PSIHOLOGIJA, 2003, Vol. 36 (3), 255-270 UDK 159.95.072:81'366

ANNOTATED CORPUS AND THE EMPIRICAL EVALUATION OF PROBABILITY ESTIMATES OF GRAMMATICAL FORMS

Nada Ševa and Aleksandar Kostić[1]

Laboratory for Experimental Psychology,
University of Belgrade, Serbia and Montenegro

The aim of the present study is to demonstrate the usage of an annotated corpus in the field of experimental psycholinguistics. Specifically, we demonstrate how the manually annotated Corpus of Serbian Language (Kostić, Đ. 2001) can be used for probability estimates of grammatical forms, which allow the control of independent variables in psycholinguistic experiments. We address the issue of processing Serbian inflected forms within two subparadigms of feminine nouns. In regression analysis, almost all processing variability of inflected forms has been accounted for by the amount of information (i.e. bits) carried by the presented forms. In spite of the fact that probability distributions of inflected forms for the two paradigms differ, it was shown that the best prediction of processing variability is obtained by the probabilities derived from the predominant subparadigm which encompasses about 80% of feminine nouns. The relevance of annotated corpora in experimental psycholinguistics is discussed more in detail .

Key words*: annotated corpora, inflected morphology, psycholinguistics*

The usage of corpora in experimental psycholinguistics is not limited to mere stimulus selection that will enable balanced factorial designs. Annotated corpora can also be used for a quantitative specification of independent variables, i.e. for probability estimates of various aspects of language such as word probability, grapheme and syllable probability, the probability of inflected forms etc. Once specified in terms of probability, an independent variable used in psycholinguistic experiments can be correlated with the subject's performance, thus providing us with deeper insights about the mechanisms responsible for language processing.

One of the main issues of experimental psycholinguistics is related to the processing of inflective morphology. How do we process words with inflectional

[1] Address of the authors: akostic@f.bg.ac.yu

suffixes, is a question related to more profound problems of cognitive mechanisms engaged in the processing of syntax. Standard paradigm in this kind of research is the lexical decision task. Participants are presented with a string of letters and their task is to answer (by pressing the yes/no key) whether the presented string is a word or not. The dependent variable is reaction time (expressed in milliseconds [ms]) measured from the stimulus onset.

A number of studies has shown that the cognitive system is sensitive to the probability of individual words: the higher probability of a word is paralleled by the shorter processing latency. By implication, it could be assumed that the same is true for inflected forms of the same word as well, i.e. the higher the probability of a suffix, the shorter the time required for processing. Research of this kind was primarily done in English and to some extent in German, French Dutch, Hebrew and Serbian. Principal models, however, are based on data obtained in English which is a language with few inflections. As a consequence, standard models of morphological processing assume that the processing of an inflected form is affected primarily by the affix frequency (Manelis, Tharp, 1977; Taft, Foster 1975; Taft, 1981; Kampley, Morton 1982; Cutler 1983; Henderson 1985; Allen, Badecker, 1999; Allegre, Gordon, 1999; Frost, Deutch, 2000). In the research referred to above, corpora (i.e. frequency dictionaries derived from corpora) were used for two kinds of probability estimates: a. word frequency and b. suffix frequency. Suffix frequency is standardly estimated independently of the word type or grammatical status of a word within a given type. This was due to the fact that frequency dictionaries used for probability estimates were compiled from corpora characterized by coarse grammatical annotation.

The aim of the present study is therefore two-fold. On one hand, it addresses specific issues of processing inflective morphology of Serbian which is a highly inflected language. On the other hand, its aim is to demonstrate the advantage of a corpus that is manually annotated up to the level of inflected morphology.

