of bigrams reaching about 93% of accuracy. Similarly, in the task of inflectional production on all inflected word classes taken together, around 92% of the inflected forms were generated correctly. This confirms that phonotactic information in the perception and production of morphologically complex words plays an important role. Therefore, this information should be taken into account when considering the emergence of larger language units and patterns. The results show that functional connections between orthography/phonology and semantics lead to successful lexical learning, but at the same time they call into question the need for positing the existence of mental lexicon in which mental representations of various language characteristics are stored.*

[1] strahinja.dimitrijevic@ff.unibl.org
[2] This paper is a substantially modified part of the author's unpublished doctoral dissertation.

Keywords: *phonotactic information, inflectional morphology, word classes discrimination, memory-based language processing, support vector machines.*

1. Introduction

This paper investigates the possibility of automatic word classes discrimination and production of inflected forms in a morphologically complex language relying on phonotactic information (PI). PI refers to phonological segments and sequences of segments in words of a given language, with unequal probability of occurrence, traditionally considered within the categories of *legal* vs. *illegal* (Crystal, 2008; Jusczyk *et al*., 1994; Vitevitch and Luce, 2004). To illustrate, in Serbian, a sequence consisting of segments 'f' and 'l' – 'fl' is allowed while the sequence 'žz' is not allowed.

The importance of phonotactic probability (PP), i.e. the probability with which phonemes co-occur (Vitevitch and Luce, 1999) in language processing has been confirmed in a number of studies. Babies prefer pseudowords containing highly frequent phoneme sequences (Jusczyk *et al*., 1994), while children rely on PP when determining the boundaries between words and speech segmentation (Cairns *et al*., 1997; Mattys and Jusczyk, 2000; Mattys *et al*., 1999; Morgan and Saffran, 1995), as well as when learning new words (Storkel, 2001, 2004; Storkel and Morrisette, 2002; Storkel and Rogers, 2000). In adults, PP affects the success in repeating pseudowords (Vitevitch and Luce, 1998, Vitevitch *et al*., 1997), phoneme identification (Pitt and McQueen, 1998), pronunciation speed (Levelt and Wheeldon, 1994), the speed of recognising pronounced words (Luce and Large, 2001; Vitevitch, 2003; Vitevitch and Luce, 1998, 1999; Vitevitch *et al*., 2002), and estimating similarity between pseudowords and words (*wordlikeness*; Bailey and Hahn, 2001; Coleman and Pierrehumbert, 1997; Frisch *et al*., 2000; Vitevitch *et al*. 1997).

Apart from PP, the number of words that are phonemically similar to a given word, so-called neighbourhood density (ND) (Vitevitch and Luce, 1999) also plays a role in word processing. Although PP and ND correlate highly positively, it has been determined that there exist individual effects of each of these variables on word processing (van der Kleij *et al*. 2016). A common way to measure ND is to count, for a specific word, its neighbours in the lexicon that can be produced by deletion, addition or replacement of *n* of the original word's phonemes, where *n* specifies the number of operations (Levenshtein, 1966). Unlike the facilitating effect of PP (except in the case of adults learning new words), the ND effect can also be inhibiting (Storkel *et al*. 2006), and in studies this effect is often absent

(Balota *et al*. 2004). Whether ND facilitates or inhibits word proccessing depends on whether we are dealing with word recognition or production (Storkel *et al*. 2006), which task type we employ, such as lexical decision, task naming etc., but also on factors such as ortographic or phonological similarity (Grainger *et al*. 2005; Ziegler *et al*. 2003). The interaction between ND and word frequency was also determined (Andrews, 1989, 1992), which further complicates the unambiguous determination of the effects of ND on word processing.

Despite the importance of PI in language processing, research aimed at the design and testing of part-of-speech (word class) taggers, i.e. algorithms that automatically give additional information to linguistic elements in the language corpus (Baker *et al*. 2006), makes a sporadic and unsystematic use of PI, with the aim of increasing the accuracy of taggers. Although this research fits in the field of Computer Linguistics and Natural Language Processing, its aim is not to develop a computer system for automatic word processing, but to determine information potential of phonotactic sequences in various language processing tasks with rich inflectional morphology. The results will indicate to what extent our cognitive system can rely on PI in various tasks of understanding and producing morphologically complex words. As a matter of fact, computational approaches allow for quantification, generation, and testing of different hypotheses with the aim of offering insights into language processes. Likewise, they are particularly relevant within debates about the mechanisms underlying production and perception of morphologically complex words (Keuleers, 2018). It was shown that the use of cognitively acceptable algorithms, such as *Memory-based Learning* (Daelemans and van den Bosch, 2005) or *Naive Discriminative Learning* (Baayen *et al*. 2011), based on the principles of animal and human learning leads to better understanding of how linguistic knowledge arises from the exposure to language and its use (Milin *et al*. 2016).

The contribution of phonotactic information to morphology was tested in two tasks: the task of automatic word classes discrimination,[3] and the task of automatic production of inflected forms of words in Serbian. In the first study, the probability of phoneme sequences served as the basis for the classification of morphosyntactic category (MSC) (e.g. masculine noun in the nominative singular; the positive form of adjective of the feminine gender in the plural genitive; infinitive, etc.)

[3] In this paper, the terms WC discrimination and MSC are used interchangeably. Namely, these terms denote the same process, seen from two different angles: discrimination of WC based on PI specified at the level of MSC, or the classification of these MSC based on the calculated frequencies into the corresponding WC.

into the corresponding word classes (WC). The classification was performed by *Support Vector Machines* (SVM) (Vapnik, 1995, 1998). Even though SVM do not reflect a probabilistic cognitive model of language processing (Baayen, 2011), this statistic tool employs pairwise comparisons as a basis for counting (Vert *et al*., 2004). This means that it relies on the information about mutual similarity of input data, whereby similarity measure for two objects is represented by their scalar product. In this case, it referes to the vicinity between certain bigrams and trigrams. If it is shown that WC can be sucessfully discriminated based on bigrams and trigrams, which implies the existance of characteristic distribution for WC, then this would imply also the 'vicinity' of bigrams and trigrams at a cognitive level.

In the second study, automatic production of inflected forms (IF) was examined by means of memory-based learning (MBL) (Daelemans and van den Bosch, 2005). This approach was chosen due to the long tradition of accounts of language production in terms of mechanisms of analogy (see: De Saussure, 1916/1966; Bloomfield, 1933; Harris, 1951, 1957), i.e. with the use of new words within existing patterns based exlusively upon stored examples in memory (Baayen, 2003; Boas, 2003; Bybee, 2010; Bybee & Eddington, 2006; Eddington, 2000; Krott *et al*., 2001; Skousen, 1989, 1992; etc.). In most cases, language analogies are based on semantic and phonological similarity with the existing forms stored in memory (Bybee, 2010). There are different ways to formalise the mechanism of analogy. Yet, what they all have in common is that they do not rely on rules in the processing of morphologically complex words, even in case of irregular forms (Skousen; 2002). One such model is MBL (Daelemans and van den Bosch, 2005), which represents a computer implementation of the approach based on examples, whose main propositions reflect not only linguistic knowledge but also a possible cognitive architecture and processes present during language use (Baayen, 2011; Keuleers, 2008; Keuleers and Dealemans, 2007; Milin *et al*, 2011; Milin *et al*., 2016).

The results of this study will show to what extent our cognitive system can rely on phonotactic information in various comprehension and production tasks of morphologically complex words. In other words, if the PI is "sufficient" to perform tasks in the domain of morphology, this puts into question the need to postulate the existence of a mental lexicon in which the representations of morphemes and / or words are stored.

