ALLOMORPHIC RESPONSES IN SERBIAN PSEUDO-NOUNS AS A RESULT OF ANALOGICAL LEARNING*

PETAR MILIN[a,c] – EMMANUEL KEULEERS[b] –
DUŠICA FILIPOVIĆ ĐURĐEVIĆ[a,c]

[a]Department of Psychology, University of Novi Sad, Serbia
[b]Department of Experimental Psychology, Ghent University, Belgium
[c]Laboratory for Experimental Psychology, University of Belgrade, Serbia

Corresponding author: Petar Milin, address: Dr Zorana Dindica 2,
Novi Sad 21000, Serbia, e-mail: pmilin@ff.uns.ac.rs

Abstract: Allomorphy is a phenomenon that occurs in many languages. Several psycholinguistic studies have shown that allomorphy, if present, co-determines cognitive processing. In the present paper we discuss allomorphic variations of Serbian instrumental singular form of pseudo-nouns as emerging from analogical learning. We compare the predictions derived from memory-based language processing models with results from a previous experimental study with adult Serbian native speakers. Results confirm that the production of suffix allomorphs in Serbian instrumental singular masculine nouns can be accounted for by memory-based learning and simple analogical inferences. The present findings are in line with a growing body of research showing that memory-based learning models make relevant predictions about the cognitive processes involved in various linguistic phenomena.

Keywords: allomorphy, memory-based learning, analogy, *wug* test, Serbian

1. Introduction

In this paper we will present a probabilistic computational model of allomorphy and demonstrate that allomorphic variation may arise from

---

analogical learning of the mapping from stems to inflected forms. We will make use of behavioral experiments that were previously conducted with adult native speakers of Serbian engaged in a computerized *wug* test (Jovanović 2008; see Berko 1958 for the original *wug* test experiment). Looking at the two allomorphic forms of the instrumental singular of Serbian masculine pseudo-nouns, we will compare the performance of native speakers with the outcomes of several simulations using computational models of analogical learning.

Allomorphy represents a variation in the form of a particular morpheme, without a change in its meaning (cf. Lieber 1982; Lyons 1968; Spencer 2001, etc.). In English, variations in the *-ed* morpheme used in the regular past tense, and the *-s* morpheme used to mark noun plurals, are well-known examples. The regular English past tense suffix appears in three different forms (or **morphs**), depending on the final sound of the verbal stem: *walk-ed* (/t/), *jogg-ed* (/d/), *trott-ed* (/ɪd/). In modern Arabic, allomorphy occurs in an **etymon**—a bi-consonantal morphological unit that carries semantic information of a given word (Ratcliffe 1998; Boudelaa–Marslen-Wilson 2001; 2004). In Dutch, the diminutive suffix has two frequent allomorphic variations (*-tje* and *-je*), and three less frequent ones (*-etje, -pje* and *-kje*) (Daelemans et al. 1997). In Finnish, allomorphy appears both in the stem (Järvikivi–Niemi 2002b) and in suffixes (Järvikivi et al. 2006). Similarly, in Hungarian, allomorphic variations occur as stem shortening or lengthening (Pléh et al. 2002), and as suffixal alternating vowels (Kertész 2003; Hayes–Cziráky Londe 2006). Finally, allomorphy is present in Slavic languages as well. Affixal allomorphy in Russian is discussed in detail by Blevins (2004), while Ivić (1990) and Zec (2006) provided linguistic analysis of the suffix allomorphy in Serbian instrumental singular masculine and neuter nouns.

1.1. Allomorphy as a cognitive phenomenon

For cognitive science, and in particular for psycholinguistics, the main question of any language phenomenon is its cognitive relevance. If a particular linguistic phenomenon can also produce critical differences in behavioral and/or neurological measures, then one can say that the linguistic phenomenon also has cognitive relevance. Although often not of central interest, the cognitive relevance of allomorphy has repeatedly been attested in behavioral research. Schreuder and Baayen (1995) stated that we may be slower in processing words with affixes that have several

allomorphs, than words containing affixes for which there is no allomorphic variation. Järvikivi et al. (2006) made a similar, but more detailed claim, using the concept of **affixal salience**—"the probability with which an affix is likely to emerge from the orthographic/phonological string" (*ibid.*, 395). They showed that affixal salience decreases as the number of affixal allomorphs increases. Conversely, however, to the inhibition that allomorphy produced to a given affix, allomorphic realizations of bound nominal stems in Finnish significantly primed the same noun in its base form—nominative singular (Järvikivi–Niemi 2002a;b). Similarly, in a priming task in Spanish, Allen and Badecker (1999) found no difference between conditions in which the prime was a true stem-homograph of the target (e.g., *placer* 'pleasure', 'to please (inf.)'–*placa* 'plate', 'panel') and conditions in which the target was preceded by a stem allomorph of the prime (e.g., *plazc*a 'to please (subjunctive, 3sg)'–*placa* 'plate', 'panel'). Finally, specific difficulties in processing allomorphic variations in Hungarian nouns were observed with normal children (Pléh 1989), and with children with Williams syndrome (Pléh et al. 2002).