A GENERAL OVERVIEW OF THE SERBIAN NOUN SYSTEM

Serbian is a highly inflected and to a great extent a free word-order language. Six out of ten word types in Serbian can appear in some of the following grammatical categories: case, number, gender, comparative, tense, person, mood, etc., each of them being marked by an inflectional suffix. An inflectional suffix added to the noun stem indicates a case, number and gender. Serbian nouns appear in seven cases, both singular and plural.2 The noun gender, on the other hand, is an

[2] Vocative case is considered by most linguists as a case with no syntactic function. Therefore, the case taxonomy presented here will not include vocative case.

intrinsic property of a noun, thus a given noun can appear in one gender only. Masculine, feminine and neuter nouns have their specific paradigms. As a consequence, case endings for masculine, feminine and neuter nouns differ. An example of the declension of Serbian feminine nouns is given in the Appendix A (Table A).

An inspection of the declension of Serbian nouns (Table A) indicates that there are several homomorphs, i.e. some cases, both singular and plural, share the same inflectional suffix. As a consequence, in a lexical decision task, a word presented in isolation may be equivocal with respect to its case. Thus, what can be experimentally manipulated is an *inflected form*, that may contain several cases. Masculine nouns appear in seven morphologically distinct forms, feminine nouns appear in six distinct forms, while neuter nouns appear in five distinct forms. Each grammatical form appears with some probability. By the same token, probabilities of inflected forms that encompass several cases also vary (Table 1).

Table 1: Unique inflected forms of feminine nouns and their probabilities

Form	**Cases**	**F%**
vod-a	nom. s. + gen. pl.	12.061
vod-e	gen. s. + nom. pl. + acc. pl.	14.201
vod-i	dat. s. + loc. s.	3.796
vod-u	acc. s.	5.480
vod-om	ins. s.	1.939
vod-ama	dat. pl. + ins. pl. + loc. pl.	1.690

As noted earlier, suffix probability is considered to be the pivotal factor that influences the processing latency of inflected word forms. In order to investigate the processing of Serbian inflected nouns, suffix probabilities were estimated from the Corpus of Serbian Language (CSL) and its derivatives (Kostić, Đ, 1965; 1999; 2001). Since the Corpus was the base that provided us with probability controls, in the forthcoming paragraphs it will be described in more detail.

THE CORPUS OF SERBIAN LANGUAGE (CSL)

The Corpus of Serbian Language (CSL) contains 11 million words, and spans Serbian language from the 12th century to contemporary language. Each word in

the corpus is manually annotated at the level of inflected morphology. The system of annotation distinguishes about 2000 different grammatical forms. The Corpus was built up by Prof. Đorđe Kostić in the mid fifties at the Institute for Experimental Phonetics and Speech Pathology in Belgrade, as a part of a broader project aimed at automatic speech and text recognition and machine translation. The work on the CSL was initiated in 1957 and lasted till 1962. About 400 collaborators (80 experts in linguistics and other related fields, together with more than 300 technical staff) participated on the CSL project. In 1996. the whole material was transferred into an electronic format. The pilot version of The Frequency Dictionary of Contemporary Serbian Language was compiled from the sample of 2 million words (subsamples of daily press and contemporary Serbian poetry) (Kostić, Đ. 1999). The Dictionary contains about 65 000 lemmata and about 240 000 grammatical forms.3 In the mid-sixties, Prof. Đorđe Kostić published several studies based on corpus materials, the most prominent one being "The probabilities of Grammatical Forms in Serbo-Croatian" (Kostić, Đ. 1965). The electronic version of the Corpus and the pilot version of the Frequency Dictionary allow almost unlimited probability estimates, ranging from the probabilities of individual words and their grammatical forms to the probabilities of phonological and syllabic structures.