2. STUDY 1. The role of phonotactic information in WC discrimination

The first study examined WC discrimination based on PI. Bigrams, combinations of two (e.g. *pa*) or trigrams, combinations of three (e.g. *pap*) phonemes/graphemes were used.[4] The probability of bigrams and trigrams was calculated at the level of MSC (for example, a masculine noun in the nominative of the plural, etc.), and their position in words was systematically varied.

SVM (Vapnik, 1995, 1998) is a reliable, mathematically well-established and efficient method for classification (Baayen, 2011; Joachims, 1998; Meyer *et al*., 2003; Steinwart and Christmann, 2008; van Gestel *et al.,* 2004), which does not make any specific assumptions about the distribution of predictor variables (Steinwart and Christmann, 2008). This was a key precondition for our study, because the largest number of values for bigrams and trigrams was zero and their distributions significantly deviated from the normal one. Their distributions do not change significantly even when, among other things, small positive values were attributed to the zero values, in the process of smoothing. In addition to frequency correction, other data transformations to overcome the problem of empty matrices, such as the reduction of raw frequencies to categorical variables (e.g. present-absent bigram/trigram) would result in a significant loss of variance. As a consequence, a lower accuracy of the classification would follow. This is why data transformations are inadequate in this situation.

2.1. Method

Material: The material necessary for checking WC discrimination on the basis of PI was distributed in eight matrices. Four matrices with frequencies were formed for bigrams: (1) bigrams at word onset (e.g. bigram *#p*, where # denotes an empty field), (2) bigrams at word endings (e.g. *p#*); (3) bigrams at any position in a word (e.g. *pa*), (4) matrix encompassing the previous three kinds of information. Similar matrices were also formed for trigrams: one matrice contained information about trigram frequency at word onset (e.g. *#pa*); another contained information about trigram frequency at word ending (e.g. *pa#*), the third one contained trigrams at any position in a word (e.g. *pap*), wheras the forth one contained information from the previous three matrices. All the matrices contained 1293 rows in total, while the number of columns depended on the number of bigrams or trigrams. There

[4] Serbian is characterised by a shallow orthography, i.e. graphemic and phonological codes are isomorphic.

were MSC in the rows, for: nouns (74), adjectives (131), pronouns (594) and verbs (494). A fragment of frequency matrices for bigrams, regardless of their position in the word, is given in Table 1.

Tabela 1. Fragment of frequency matrices for bigrams, regardless of their position in a word

MSC code	...	ra	ke	je	no	an	re	ar	st	...
.....	...	...	...	...	...	...	...	...	...	...
100213	...	1136	4	174	441	391	1107	535	1496	...
100221	...	5101	30	336	794	1410	789	2768	1215	...
100222	...	934	5	216	237	964	505	445	750	...
100223	...	171	0	31	5	22	399	49	347	...
100231	...	16	2	2	0	14	1	5	0	...
100232	...	25	0	77	12	58	0	10	20	...
100233	...	16	0	0	0	1	0	0	6	...
.....	...	...	...	...	...	...	...	...	...	...

Note: The first number in a code refers to word type (1-noun), the fourth number refers to case features (2-genitive), the fifth number refers to number features (1-singular; 2-plural; 3-pluralia tantum), the sixth number refers to gender features.

To calculate the frequencies of bigrams and trigrams, a subsample of one million words from the daily press from *Korpus srpskog jezika* (*Corpus of Serbian Languge*; Kostić, 2001) was used. After calculating the frequencies for all the bigrams and trigrams, their number, apart from the bigrams at word onset and at the word ending, was reduced by keeping those whose average frequency for one MSC (the sum of frequencies/number of different MSC in which a given bigram or trigram was reported) was greater than 25 (Table 2). This value was determined in order to to preserve more input, while simultaneously reducing the fragmentation of the matrix (controlling the total number of empty cells). The choice of this value was based on a graphical estimate of the number of 'important' bigrams and trigrams, and the procedure was similar to Catell's *scree test* (1966).

Table 2. Number of bigrams and trigrams, before and after reduction

	bigrams				trigrams			
	#x	x#	xy	all the bigrams	#xy	xy#	xyz	all the trigrams
before reduction	30	30	672	732	410	396	6987	7793
after reduction	30	30	221	273	62	179	374	615
%	100	100	32.89	37.24	15.12	45.20	5.35	7.89

Note: The labels *x*, *y* and *z* refer to any phonemes, the only possible combinations being of two or three phonemes appearing in the Serbian language. When # is in front of *x* it marks the beginning of a word, and when behind *x* or *y*, it indicates the end of the word. % raw refers to the percentage of maintained bigrams and trigrams out of the total number of bigrams and trigrams prior to its reduction.

Preparation of material for analysis: The existence of a large number of cells in the matrices in which the frequency equals zero (sparse data) does not represent a special challenge for SVMs (Farquad *et al*., 2010). However, SVMs are more effective if the characteristics that rarely occur, e.g. only once, are excluded from the analysis (Nakagawa *et al*., 2001). In addition, some authors suggest that linear scaling of input data within the range [-1, 1] or [0, 1] should be adopted before applying SVM (Hsu *et al*., 2010). For that reason, the frequences of bigrams and trigrams were calculated using the modification of the *simple Good-Turing discounting* (Gale and Sampson, 1995), suggested by Milin and Moscoso del Prado Martín (p.c.). The modification of Good-Turing correction for frequencies allowed avoiding uniform attribution of one value to all variables whose frequency was zero for the given MSC. It was assumed that the proportions obtained by the Good-Turing procedure (the number of bigrams and trigrams appearing only once, divided by the number of bigrams and trigrams whose frequency is zero) multiply by the probability of the rows and probability of the empty cell column ($f = 0$) and are divided by the sum of all products of the probabilities of rows and columns containing empty cells. Thus, the corrected value depends on the total frequency of the variable (the sum of bigram or trigram occurence in all MSC), but also on the frequency of the MSC (the sum of all bigrams or trigrams occurences in one MSC).

Training and test dataset: The training sample size was 75% of the total sample or 969 MSC, while the test sample contained 324 MSC (Table 3).

Table 3. Structure of training sample and test-sample

WC	N	training sample		test-sample	
		n	%	n	%
nouns	74	54	5.57	20	6.17
adjectives	131	103	10.63	28	8.64
pronouns*	594	438	45.20	156	48.15
verbs	494	374	38.60	120	37.04
TOTAL	1293	969	100.00	324	100.00

* According to the traditional classification, prounouns included the following categories: personal pronouns, possessive pronouns, demonstrative pronouns, interrogative pronouns, indefinite pronouns, negative pronouns, and general pronouns.

Statistical analysis: The justification of the use of PI in the task of discriminating WC was verified using support vector machines C-SVM, with a *linear kernel function*, implemented in the STATISTICA software version 7 (StatSoft Inc., 2004). The linear kernel function implies the adjustments of one parameter – *C* parameter, which determines the "punishment" for incorrectly classified objects. It has been estimated by means of a network search, recommended by a number of authors (e.g. Olson and Dulen, 2008; Hsu *et al*., 2010). The first step meant varying size *C* within a range $[2^{-5}, 2^{15}]$, increasing the exponent by two. When the area where the best result is established was found, the *C* parameter was varied within a given range, with an increase in the exponent by 0.25, using cross-validation with a sample divided into five equal parts. In the end, for the optimal value of *C* parameter, a new classification was made on a sample divided by the ratio 75% : 25%, most of which was the training sample, and a smaller part the test sample.