One of the most common instances of allomorphy in Serbian is the suffix allomorphy (*-em* vs. *-om*) occurring for instrumental ('making use of') singular masculine nouns. For instance, Serbian native speakers may be somewhat puzzled whether to say *nos-om* or *nos-em* 'using the nose', *malj-om* or *malj-em* 'using an odor', *obruč-om* or *obruč-em* 'using a hoop', *pištolj-om* or *pištolj-em* 'using a revolver', and so on. Jovanović et al. (2008) directly addressed this form of allomorphy using two experimental tasks. First, using a **sentence completion task**, the authors confirmed that suffix allomorphy in Serbian instrumental singular masculine nouns occurred only when a noun stem ended in a particular subset of consonants: **palato-alveolars** or **back coronals**.[1] Second, using a visual lexical decision task, they showed that suffix allomorphy in Serbian masculine nouns with stems ending in back coronals elicits significant differences in processing latencies: for words with the *-om* suffix, an increase in observed form frequency was associated with an increase in processing latency, while for the *-em* suffix, an increase in form frequency was associated with a decrease in reaction time. This interaction between a particular suffix realization (*-om* or *-em*) and its probability in produc-

---

[1] Different subset labels come from two means of consonant classification. Front coronals match alveolars: *n* (/n/), *l* (/l/), and *r* (/r/) and include five additional consonants: *t* (/t/), *d* (/d/), *s* (/s/), *z* (/z/) and *c* (/ts/). Back coronals match palato-alveolars: *č* (/tʃ/), *ć* (/tɕ/), *dž* (/dʒ/), *đ* (/dʑ/), *nj* (/ɲ/), *lj* (/ʎ/), *j* (/j/), *š* (/ʃ/), and *ž* (/ʒ/).

tion task showed that even though *-om* is the most frequent suffix in the Serbian instrumental singular, it is processed slower if encountered within the phonological domain for which *-em* is preferred. Although such complementarity might suggest rule-based derivation of the two allomorphic forms, we will advocate that this pattern can emerge from a more parsimonious learning principle.

1.2. Modeling allomorphic response as analogical learning

Jovanović and her collaborators (2008) discussed their results in respect to previous findings of Mirković et al. (2011),[2] who used a connectionist network to model the production of Serbian case-inflected morphology. This model used a training set of 3244 Serbian nouns, and learned to produce the correct case-endings by developing particular probabilistic constraints at the level of phonology and semantics. At the end of learning, the error rate for masculine instrumental singular nouns, our target case, was still approximately 4%. However, the model excluded the possibility of having both suffixes applied to the same stem with different probabilities, but implemented a simple rule that attached either *-om* or *-em* to a given stem. For instance, all masculine nouns with a stem ending in an alveolar or palato-alveolar consonant, necessarily took the *-em* suffix, while all nouns with other terminating consonants used *-om* instead (Mirković et al. 2011).

In contrast, a study by Zec (2006) showed that masculine noun stems ending with a coronal can, and usually do, have allomorphic realizations in the instrumental singular: both *-om* and *-em* can apply. Jovanović et al. (2008) and Jovanović (2008) confirmed the analysis of Zec (2006), both in a lexical decision task and in a computerized modification of the *wug* test (Berko 1958), administered to adult native speakers of Serbian. More specifically, masculine nouns ending in back coronals (or palato-alveolars) were significantly more likely to allow for both suffix allomorphs (*-om* and *-em*).

In principle, connectionist networks should be capable of modeling allomorphy. In particular, a probabilistic version of the model of Mirković et al. (2011) could account for allomorphic variation in Serbian nouns. However, the immense power or flexibility that is typical for artificial

---

[2] Jovanović et al. (2008) based their comments on a 2005 manuscript by Mirković et al. that was available to them at that time. The paper by Mirković et al. (2011) referred to in the text is a later version of the same manuscript.

neural networks, comes at the cost of a lack of insight in how a given network achieved a particular morphological mapping. As Norris (2005) suggested, the true contribution of connectionist models should not come from their performance, but from understanding the principles that guide the performance of the networks (see also Baayen (2003) for a more elaborate discussion). Thus, the question is whether more directly addressable learning mechanisms could meet the same goal. In particular, we are interested in testing whether we could model allomorphic variation in the Serbian instrumental singular by using a very simple analogical approach. However, before we go any further, a note of caution is in order: it is perfectly possible to successfully model the same phenomenon using different **machine learning** approaches. What is important is the contribution that different approaches give to our understanding of the phenomenon. Following Marr (1982), we can say that analogical learning improves our understanding mostly at the **algorithmic level**, revealing the processes and representations of this task. At the same time, a connectionist network improves our understanding mostly at the **implementational level**, showing how neural structures and neuronal activities might implement a given cognitive task.