EXPERIMENTS WITH SERBIAN INFLECTED NOUN FORMS

In a number of lexical decision type experiments, the processing of Serbian masculine, feminine and neuter nouns had been investigated (Kostić, A. 1991; 1995; 2001; 2003). Due to the fact that each Serbian noun can appear in a number of morphologically distinct forms, the first step was to estimate the probability of inflectional suffixes. This proved to be a nontrivial task, because the probability of a suffix can be specified at a number of levels. Take, for example, the suffix "i". If attached to a verb, it specifies the third person singular present tense. If attached to an adjective, it specifies the nominative plural masculine gender, if attached to a noun it specifies a dative and locative singular feminine noun and a nominative plural masculine noun. In other words, the suffix per se is equivocal, with respect to the grammatical form. Consequently, there is a number of possible probability estimates: a. irrespective of a word type, b. with respect to a word type (e.g. suffix "i" attached to noun), c. with respect to a defined paradigm within a given word type (e.g. the probability of suffix "i" to be attached to a feminine noun) and d. the probability of a suffix x attached to a word y of type z. Prior to designing experiments, probabilities for each suffix were estimated at all levels.

[3] More information about this project is available at **www.serbian-corpus.edu.yu**.

In a series of lexical decision experiments, all inflected forms of nouns of a given gender were presented to the participants (Kostić, A. 2003, submitted). The dependent variable was the reaction time expressed in milliseconds (ms). The analyses were performed on mean reaction times for each presented form. In other words, if we present six inflected forms in an experiment, the analysis of variance indicates whether there is a main effect of form, i.e. whether there is a systematic factor responsible for differences in mean reaction time among the six presented forms. However, if we want to estimate the effect of the form's probability on the processing latency, the regression analysis is required. The regression analysis allows us to estimate the proportion of explained variance of reaction time due to the variation in the form's probability. In other words, in order to evaluate which level of suffix probability specification is the one which our cognitive system is sensitive to, we have to correlate different probability estimates, which we referred to earlier, with mean reaction times to inflected noun forms.

The analyses had shown that the highest correlation between mean reaction time and suffix probabilities was obtained for probabilities specified within a defined paradigm for a given word type (e.g. probability of suffix "i" attached to a feminine noun), the implication being that the cognitive system is not sensitive to the probability of a suffix per se. This was true for nouns of all three genders. While for feminine nouns there was a significant correlation, for masculine and neuter nouns correlation did not reach significance. In order to increase the proportion of explained variability of processing latency to inflected noun forms, an additional parameter, namely, the number of syntactic functions and meanings covered by an inflected noun form of a given paradigm, was introduced (Kostić, Đ. 1965b) (see Table A, Appendix A). If we divide the frequency of an inflected form by the number of its syntactic functions and meanings, the obtained unit is *the average frequency per syntactic function/meaning for a given inflected form*. Since the frequency specifies probability, the new unit can be expressed in terms of the amount of information (bits) carried by an inflected noun form. In order to obtain the amount of information, the average frequency per function/meaning for a given noun form should be expressed as a proportion, relative to a sum of average proportions per function/meaning for other noun forms for a given gender. In order to express it in terms of the amount of information, the obtained proportion must undergo a log transform. This will provide us with bits carried by each grammatical form (Equation 1).

$$I_m = \left[-\log_2 \left(\frac{\frac{F_m}{R_m}}{\sum_{m=1}^{M} \frac{F_m}{R_m}} \right) \right] \qquad (1)$$

In Equation 1, I stands for the amount of information carried by an inflected noun form (m), F stands for the frequency of a form, and R stands for the number of functions/meanings encompassed by a form. The obtained unit is the amount of information derived from the average frequency per function/meaning for a given noun form. This descriptor refers to the relative complexity of a noun form: the higher the value of I, the higher the complexity of a form. Consequently, the increase in the amount of information, should be paralleled by the increases in the processing time (Kostić, A. 1991; 1995; 2003, submitted).

Values derived from Equation 1 accounted for almost all of the processing time variability (of) inflected forms of Serbian feminine nouns (Kostić, A., 2003, submitted). Likewise, extremely high proportion of explained processing variability was observed for masculine and neuter nouns as well. For all three genders r^2 varies between .93 and .98. The fact that almost all of the processing variability has been accounted for by the amount of information as specified by Equation 1 indicates, that both, form's probability and the number of syntactic functions/meanings carried by a form, affect the processing latency of Serbian inflected noun forms.