2. 2. Results and discussion

Efficacy of WC discrimination based on PI generated at the level of MSC can be classified into three groups. The highest percentage of accurately separated WC

ranges between 91.4 and 93.2%, and it is obtained by cases of bigrams with no additional information about their position in words [xy], all the trigrams taken together [#xz, xyz, xz#], and all bigrams taken together [#x, xy, x#] (Figure 1). Between these three types of PI there was no statistically significant difference in the efficiency of discrimination. In the second group there is PI based on which a lower discrimination accuracy is obtained: bigrams at word beginnings [#x], trigrams at word beginnings [#xy] and at word endings [xy#], and trigrams at any position in a word [xyz].[5] The smallest percentage of accurately processed WC (67.28%) was obtained on the basis of probability information of bigrams at word beginnings [x#] (Figure 1), which is less statistically different from discrimination, which relied on probability information of bigrams at word beginnings [#x] (*McNemar's* χ^2 (1) = 13.6, $p < .001$).

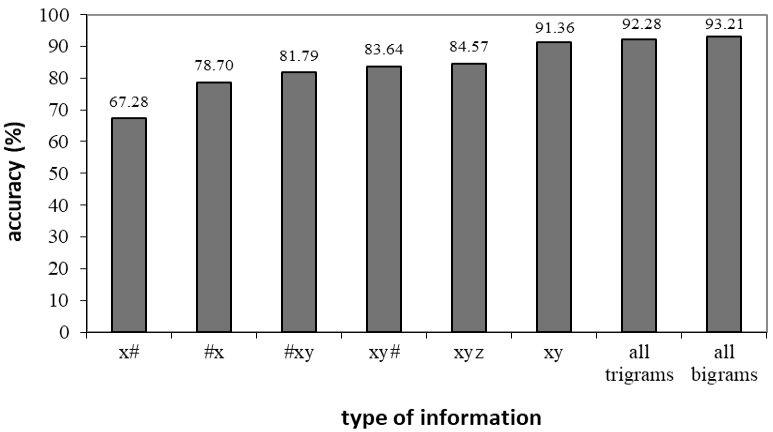

Figure 1. Accuracy of WC discrimination based on bigrams and trigrams. The labels x, y and z refer to any phonemes, the only possible combinations being of two or three phonemes appearing in Serbian. When # is in front of *x* it marks the beginning of a word, and when behind *x* or *y*, it indicates the end of the word.

When examining discrimination accuracy depending on the word class and type of PI on which the discrimination relies, similar results are obtained. The greatest number of mistakes was obtained in cases where discrimination was done on the basis of bigrams and trigrams at word beginnings and endings (Figure 2). The exceptions are adjectives, in which the best result was achieved when used on

[5] The final judgment of the efficacy of the classification based on the trigrams [xyz], should take into account that for this type of PI, only 5.35% of all trigrams that appeared in the sample was kept in the analysis.

trigrams from word beginnings. [#xy] (Figure 2). The question is whether this is a consequence of a greater discrimination of adjectives on the basis of the initial trigrams, or the reason lies in the structure of the test sample, since the adjectives are represented in a relatively small number of cases (8.6%), which is why one should be cautious in making the final conclusion.

When looking at discrimination efficacy of individual WC, regardless of the PI used, the largest percentage of errors was found in nouns, where the error ranged from 35% to 80%, and the smallest in pronouns, ranging from 0.64% to 15% (Figure 2). The reason why pronouns are the most accurately processed word class lies in the fact that in case of pronouns we are dealing with a finite set of words, which get distributed through a big number of MSC, and this facilitates their classification.

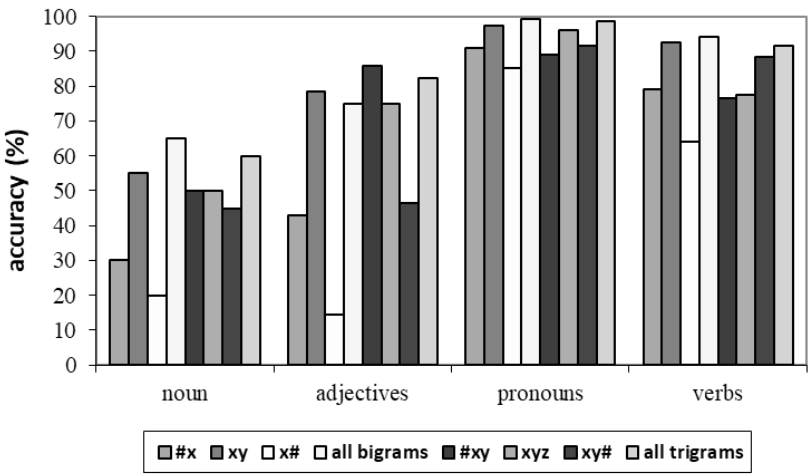

Figure 2. Accuracy of discrimination depending on WC. The labels *x, y* and *z* refer to any phonemes, the only possible combinations being of those appearing in Serbian. When the label # is in front of *x,* it marks the beginning of a word, and when it is found behind *x* or *y*, it indicates the end of the word.

The results obtained in this study showed that, based on the PI frequencies calculated at the level of MSC, the inflected WC that these MSC belong to can be successfully determined. However, not all types of PI are sufficiently informative to successfully accomplish that task. The four least effective classifications were obtained for cases of bigrams and trigrams at word beginnings and word endings. Baayen *et al*. (2011) analyse the factors that influence potential for separation of MSC, noting that in languages with rich inflectional morphology suffixes do not

have an important role in word processing, because more MSC can share the same inflectional suffix.

This suggests that inflectional suffixes can not be useful as bearers of specific meanings, but can be bearers of useful redundancy, thus reducing the level of uncertainty. Within the discussion about the development of linguistic structures, Kostić (2004) makes a similar claim regarding the role of suffixes in the marking of syntactic functions and the word meaning. Increasing the complexity of the system (with a growing number of lexemes and their relation) within limited cognitive capacities results in a constant reorganisation of language. The basic mechanisms on which the reorganisation is based are sharing and introducing, above all, new subclasses. Kostić points out that these changes, from the point of descriptive linguistics, seem unmotivated. However, if considered from the perspective of language development, seen as a dinamic complex system (Beckner *et al*., 2012), processes of self-regulation aiming at optimal distribution of information load are significant (Kostić, 2004 : 51).

It should be noted that in this study the emphasis was not on reaching the maximum, i.e. the upper limit of the accuracy that can be achieved in the task of discriminating inflected WC based on PI. If this is set as a goal, it is necessary to systematically vary the number of bigrams and trigrams in order to determine their optimal number: the smallest sample in which a maximum percentage of accurately classified MSC is obtained.[6] It would also be useful to determine a subset of those bigrams and tirigrams that have the greatest discriminatory power. This would provide additional information on bigrams and trigrams, as well as the possibility to correct the accuracy of discrimination based on PI. Moreover, if the goal is maximum accuracy, during the use of SVM it is necessary to check other kernel functions as well, like the *gaussian radial basis function, exponential radial basis function*, *polynomial*, *neural/sigmoid or tanh* etc., and not only the linear function.

3. STUDY 2. The role of phonotactic information in the production of inflectional forms

In the previous study, the possibility of discrimination of inflected WC based on bigrams and trigrams was examined. This problem refers to the domain of automatic processing, i.e. the understanding of language. However, there remains the question of what is the role of PI and how it is used in language production. In

[6] On the problem of stability of language distributions and achieving optimal sample size for different tasks in language processing the reader is referred to Dimitrijević *et al*., 2009. and Kostić *et al*., 2008.

the second study, automatic production of IF was tested, using a model that relies on phonological similarity with examples from experience.

The task set for the model was the production of the correct (attested) inflected form for a given lemma. The model relied on PI from the last four syllables of the base form (lemma) of the word.[7] For each word to be processed, a target of MSC was also set.

To accomplish this task, memory-based learning (MBL) (Daelemans and van den Bosch, 2005) was used. It is a model of language acquisition and processing, which can be regarded as operationalisation of the approaches relying on analogy. MBL is based on the assumption that in the process of executing cognitive tasks, no mental rules or other abstract representations derived from the experience are being used. Instead, conclusions about a new situation are made directly on the basis of its similarity to the events from the past (see also Daelemans and van den Bosch, 2005; Daelemans *et al.,* 2010). Since this approach does not make a difference between storing correct and incorrect forms, MBL belongs to the class of one-route models, which presuppose one unique mechanism of IF production. This mechanism primarily relies on phonological similarity (Rumelhart and McClelland, 1986), but also on other, nonphonological information (Keuleers *et al*., 2007).

MBL proved successful in different tasks of automatic language processing, from morpho-phonological analyses and classifications to text and discourse analyses. Automatic text processing relying on MBL was also tested on the phenomena of inflection, e.g. in forming past tense in English (Keuleers, 2008), forming plural in Dutch (Keuleers and Daelemans, 2007; Keuleers *et al*., 2007) and German (Hahn and Nakisa, 2000), showing gender in Spanish (Eddington, 2002a), diminutive in Spanish (Eddington, 2002b) and Dutch (Daelemans *et al*., 1997), prediction of infixes in Dutch compouds (Krott *et al*., 2001) and German compounds (Krott *et al*., 2007). Also, generating alomorphic variants of instrumental singular masculine nouns in Serbian, was also examined using MBL (Milin *et al*., 2011).

3. 1. Method

Sample: The efficacy of the model was tested on the sample of 28.971 different words, appearing in 89.024 different forms. The words in the sample belonged to

[7] In these tasks it is usual to rely on syllables and their constituents: *onset*, *nucleus*, *coda* (Keuleers *et al*., 2007). Considering the requirement that all examples in memory contain the same number of elements (*alignment method*), which is a prerequisite for the use of MBL, in cases in which a word contained less than four syllables, the lack of phonemes in a certain position within the vector is signalled by a special sign.

the group of inflected WC (nouns, adjectives, pronouns and verbs), and adverbs were also included, since they can be inflected in comparison. The sample was formed on the basis of *Frekvencijski rečnik dnevne štampe* (*Frequency dictionary of daily press*), a part of *Frekvencijski rečnik savremenog srpskog jezika* (*Frequency dictionary of Contemporary Serbian language*; Kostić, 1999). From this subsample all atypical word forms were removed, such as words with a hyphen, numbers, etc. For each word form, which was also the target form that was desired to be obtained by applying the model, there was information about the lemma and the code. This code precisely determined the morphosyntactic status of the target form. To illustrate, in the code 201221, the first number indicates the word is an adjective, the third tells it is a basic form, the fourth indicates it is genitive, the fifth signals it is plural, and the last number says it is an adjective in masculine form. One word could appear only once in the determined grammatical form.

The words in the sample were distributed in 1160 different MSC. Such a large number of MSC is a consequence of the rich inflectional morphology of Serbian (Table 4). Out of 1160 MSC one half contained one word, 20% (233 MSC) had two to five words, and the remaining 30% (345 MSC) varied from six to 6783 different words.

Table 4. *Distribution of word forms in the sample depending on WC*

WC	Word forms (n)	Percentage	Morphosyntactic category (n)
nouns	41062	46.12	70
adjectives	25189	28.30	117
pronouns	906	1.02	534
verbs	20618	23.16	434
adverbs	1249	1.40	5
Total	89024	100.00	1160

Instrument: The model intended for the production of IFs in Serbian was implemented using TiMBL software – *Tilburg memory based learner* (Daelemans *et al.*, 2010), which allows the application of several different memory-based learning

algorithms and *k-nearest neighbour classification*, tailored to the tasks performed on the language material.

During the implementation of the model, as a measure of the similarity of the examples stored in memory, the MVDM metric was used (*modified value difference metric*; Cost and Salzberg, 1993), when the closest distance is $k = 7$. Unlike the k-NN algorithm, where k refers to the number of closest neighbours, in memory-based learning, this value refers to the *k*-nearest distance (Daelemans and van den Bosch, 2005; Daelemans *et al.*, 2010). Therefore, the number of examples in the set nearest neighbours may be even greater than seven. This value is taken because the 7-NN model proved more robust than other models in simulating the results obtained in the production of plural nouns in Dutch and the past of novel forms in English, (Keuleers, 2008; Keuleers and Dealemans, 2007), which was confirmed in the task of generating alomorphs in Serbian (Milin *et al.*, 2011).

During the implementation we used a variant of the model in its simplest, basic form. This means that each of the *k* exemplars from the nearest neighbours had the same weight (*zero decay weighting*), i.e. they are not additionally weighted depending on the degree of similarity (*size of distance*) with the object being classified. No additional optimisation of the process was performed in terms of improving performances such as, for example, representing exemplars in memory with more abstract properties (e.g. prototypes) or a selection of exemplars based on their usefulness, which would make the model more complex.

Procedure: In the first step, exemplars representing knowledge stored in memory, necessary for automatic language production, are created. Exemplars are formed for each word from the sample, and are represented by one-dimensional vectors with 14 elements, such as, for example, the genitive form of the masculine plural for the adjective *opasan*:

=, =, =, =, o, =, p, a, =, s, a, n, 201221, 987654320*ih*

In the first twelve vector cells, the last four syllables are placed. They were obtained by the decomposition of the lemma; the thirteenth contains the code, i.e. a detailed MSC specification of the word, and the fourteenth contains the inflection form. During the decomposition of the lemma into syllables, marked cases included lemmas in which the syllable had a consonant beginning or ending. An inflection form represents the way by which the assigned word form is obtained. In the inflection form, letters mark the inflection suffix, while the numbers from zero to nine refer to the phonemes from the basic word form, so that the null marks the last one, one marks the one before the last phoneme etc. The absence of some

of the digits within the IF means that the morphosyntactic form does not contain the phoneme which in the basic form of the word was located in the place indicated by that number. For example, the correct form of the genitive of plural masculine form of the adjective *opasan* is obtained by the application of IF *987654320ih*. So, from the basic form of the word, the phoneme in the place before the last one was ommitted, and then the suffix *ih* added: opas(a)n = opasn + *ih* = opasnih.

When testing the model, a leave-one-out procedure was used. The model starts from the information about the phonological structure of the word and the information about the morphosyntactic category to be tested (the first 13 positions for each of the word forms in memory). In the first stage of modeling, a set of neighbours is formed, with the maximum distance *k*, which, in this case, represents the number of allowed transformations in relation to the tested word. The next phase involves deciding the IF for the required form. This is achieved by calculating the probabilities of each IF appearance in the nearest neighbours set (f / K), where the IF with highest probability is the expected class for the tested form.

3. 2. Results and discussion

Relying solely on the PI and information on the morphosyntactic status of words, in the task of inflectional production, MBL successfully generates 89% of the words (Table 5). This percentage is even higher, since there are no marked cases in the sample when two forms are equal, i.e. accurate (for instance, double noun forms *vukovi – vuci*; shorter or longer forms of the adjective *plav – plavi*, adverb *kad – kada*; pronoun *kog – koga*; *ekavica* and *ijekavica* dialect *cvet – cvijet* etc.). On a random subsample of 537 words that were incorrectly processed (around 5.5% of all incorrectly processed words) there were about 29% of such cases. If this correction is taken into account, the percentage of successfully produced IFs is around 92.2%.

Table 5. *Efficacy of memory-based learning in the production of inflectional forms, depending on WC*

WC	Accurate		Total
	n	%	
nouns	38190	93.01	41062
adjectives	21820	86.63	25189
pronouns	549	60.60	906

WC	Accurate		Total
verbs	17519	84.97	20618
adverbs	1188	95.12	1249
TOTAL	79266	89.04	89024

In order to get a better insight into the factors that affect the accuracy of automated IF production using MBL, a separate analysis was made for nouns (Table 6). Several factors could be responsible for poorer results in certain MSCs: (a) an insufficient number of exemplars,[8] (b) the appearance of phonological alternatives (palatalisation, loss of the vowel *a*), (c) the existence of doublet forms, (d) different suffixes used for the same word, etc.