THE PROCESSING OF SUBPARADIGMS OF FEMININE NOUNS

In the summarized experiments, only nouns typical for the respective gender paradigm were presented. The selection criterion was intuitive and stimulus material consisted of nouns of type "voda" (water) only (see Table 1). However, within the regular feminine paradigm there are two types of exceptions:

a. The change in the distribution of the same suffixes to different cases. Noun *voda* (water), which is a typical example of the paradigm of regular feminine nouns, ends with the suffix "a" in genitive plural. In contrast, nouns like *bajka* (fairytale) and *ruka* (hand) end with the suffix – "i" (bajk-i) and – "u" (ruk-u) in genitive plural.

b. The second exception is related to the voice alternations which appear in the stem of the word. There are two types of voice alternations within this paradigm: sibilarisation and fleeting -A.

Cross-reference of these two types of exceptions generates 10 different classes (subparadigms) of regular feminine nouns. A similar type of exceptions can be seen for nouns of masculine and neuter gender as well. Case alternations for different classes of regular feminine nouns are presented in the Appendix A (Table C). Note that within a single paradigm of regular feminine nouns there are differences in the suffix distribution across different cases with respect to specific subparadigms. This implies that specification of the suffix probability within a given paradigm, may not be as simple as assumed. Not only is the same suffix shared across different word types, it is also shared by different cases within the same noun paradigm. In other words, there is no consistent suffix-to-case mapping within a paradigm of regular feminine nouns.

In order to determine the probability distribution of the respective classes of regular feminine nouns, 2125 lemmata were taken from the Corpus of Serbian Language and the pilot version of the Frequency Dictionary of Contemporary Serbian Language (Kostić, Đ. 1999; 2001). An inspection of Table B in the Appendix A indicates that the nouns of paradigm "voda", which was used in the summarized experiments with feminine nouns, is the dominant subparadigm for regular feminine nouns.

Having the above properties of feminine nouns in mind, the question is, what is the proper specification of case (i.e. inflected form) probability that our cognitive system is sensitive to? The fact that probability distribution of inflected forms is class dependent, may imply that the processing latency patterning of inflected forms will differ with respect to class. If so, the implication is that the cognitive system is sensitive to probability distribution within subparadigms. On the other hand, if no difference is observed, it may imply that the cognitive system is sensitive to global probability distribution, i.e. the one derived from all feminine nouns in the Corpus. Note that due to the fact that dominant subparadigm encompasses almost 80% of feminine nouns (see Table B in the Appendix A), the global probability distribution is predominantly affected by this subparadigm.

Experiment

In order to evaluate the hypotheses above, three inflected forms of nouns of subparadigm "voda" and three forms of nouns of subparadigm "bajka" were presented in a lexical decision experiment. The difference between the three forms of paradigm "voda" and paradigm "bajka" is presented in Table C in the Appendix A.

Method

Stimuli and procedure: 30 nouns of type "bajka" (fairytale), 30 nouns of type "voda" (water) and 60 pseudonouns were presented in 3 forms, ending with suffixes A, E and AMA. Stimuli were exposed for 1500 ms on a computer screen (PC PentiumII). The participant's task was to answer as quickly and as accurately as possible (by pressing yes/no keys) whether the presented string of letters is a word or not.

Participants: 45 first-year undergraduate students from the Department of Psychology, University of Belgrade, participated in the experiment as part of their academic requirements. Subjects were divided into three groups.

Results

The mean reaction time and normalized reaction time for three forms of the two paradigms of nouns is presented in Table 2.

Table 2: Mean reaction time and normalized reaction time for three inflected forms of the two paradigms of feminine nouns.