Table 6. *Efficacy of memory-based learning in the production of inflectional forms of nouns, depending on the grammatical number and gender*

CASE	Efficacy (%)	NUMBER		GENDER			Total of exemplars
		Singular	Plural	Masc. gender	Femin. gender	Neutral gender	
Nominative	97.28	98.83	91.24	96.92	97.98	97.61	13347
Genitive	92.00	94.25	86.99	90.77	93.60	92.62	9494
Dative	93.01	92.78	93.54	91.57	94.23	96.00	2159
Accusative	88.72	87.57	91.58	78.41	97.26	96.97	7581
Vocative	64.29	55.78	89.80	65.25	61.54	66.67	196
Instrumental	92.09	92.32	91.59	86.97	96.76	95.69	3666
Locative	91.69	92.26	90.33	89.20	92.13	96.17	4619
	93.01	94.02	90.26	90.63	95.59	95.75	41062

In order to illustrate challenges arising from the task in the process of automatic production of IF in a single MSC, a nominative of the plural of masculine nouns can be used. For this form the model scored an accuracy of 87%. In the sample

[8] For this reason the worst result was obtained for pronouns. The pronouns sample contained 906 examples distributed within 534 different MSC, which gives 1.7 examples per category on average. For nouns, this ratio was 586.6 per category.

of 1490 units, there were 30 nouns of masculine gender which form a plural nominative by adding an inflection suffix *–i* with infix *–ev* and 142 nouns that form a plural by adding a suffix *–i* and infix *–ov*. Other cases were related to nouns whose plural is obtained with the suffix *–i*. In 90% of cases, MBL wrongly classified the masculine nouns that form the plural with a suffix *–ev+i*. Witin these errors, the model wrongly attributed the suffix *–i* in 11 cases (e.g. *kursi* instead of *kursevi*), and in 16 cases the suffix *–ov-i* (e.g. *žuljovi, lešovi* instead of *žuljevi*, *leševi*). The error on the nouns whose plural is obtained by means of suffix *–ov+i* was about 23%. Within these errors, around 42% refers to the forms in which phonological alternations concern the nominative plural (mostly the unstable *a*). On the other hand, the forms with phonological alternatives give a correct output in only about 3% of cases. In almost all incorrectly processed masculine nouns forming a plural with the suffix *–ov+i,* the error is reflected in an attempt to make the desired form with an inflectional suffix *-i* (e.g. *ritami*, *pojami* instead of *ritmovi*, *pojmovi* etc.).

In incorrectly processed masculine nouns, where the plural is made by adding the suffix *–i* (134 forms), in one case a wrong suffix is added *–evi* (*konjevi*), and in 28 cases (about 21%) the wrong suffix relates to *–ovi* (e.g. *pasovi*, *mravovi* etc.). Other errors are mostly related to phonological alternations within plural forms, such as: unstable *a* (*pucanji* instead of *pucnji*), palatalisation (*hirurgi*, *potoki* instead of *hirurzi*, *potoci*) or their combinations (*čuperaki* instead of *čuperci*), etc.

These examples indicate that phonological alternations pose a challenge in the task of automated inflectional morphology production using TiMBL. In addition, the model seems to have a tendency to facilitate processing by using more frequent suffixes. This is in line with the conclusion that linguistic production, at least to some extent, depends on the frequency of the grammatical form (*type frequency*; Bybee, 2001, 2010). According to Bybee, the higher the frequency of the grammatical form, the greater the probability that it will be applied to new cases.

If the goal is to reach the maximum accuracy, the MBL efficacy should be checked depending on the PI specification method. Namely, although the usual procedure for obtaining exemplars of the same length is to break words into syllables and their elements (Keuleers *et al*., 2007), some findings argue against the claim that phonotactic constraints depend solely on the structure of the syllable (e.g. Blevins, 2003; 2006). This is supported by the results obtained in this study, in the task of the discrimination of WC by means of SVM, but also by the efficacy of *Naive Discriminative Learning* (Baayen *et al*., 2011), which also relies on bigrams and trigrams.

4. General discussion

The present study examined the extent to which our cognitive system can rely on PI in word processing in a language with rich inflectional morphology. Informativeness of phonotactic sequences has been tested in two ways: (1) in the task of discriminating the inflected WC, by means of SVM (Vapnik, 1995, 1998); (2) in the task of automatic production of IF, by means of TiMBL (Daelemans *et al*., 2010).

Discrimination efficacy of the inflected WC based on PI, specified on the basis of a morphosyntactic category, which was obtained in all bigrams [#x, x#, xy] was about 93%. On the other hand, in the task of inflectional production by means of MBL (Daelemans and van den Bosch, 2005), approximately 92% of IFs was generated correctly on all the inflected WC taken together. This percentage of accurately processed words is slightly smaller than the accuracy obtained by means of the most effective taggers used in these or similar tasks (see, e.g. Güngör, 2010; Manning, 2011; Manning and Schuetze, 2000), and is in line with the results obtained for Serbian (Sečujski and Kupusinac, 2009). At the same time, we should not forget that the PI is a simple, if not even the simplest form of information that can be used in these and similar tasks. Also, MBL (Daelemans and van den Bosch, 2005) was applied to all inflected words, which, otherwise, is not a common practice. Typically, MBL is applied to a single linguistic phenomenon, e.g. generating allomorphy (Milin *et al*., 2011), plurality (Hahn and Nakisa, 2000; Keuleers and Daelemans, 2007; Keuleers *et al*., 2007) and alike.

The results show that perception and production of morphologically complex words can be performed on the basis of phonological/orthographic and semantic information, and that there is no need to postulate a special domain in charge of morphology processing (Baayen *et al*., 2011; Davis, 2004; Harm and Seidenberg, 1999, 2004; Seidenberg and Gonnerman, 2000, among others). This is consistent with Bybee's assumption that the connection between words in a cognitive system is phonological and semantic by nature (Bybee, 1985, 2001, 2010). In one such model, information about the sequences of phonemes/graphemes is necessary, so that the empirical findings about the role of phonotactic constraints and the size of similar neighbours sets in language processing become integrated. Both of these factors are not independent of words. On the other hand, the connections between words are not only based on the similarity of the shared phonemes/graphemes, but are also semantic in nature. In fact, this is what all one-route models and models in favour of direct mapping between form and meaning point out (Baayen *et al*., 2011; Bybee,

1985, 1999, 2010; Plaut and Gonnerman, 2000; Rumelhart and McClelland, 1986; Seidenberg and Gonnerman, 2000). Ultimately, models must include, at least implicitly, a knowledge of the world (Dimitrijević, 2007; Milin, 2004). In the case of automated production of IF using MBL, this implicit knowledge is reflected in the information on the required morphosyntactic category. The morphosyntactic category is not a bearer of grammatical meaning only, but it also contains information about the function of the requested form, i.e. its function in a sentence, such as the subject, which includes a "knowledge of the world". Therefore, grammar is not represented by a set of abstract rules, as indirectly related to linguistic experience, but in the form of *patterns* that *emerge*. These patterns arise as a direct consequence of the use of language information stored in the cognitive system (Bybee, 2001; Beckner *et al*., 2012). To put it differently, grammatical knowledge is procedural knowledge, which also refers to phonology. In this way, phonology takes part in the knowledge related to production and perception of grammatical constructions (Bybee, 2001).

The results obtained for the processing and production of complex words in a language with rich inflectional morphology show that functional links between orthography/phonology and learning-based models, which do not assume the cognitive reality of morphemes, enable success in lexical learning. At the same time, these findings have far-reaching implications, considering that they point to a number of constraints of the structural theories of language. Besides, they call into question large and expensive postulates on the mental lexicon and mental representation of words containing a complex but not sufficiently specified organisation of various linguistic features.