Form	**RT (ms)**	***Form***	**Normalized RT (ms)**
bajka	667	bajka	643
bajke	669	bajke	652
bajkama	712	bajkama	701
voda	639	voda	641
vode	648	vode	654
vodama	683	vodama	701

The analysis of variance, performed on the subjects' mean response latencies, indicated a significant main effect of noun type: $F(1, 44)=62.527$ $p<.001$; nouns of type "voda" were processed faster than nouns of type "bajka". Also, there was a significant effect of noun form: $F(2, 88)=42.463$, $p<.001$; some forms, irrespective of noun type, were processed faster than others. There was no significant form by type interaction, indicating no difference in the patterning of response latencies to inflected forms between two types of nouns.

In order to evaluate whether the cognitive system is sensitive to probability distribution relative to subparadigm or to probability distribution of the dominant subparadigm, a single regression analysis on all six mean reaction times has to be applied. Mean reaction times will be correlated with two distinct probability counts. The one that is specific to subparadigms (I1), and the other derived from the paradigm of feminine nouns, applied to both word types (I2) (see Table 3).

Table 3: Inflected forms, cases, and two ways (I1 and I2) of specifying the amount of information carried by three forms of the two subparadigms of regular feminine nouns.

Inflected forms	**Cases**	**I 1**	**I 2**
voda	n.s. +g.p.	1.464	1.464
vode	g.s. + n.p. +a. p.	2.280	2.280
vodama	d.p.+l.p. +i. p.	4.773	4.773
bajka	n.s.	0.207	1.464
bajke	g.s. +n.p. +a.p.	4.746	2.280
bajkama	d.p.+l.p. +i.p.	7.239	4.773

Since there was a significant main effect of noun type, all mean reaction times for different forms in each group have to be normalized to a grand mean (see Table 2). When normalized, mean reaction times for the three inflected forms of the two groups (6 points) were regressed on the amount of information specified with respect to subparadigm (I1) 54% of processing variability for both noun types was accounted for by the informational values: $r^2 = .535$, $F(1,4)=6.7451$ $p<.06$. However, when the normalized RTs were regressed on the amount of information derived from the paradigm of feminine nouns (I2, type "voda"), the amount of explained variance was increased up to 0.97, $F(1,4)=156.59$, $p<.001$ (Figure 1).

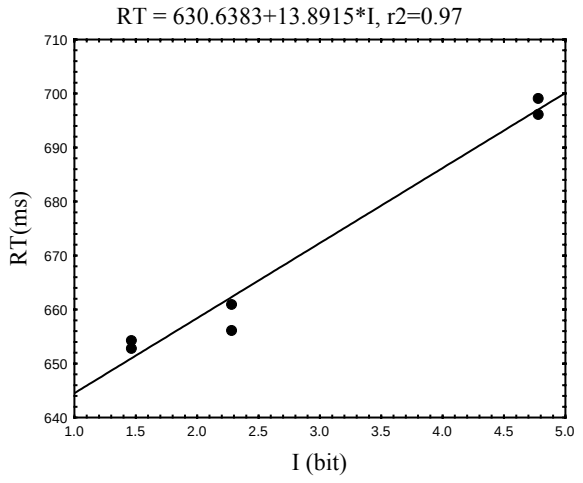

Figure 1: The relation between RTs for three forms of paradigm "bajka" and three forms of paradigm "voda" and the amount of information carried by those forms.

The outcome of the experiment shows that the cognitive system is sensitive to probabilities derived from the dominant subparadigm and insensitive to local

probability distributions defined relative to given subparadigms of nouns of particular gender.

GENERAL DISCUSSION

The outcome of the present experiment shows that the sensitivity of the cognitive system is finely tuned to the probabilities at the level of grammatical forms within a defined paradigm (feminine nouns). It was also shown that the cognitive system is not sensitive to the distinctions within a given paradigm, because a better prediction of processing latency variation was obtained with probabilities derived from a paradigm defined in terms of "feminine nouns" rather than in terms of "feminine nouns of type x".