Acknowledgments

I would like to thank Emmanuel Keuleers for his significant assistance in the preparation of data and implementation of TiMBL software, and my mentor Petar Milin for continuous support in the development of the thesis.

References

1. Andrews, Sally (1989), „Frequency and neighborhood size effects on lexical access: Activation or search?“, *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, 15, 802-814. DOI: 10.1037/0278-7393.15.5.802

2. Andrews, Sally (1992), „Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy?, *Journal of Experimental Psychology*: *Learning, Memory, and Cognition,* 18, 2, 234-254. DOI: 10.1037//0278-7393.18.2.234.
3. Baayen, Harald (2003), Probabilistic approaches to morphology. In: Rens Bod, Jennifer Hay and Stefanie Jannedy (Eds.), *Probability theory in linguistics* (pp. 229-287). Cambridge: MIT Press.
4. Baayen, Harald (2011), „Corpus linguistics and naive discriminative learning." *Brazilian Journal of Applied Linguistics,* 11, 2, 295-328. DOI: 10.1590/S1984-63982011000200003.
5. Baayen, Harald, Petar Milin, Dušica Filipović Đurđević, Peter Hendrix and Marco Marelli (2011), „An amorphous model for morphological processing in visual comprehension based on naive discriminative learning", *Psychological Review,* 118, 438-481. DOI: 10.1037/a0023851.
6. Bailey, Tod and Urlike Hahn, U. (2001), „Determinants of wordlikeness: Phonotactics or lexical neighborhoods?", *Journal of Memory and Language*, 44, 4, 568-591. DOI: 10.1006/jmla.2000.2756.
7. Balota, David, Michael Cortese, Susan Sergent-Marshall, Daniel Spieler and Melvin Yap (2004), „Visual word recognition of single-syllable words", *Journal of Experimental Psychology*: *General*, 133, 2, 283-316. DOI: 10.1037/0096-3445.133.2.283.
8. Baker, Paul, Andrew Hardie and Tony McEnery (2006), *A glossary of corpus linguistics*, Edinburgh: Edinburgh University Press.
9. Beckner, Clay, Nick Ellis, Richard Blythe, John Holland, Joan Bybee, Jinyun Ke, Morten Christiansen, Diane Larsen-Freeman, William Croft and Tom Schoenemann (2012), „Language is a complex adaptive system: Position paper", *Language Learning*, 59, 1, 1-26. DOI: 10.1111/j.1467-9922.2009.00533.x.
10. Blevins, Juliette (2003), The independent nature of phonotactic constraints: An alternative to syllable-based approaches, In: Caroline Féry and Ruben Van de Vijver (Eds.), *The syllable in optimality theory* (pp. 375-403). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511497926.016.
11. Blevins, Juliette (2006), „Word-based morphology", *Journal of Linguistics*, 42, 3, 531-573. DOI: 10.1017/S0022226706004191.
12. Bloomfield, Leonard (1933), *Language*, New York: Holt, Rinehart and Winston.
13. Boas, Hans (2003), *A constructional approach to resultatives*. *Stanford monographs in linguistics,* Stanford: CSLI Publications.
14. Bybee, Joan (1985), *Morphology*: *A study of the relation between meaning and form*, Amsterdam: John Benjamins.
15. Bybee, Joan (2001), *Phonology and language use*, Cambridge: Cambridge University Press.
16. Bybee, Joan (2010), *Language, usage and cognition*, New York: Cambridge University Press.
17. Bybee, Joan & David Eddington (2006), „A usage-based approach to Spanish verbs of 'becoming'", *Language,* 82, 2, 323-55. DOI: 10.1353/lan.2006.0081.

18. Cairns, Paul, Richard Shillcock, Nick Chater and Joseph Patrick Levy (1997), „Bootstrapping word boundaries: A bottom-up corpus based approach to speech segmentation", *Cognitive Psychology*, 33, 2, 111-153. DOI: 10.1006/cogp.1997.0649.
19. Cattell, Raymond (1966), „The scree test for the number of factors", *Multivariate Behavioral Research*, 1, 2, 245-276. DOI: 10.1207/s15327906mbr0102_10.
20. Coleman, John and Janet Pierrehumbert (1997), „Stochastic phonological grammars and acceptability", *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology* (pp 49-56). Somerset: Association for Computational Linguistics.
21. Cost, Scott and Steven Salzberg (1993), „A weighted nearest neighbor algorithm for learning with symbolic features", *Machine Learning*, 10, 1, 57-78. DOI: 10.1007/BF00993481.
22. Crystal, David (2008), *An encyclopedic dictionary of language and languages,* Oxford: Blackwell Publishing.
23. Daelemans, Walter, Peter Berck and Steven Gillis (1997), „Data mining as a method for linguistic analysis: Dutch diminutives", *Folia Linguistica*, 31, 1-2, 57-75. DOI: 10.1515/flin.1997.31.1-2.57.
24. Daelemans, W. and van den Bosch, A. (2005). *Memory-based language processing*, Cambridge: Cambridge University Press.
25. Daelemans, Walter, Jakub Zavrel, Ko Van der Sloot and Antal van den Bosch, (2010), *TiMBL*: *Tilburg memory based learner, version 6.3, reference guide,* ILK Research Group Technical Report Series no. 10-01. Retrieved from http://ilk.uvt.nl/downloads/pub/papers/ ilk.1001.pdf.
26. Davis, Stuart (1988), *Topics in syllable geometry*, New York: Garland.
27. De Saussure, Ferdinand (1916/1966), *Course in general linguistics*, New York: McGraw-Hil.
28. Dimitrijević, Strahinja (2007), *Kognitivne strategije u obradi jezika*: *Primjena kontekstualnih jezičkih informacija u zadatku automatske lematizacije*, (Nepublikovani magistarski rad), Filozofski fakultet, Univerzitet u Banjoj Luci, Banja Luka.
29. Dimitrijević, Strahinja, Aleksandar Kostić and Petar Milin (2009), „Stability of the syntagmatic probability distributions", *Psihologija*, 42, 1, 107-119.
30. Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua*, *110*(4), 281-289. DOI: 10.1016/S0024-3841(99)00043-1.
31. Eddington, David (2002a), „Spanish gender assignment in an analogical framework", *Journal of Quantitative Linguistics,* 9, 1, 49-75. DOI: 10.1076/jqul.9.1.49.8482
32. Eddington, David (2002b), „A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling", In: Royal Skousen, Deryle Lonsdale and Dilworth Parkinson (Eds.), *Analogical modeling*: *An exemplar-based approach to language* (pp. 141-156). Amsterdam: John Benjamins.