It should be emphasized that the above insights would not be possible without a corpus *manually* annotated at the level of inflected morphology, where the system of annotation encompasses a vast number of grammatical forms. This statement may need further clarification. Due to homography, automatic annotation of highly inflected languages with an acceptable error margin, may prove to be extremely difficult, if not impossible. Take, for example, the disambiguation of case in Serbian. A serbian adjective with the suffix "im" can have 11 possible interpretations, depending on grammatical number and gender. In order to make a proper case disambiguation, the meaning of a sentence may need to be taken into consideration. This, on the other hand can hardly be done automatically. Likewise, word type specification and additional grammatical disambiguation within a given word type may also imply semantic constraints. Without these specifications, probability estimates are necessarily coarse and, more often than not, erroneous. As already noted, research presented in this study would not be possible if the probabilities at all levels were not available and reliably estimated. Once being available, they become a powerful tool for all kinds of precise controls, that enable a clear choice between plausible alternatives. In that respect, annotated corpora are an indispensable tool for research in experimental psycholinguistics.

REFERENCES

Allegre, M., Gordon, P. (1999). Frequency effects and representational study of regular inflections. *Journal of Memory and Language*, **40**, 41-61.

Allen, M., Badecker, W. (1999). Stem homograph inhibition and stem allomorphy; representing and processing inflected forms in multilevel lexical system. *Journal of Memory and Language*, **41**, 105-123.

Cutler, A. (1983). Lexical complexity and sentence processing. In G.B. Flores d'Arcains&R.Jarvella (Eds.). *The Process of Language Understanding*. New York, Wiley

Frost, R., Deutch, A. (2000). Decomposing morphologically complex words in nonlinear morphology. *Journal of Experimental Psychology* (LMC), **26** (3), 751-765.

Henderson, L. (1985) Towards a psychology of morphemes. In A.W. Ellis (Ed.), *Progress in the psychology of language*. *Vol. I.* London, Lawrence Erlbaum Associates Ltd.

Kempley, M. , Morton, J. (1982). The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology,* **73**, 441-454.

Kostić, A. (1991). Informational approach to processing inflected morphology: Standard data reconsidered. *Psychological Research*, **53**, (1), 62-70.

Kostić, A. (1995). Informational load constraints on processing inflected morphology. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing*. New Jersey, Lawrence Erlbaum, Inc., Publishers.

Kostić, A. (2003). *The effects of the amount of information on processing of inflected morphology* (submitted).

Kostić, Đ. (1965a). *Struktura upotrebne vrednosti gramatičkih oblika u srpskohrvatskom jeziku.* (Probability Structure of Grammatical Forms in Serbo-Croatian) Beograd, Institut za eksperimentalnu fonetiku i patologiju govora.

Kostić, Đ. (1965b). *Sintaktičke funkcije padežnih oblika u srpskohrvatskom jeziku* (Syntactic Functions and Meanings of Serbo-Croatian cases) Beograd, Institut za eksperimentalnu fonetiku i patologiju govora.

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika. Tom I – VII.* (Frequency Dictionary of Contemporary Serbian Language. Vol. I – VII.). Institut za eksperimentalnu fonetiku i patologiju govora, Beograd i Laboratorija za eksperimentalnu psihologiju Filozofskog fakulteta u Beogradu.

Kostić, Đ. (2001). *Kvantitativni opis strukture srpskog jezika – Korpus srpskog jezika*. (Quantitative Description of Serbian Language Structure – Corpus of Serbian Language). Institut za eksperimentalnu fonetiku i patologiju govora, Beograd i Laboratorija za eksperimentalnu psihologiju Filozofskog fakulteta u Beogradu.

Manelis, L., Tharp, D. (1977). The processing of affixed words. *Memory and Cognition,* **5** (6), 690-695.

Taft, M., Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behaviour*, **14**, 638-647.

Taft, M. (1981). Prefix stripping revisited. *Journal of Verbal Learning and Verbal Behaviour*, **20**, 289-297.