33. Farquad, M. A. H., Ravi, V. and Raju Bapi (2010), „Support vector machine based hybrid classifiers and rule extraction thereof: Application to bankruptcy prediction in banks", In: Emilio Soria Olivas, Jose David Martin Guerrero, Marcelino Martinez Sober, Jose Rafael Magdalena Benedito, Antonio Jose Serrano Lopez, (Eds.), *Handbook of research on machine learning applications and trends*: *Algorithms, methods, and techniques* (pp. 404-426). Hershey: IGI Global.
34. Frisch, Stefan, Nathan Large and David Pisoni (2000), „Perception of wordlikeness: Effects of segment probability and length on processing non-words", *Journal of Memory and Language*, 42, 481-496.
35. Gale, William and Geoffrey Sampson (1995). „Good-Turing frequency estimation without tears", *Journal of Quantitative Linguistics,* 2, 3, 217-237. DOI: 10.1080/092961795 08590051.
36. Grainger, Jonathan, Mathilde Muneaux, Fernand Farioli and Johannes Ziegler (2005), „Effects of phonological and orthographic neighborhood density interact in visual word recognition", *Quarterly Journal of Experimental Psychology*, 58A, 981-998 DOI: 10.1080/02724980443000386.
37. Güngör, Tunga (2010), „Part-of-speech tagging", In Nitin Indurkhya and Fred Damerau (Eds.), *Handbook of natural language processing* (pp. 205-236), Boca Raton: Taylor and Francis Group, LLC.
38. Hahn, Ulrike and Ramin Charles Nakisa (2000), "German inflection: Single route or dual route?", *Cognitive Psychology,* 41, 4, 313-360. DOI: 10.1006/cogp.2000.0737.
39. Harm, Michael and Mark Seidenberg (1999), „Reading acquisition, phonology, and dyslexia: Insights from a connectionist model", *Psychological Review*, 106, 3, 491-528. DOI: 10.1037//0033-295X.106.3.491.
40. Harm, Michael and Mark Seidenberg (2004), „Computing the meanings of words in reading: Division of labor between visual and phonological processes", *Psychological Review*, 111, 3, 662-720. DOI: 10.1037/0033-295X.111.3.662
41. Harris, Zelling. (1951), „*Methods in structural linguistics*", Chicago: University of Chicago Press.
42. Harris, Zelling (1957), „Co-occurrence and transformation in linguistic structure", *Language*, 33, 3, 283-340. DOI: 10.2307/411155
43. Hsu, Chih-Wei, Chih-Chung Chang and Chih-Jen Lin (2010), *Practical guide to support vector classication*, Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/ guide.pdf.
44. Joachims, Thorsten (1998), „Text categorization with support vector machines: Learning with many relevant features", *European Conference on Machine Learning* (*ECML*). DOI: 10.1007/BFb0026683.
45. Jusczyk, Peter, Paul Luce and Jan Charles-Luce (1994), „Infants' sensitivity to phonotactic patterns in the native language", *Journal of Memory and Language*, 33, 630-645.

46. Keuleers, Emmanuel (2008), *Memory-based learning of inflectional morphology*, Unpublished doctoral thesis, Faculteit Lettern en Wijsbergeerte, Universiteit Antwerpen, Antwerpen.
47. Keuleers, Emmanuel (2018), „Computational Approaches to Morphology", *Oxford Research Encylopedia of Linguistics*, Oxford: Oxford University Press.
48. Keuleers, Emmanuel and Walter Daelemans (2007), „Memory-based learning models of inflectional morphology: A methodological case study", *Lingue e Linguaggio*, 6, 151-174.
49. Keuleers, Emmanuel, Dominiek Sandra, Walter Daelemans, Steven Gillis, Gert Durieux and Evelyn Martens (2007), „Dutch plural inflection: The exception that proves the analogy", *Cognitive Psychology*, 54, 4, 283-318. DOI: 10.1016/j.cogpsych.2006.07.002.
50. Kostić, Aleksandar (1991), „Informational approach to processing inflected morphology: Standard data reconsidered", *Psychological Research*, 53, 1, 62-70. DOI: 10.1007/ BF00867333.
51. Kostić, A. (2004). Kognitivna ograničenja i obrada jezika. U A. Kostić, D. Todorović, S. Marković (Eds.), *Jezik i opažanje. Tri studije iz eksperimentalne psihologije* (pp. 7-51). Beograd: Filozofski fakultet, Univerzitet u Beogradu.
52. Kostić, Aleksandar, Svetlana Ilić i Petar Milin (2008), „Aproksimacija verovatnoća i optimalna veličina jezičkog uzorka", *Psihologija*, 41, 1, 35-51.
53. Kostić, Đorđe (1999), *Frekvencijski rečnik savremenog srpskog jezika*, Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju, Univerzitet u Beogradu.
54. Kostić, Đorđe (2001), *Korpus srpskog jezika*, Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju, Univerzitet u Beogradu.
55. Krott, Andrea, Harald Baayen and Robert Schreuder (2001), „Analogy in morphology: Modeling the choice of linking morphemes in Dutch", *Linguistics,* 39, 1, 51-93. DOI: 10.1515/ling.2001.008.
56. Krott, Andrea, Robert Schreuder, Harald Baayen and Wolfgang Dressler (2007), „Analogical effects on linking elements in German compounds", *Language and Cognitive Processes*, 22, 1, 25-57. DOI: 10.1080/01690960500343429.
57. Levelt, Willem and Linda Wheeldon (1994), „Do speakers have access to a mental syllabary?", *Cognition*, 50, 1-3, 239-269. DOI: 10.1016/0010-0277(94)90030-2.
58. Levenshtein, Vladimir (1966), „Binary codes capable of correcting deletions, insertions, and reversals", *Soviet physics doklady*, 10, 707–710.
59. Luce, Paul and Nathan Large (2001), „Phonotactics, density, and entropy in spoken word recognition", *Language and Cognitive Processes*, 16, 5-6, 565-581. DOI: 10.1080/01690960143000137.

60. Manning, Christopher and Hinrich Schuetze (2000), *Foundations of statistical natural language processing*, Cambridge: MIT Press.
61. Manning, Christopher (2011), „Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?", In Alexsander Gelbukh (Ed.), *Proceedings in the 12th International Conference Computational Linguistics and Intelligent Text Processing, CICLing*, 171-189, Berlin: Springer.
62. Mattys, Sven and Peter Jusczyk (2000), „Phonotactic cues for segmentation of fluent speech by infants", *Cognition*, 78, 91-121.
63. Mattys, Sven, Peter Jusczyk, Paul Luce and James Morgan (1999), „Phonotactic and prosodic effects on word segmentation in infants", *Cognitive Psychology*, 38, 4, 465-494. DOI: 10.1006/cogp.1999.0721.
64. Meyer, David, Friedrich Leisch and Kurt Hornik (2003), „The support vector machine under test", *Neurocomputing*, 55, 1-2, 169-186. DOI: 10.1016/S0925-2312(03)00431-4.
65. Milin, Petar (2004). *Probabilistički pristup određivanju gramatičkog statusa reči i kognitivne strategije u obradi jezika,* Nepublikovana doktorska disertacija, Filozofski fakultet, Univerzitet u Beogradu; Beograd.
66. Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević and Harald Baayen (2016). „Towards cognitively plausible data science in language research", *Cognitive Linguistics, 27*(4), 507-526. DOI: 10.1515/cog-2016-0055.
67. Milin, Petar, Emmanuel Keuleers and Dušica Filipović Đurđević (2011), „Allomorphic responses in Serbian pseudo-nouns as a result of analogical learning", *Acta Linguistica Hungarica*, 58, 1, 65-84. DOI: 10.1556/ALing.58.2011.1-2.4.
68. Morgan, James and Jenny Saffran (1995), „Emerging integration of sequential and suprasegmental information in preverbal speech segmentation", *Child Development*, 66, 4, 911-936. DOI: 10.2307/1131789.
69. Nakagawa, Tetsuji, Taku Kudo and Yuji Matsumoto (2001), „Unknown word guessing and part-of-speech tagging using support vector machines", *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium* (NLPRS-2001), 325-331, Tokyo, JP.
70. Olson, David and Dursun Dulen (2008), *Advanced data mining techniques,* Berlin: Springer-Verlag.
71. Pitt, Mark and James McQueen (1998), „Is compensation for coarticulation mediated by the lexicon?", *Journal of Memory and Language*, 39, 3, 347-370. DOI: 10.1006/jmla.1998.2571.
72. Plaut, David and Laura Gonnerman (2000), „Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?", *Language and Cognitive Processes*, 15, 4-5, 445-485. DOI: 10.1080/016909600501196 61.
73. Rumelhart, David and James McClelland (1986), „On learning the past tenses of English verbs", In James McClelland, David Rumelhart and The PDP Research Group

(Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition: II. Psychological and Biological Models* (pp. 216-271), Cambridge: MIT Press.
74. Seidenberg, Mark and Laura Gonnerman (2000), „Explaining derivational morphology as the convergence of codes", *Trends in Cognitive Sciences*, 4, 353-361.
75. Sečujski, Milan i Aleksandar Kupusinac (2009), „Poređenje postupaka automatske morfološke anotacije tekstova na srpskom jeziku", *XVII Telekomunikacioni forum TELFOR*. Beograd: Društvo za telekomunikacije. Retrieved from http://2009.telfor.rs/files/radovi/ 09_44.pdf.
76. Skousen, Royal (1989), *Analogical modeling of language*, Dordrecht: Kluwer Academic Publishers.
77. Skousen, Royal (1992), *Analogy and structure,* Dordrecht: Kluwer Academic Publishers.
78. Skousen, Royal (2002), Introduction. In Royal Skousen, Deryle Lonsdale and Dilworth Parkinson (Eds.), *Analogical modeling: An exemplar-based approach to language* (pp. 1-8), Amsterdam: John Benjamins.
79. StatSoft Inc. (2004). STATISTICA system reference version 7.0. StatSoft Inc, Tulsa.
80. Steinwart, Ingo and Andreas Christmann (2008), *Support vector machine*s, New York: Springer.
81. Storkel, Holly (2001), „Learning new words: Phonotactic probabilities in language development", *Journal of Speech, Language, and Hearing Research*, 44, 6, 1321-1337. DOI: 10.1044/1092-4388(2001/103).
82. Storkel, Holly (2004), „The emerging lexicon of children with phonological delays", *Journal of Speech, Language, and Hearing Research*, 47, 5, 1194-1212. DOI: 10.1044/1092-4388(2004/088).
83. Storkel, Holly, Jonna Armbrüster and Tiffany Hogan (2006), „Differentiating phonotactic probability and neighborhood density in adult word learning", *Journal of Speech, Language, and Hearing Research*, 49. DOI: 1175-1192 10.1044/1092-4388(2006/085).
84. Storkel, Holly and Michele Morrisette (2002), „The lexicon and phonology: Interactions in language acquisition", *Language, Speech, and Hearing Services in Schools*, 33, 1, 24-37. DOI: 10.1044/0161-1461(2002/003).
85. Storkel, Holly and Margaret Rogers (2000), „The effect of probabilistic phonotactics on lexical acquisition", *Clinical Linguistics and Phonetics*, 14, 407-425.
86. van der Kleij, Sanne, Judith Rispens and Annette Scheper (2016), „The effect of phonotactic probability and neighbourhood density on pseudoword learning in 6- and 7-year-old children", *First Language*, 36, 2, 93-108.
87. van Gestel, Tony, Johan Suykens, Bard Baesens, Stijn Viaene, Jan Vanthienen, Guido Dedene, Bart De Moor and Joos Vandewalle (2004), „Benchmarking least squares support vector machine classifiers", *Machine Learning*, 54, 1, 5-32. DOI: 10.1023/B:MACH.0000008082. 80494.e0.

88. Vapnik, Vladimir (1995), *The nature of statistical learning theory,* New York: Springer-Verlag New York.
89. Vapnik, Vladimir (1998), *Statistical learning theory*, New York: John Wiley and Sons, Inc.
90. Vert, Jean-Philippe, Koji Tsuda and Bernhard Schölkopf (2004), „A primer on kernel methods", In Bernhard Schölkopf, Koji Tsuda and Jean-Philippe Vert (Eds.), *Kernel methods in computational biology* (pp. 35-70), Cambridge: MIT Press, A Bradford Book.
91. Vitevitch, Michael (2003), „The influence of sublexical and lexical representations on the processing of spoken words in English", *Clinical Linguistics and Phonetics*, 17, 6, 487-499. DOI: 10.1080/0269920031000107541.
92. Vitevitch, Michael and Paul Luce (1998), „When words compete: Levels of processing in perception of spoken words", *Psychological Science*, 9, 4, 325-329. DOI: 10.1111/1467-9280.00064.
93. Vitevitch, Michael and Paul Luce (1999), "Probabilistic phonotactics and neighborhood activation in spoken word recognition", *Journal of Memory and Language*, 40, 3, 374-408. DOI: 10.1006/jmla.1998.2618.
94. Vitevitch, Michael and Paul Luce (2004), „A web-based interface to calculate phonotactic probability for words and nonwords in English", *Behavior Research Methods, Instruments and Computers*, 36, 3, 481-487. DOI: 10.3758/BF03195594.
95. Vitevitch, Michael, Paul Luce, Jan Charles-Luce and David Kemmerer (1997), „Phonotactics and syllable stress: Implications for the processing of spoken nonsense words", *Language and Speech*, 40, 47-62. DOI: 10.1177/002383099704000103
96. Vitevitch, Michael, David Pisoni, Karen Iler Kirk, Marcia Hay-McCutcheon and Stacy Yount (2002), "Effects of phonotactic probabilities on the processing of spoken words and nonwords by postlingually deafened adults with cochlear implants", *Volta Review*, 102, 283-302.
97. Ziegler, Johannes, Mathilde Muneaux and Jonathan Grainger (2003), „Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation", *Journal of Memory and Language*, 48, 4, 779-793. DOI: 10.1016/S0749-596X(03)00006-8.

Originalni naučni članak
UDK 811.163.41`366:811.163.41`282
DOI 10.21618/fil1919036d
COBISS.RS-ID 8214040

Strahinja B. Dimitrijević
Univerzitet u Banjoj Luci
Filozofski fakultet
Laboratorija za eksperimentalnu psihologiju

ULOGA FONOTAKTIČKIH INFORMACIJA U RAZUMIJEVANJU I PRODUKCIJI FLEKTIVNE MORFOLOGIJE

Rezime

U dvije odvojene studje provjeravano je u kojoj mjeri se naš kognitivni sistem može osloniti na fonotaktičke informacije, tj. dozvoljene kombinacije fonema/grafema, u zadacima automatske percepcije i produkcije riječi u jezicima sa bogatom flektivnom morfologijom, kao što je srpski jezik. U prvoj studiji, uz pomoć mašina sa vektorima podrške, obavljena je diskriminacija promjenljivih vrsta riječi na osnovu bigrama i trigrama generisanih na nivou morfosintaksičkih kategorija. U drugoj studiji izvedena je produkcija infleksionih oblika uz pomoć učenja zasnovanog na memoriji, oslanjanjem na fonotaktičke informacije iz posljednja četiri sloga osnovnih oblika riječi (lema). Maksimalna tačnost diskriminacije promjenljivih vrsta riječi dobijena je na osnovu bigrama i kretala se oko 93%. Slično, u zadatku infleksione produkcije na svim promjenljivim vrstama riječi uzetim zajedno, ispravno je generisano oko 92% infleksionih oblika. Potvrđena je značajna uloga fonotaktičkih informacija u percepciji i produkciji morfološki složenih riječi, te potreba da se ove informacije uzmu u obzir kada se razmatra pojavljivanje većih jezičkih jedinica i obrazaca. Rezultati pokazuju da funkcionalne veze ortografije/fonologije i semantike omogućavaju uspješno leksičko učenje, ali istovremeno dovode u pitanje potrebu za postojanjem mentalnog leksikona u kojem bi bile uskladištene mentalne predstave najrazličitih jezičkih karakteristka.

Ključne riječi: fonotaktičke informacije, infleksiona morfologija, diskriminacija vrsta riječi, učenje zasnovano na memoriji, mašine sa vektorima podrške.

Preuzeto 9. 7. 2018.
Korekcije 22. 1. 2019.
Prihvaćeno 11. 2. 2019.