APPENDIX

Table A: An overview of the Serbian noun system: an example for a feminine noun, frequency for each case in all paradigms (F%) and the number of function and meanings (R)

Case	Number				Number of functions and meanings (R)
	Singular		Plural		
	Form[a]	F%[4]	Form	F%	
nominative	vod-a	8.841	vod-e	3.577	3
genitive	vod-e	7.876	vod-a	3.220	51
dative	vod-i	0.377	vod-ama	0.157	22
accusative	vod-u	5.480	vod-e	2.570	58
instrumental	vod-om	1.939	vod-ama	0.734	32
locative	vod-i	3.419	vod-ama	0.799	21

Table B: Subparadigms of regular feminine nouns and their probabilities

	a.	b.	c.	d.	e.	f.	
voda	*						77.929
jabuka	*				*		5.035
zemlja	*					*	0.847
devojka	*				*	*	0.423
pretnja		*					12.659
bajka		*			*		2.353
ruka			*				0.094
kretnja				*			0.282
tetka				*	*		0.282
olovka				*			0.094

voda – water; jabuka – apple; zemlja – land; devojka – girl; pretnja – threat; bajka – fairytale ; ruka – hand; kretnja – movement; tetka – aunt; olovka – pencil.

(a) gen. pl. – A; (b) gen. pl. – I; (c) gen. pl. – U; (d) double form of gen. pl. (A nad I); (e) sibilarization; (f) fleeting A.

[4] Frequency values (F%) and number of functions and meanings adapted from Kostić, Đ. (1965a, 1965b).

Table C: The amount of information (I) carried by inflected forms of nouns belonging to subparadigms "voda" i "bajka"

Inflected forms	Cases	I	Inflected forms	Cases	I
voda[a]	**n. s. +g. p.**	1.464	**bajka**	**n. s**	0.207
vode	g. s. +n. p. +a. p.	2.280	bajke	g. s. +n. p. +a. p.	4.746
vodi	**d. s. +l. s.**	2.803	**bajci**	**d. s. +l. s.**	5.269
			bajki	**g. p**.	5.752
vodu	a. s.	2.705	bajku	a. s.	5.171
vodom	i. s.	3.346	bajkom	i. s.	5.811
vodama	d. p. +l. p. +i. p.	4.773	bajkama	d. p. +l. p. +i. p.	7.239

ABSTRACT

ANOTIRANI KORPUS I EVALUACIJA PROCENE VEROVATNOĆA GRAMATIČKIH OBLIKA

Nada Ševa i Aleksandar Kostić

U radu je demonstrirana upotreba anotiranog jezičkog korpusa u oblasti eksperimentalne psiholingvistike. Pokazano je kako je ručno anotirani Korpus srpskog jezika (Kostić, Đ., 2001) moguće koristiti za pocenu verovatnoća gramatičkih oblika koje služe kao kontrola nezavisnih varijabli u psiholingvističkim eksperimentima. Upotreba Korpusa demonstriarana je u okviru eksperimenta sa zadatkom leksičke odluke u kome je ispitana obrada dve paradigme imenica ženskog roda u srpskom jeziku, od kojih je jedna dominantna (80% imenica ženskog roda pripada toj paradigmi). Standardni nalazi vezani za obradu infleksione morfologije pokazuju da je frekvenca inflektivnog oblika jedan od dominatnih činilaca koji utiče na vreme njegove obrade. Iako se dve paradigme razlikuju po distribuciji verovatnoća svojih inflektivnih oblika, nalazi ogleda pokazuju da je vreme obrade determinisano verovatnoćama dominantne paradigme. Pošto kontrola nezavisne varijable u ovakvoj vrsti eksperimenata nije moguća bez prethodnog uvida u verovatnoće gramatičkih oblika, posebno je razmotren značaj korpusa anotiranih na nivou inflektivne morfologije.