Our claim is that allomorphy can take place from analogical inference, where sources of analogy (existing stem forms) compete with each other in providing one or the other suffix allomorph—possible inflected forms of instrumental singular masculine nouns. Acquisition and processing of linguistic knowledge by means of memory and analogy has a long history in twentieth-century linguistics (Saussure [1916] 1966; Bloomfield 1933; Harris 1951; 1957, etc.). Recently, the idea has been further developed in usage-based models of language (Bybee 2007). In psychology, the concept of analogy can be seen in exemplar-based accounts of human categorization behavior (Smith–Medin 1981; Nosofsky 1986; Estes 1994). According to these accounts, categories are formed by storing exemplars in memory, and categorization decisions are made by relying on similarities of target stimuli to exemplars stored in memory. In computational linguistics, these ideas have been applied in memory-based learning (Daelemans–van den Bosch 2005) and Analogical Modeling of Language (Skousen 2002).

According to the memory-based learning approach, a categorization decision (e.g., the choice of allomorph) is resolved by re-use of existing exemplars and analogical reasoning. In order for this process to take place, three conditions need to be fulfilled. In the case we are studying here, firstly, we need a store of exemplars (stems) with assigned exponent (the

instrumental ending). These exemplars can be represented as vectors of phonological features at the subsyllabic level (i.e., the onset, nucleus, coda of each syllable). Secondly, a distance function is required to compute the similarity of the target form to the forms stored in memory. Finally, in order to assign a class to the novel exemplar, a decision function is required. The decision function is adopted from the field of artificial intelligence and is based on the k nearest neighbor classifier method (k-NN). This implies that the outcome of the decision function is determined by the class of the k nearest neighbors (e.g., if $k = 1$, a novel exemplar is assigned a class of the exemplar most similar to it). Memory-based learning has a long history of application within the field of computational linguistics. Recently, the method has also been successfully applied in psycholinguistic research, where the aim is to approach the performance of native speakers, that is, to simulate the functioning of the cognitive system. By now, a considerable body of empirical data demonstrated the efficiency of memory-based learning. Keuleers et al. (2007) and Keuleers–Daelemans (2007) have demonstrated that outcomes of simulations based on the memory-based learning paradigm mimic performance of native speakers in the production of Dutch noun plurals. Similar findings have been reported for Italian verb conjugations (Eddington 2002a), Spanish gender assignment (Eddington 2002b), linking elements in German compounds (Krott et al. 2007), and so on.

1.3. Problem

In this paper, we will compare the predictions derived from memory-based learning models to experimental results by looking at the production of allomorphic variations using pseudo-nouns in the domain of the Serbian instrumental singular. Attempts have been made to describe orthographic/phonological properties of stems that lead to the production of each of the two allomorphic variations (Zec 2006 in particular). These descriptions were moderately successful in predicting responses collected from native speakers, and can be seen as rules for choosing an allomorph. In this study, we will not compare the predictions derived from these rules to the results obtained by means of exemplar-based modeling. Our aim is to demonstrate that analogical learning can account for allomorphic variation at least as well as the rule-based descriptions. Moreover, the difference between the analogical models and the rule-based descriptions is that the former operate in a completely inductive manner. By this

we imply that the model does not rely on a priori knowledge of which features are important and which ones are not.

The predictive power of the memory-based learning models will be tested by comparing the outcomes of simulations to behavioral responses collected from native speakers. In particular, for each allomorph, we will be looking at the correlation coefficients between probabilities assigned by the model and the probabilities observed in behavior of native speakers (by dividing speakers preferring one allomorph with total number of speakers in a given sample). Because the simulations are based on the principles of memory-based learning, high correlation coefficients between these probabilities would suggest that these principles have a cognitive relevance.

Finally, the memory-based learning models will only use similarity between forms at the level of orthography/phonology.[3] Although a clear improvement in predictions is to be expected if additional similarities were included, we shall opt for simplicity, and examine the explanatory potential of a simple measure.

2. Method

2.1. Experimental data

The experimental data are taken from Jovanović (2008). In total, 42 adult participants, first year students of Psychology in Novi Sad, mainly females, with normal or corrected-to-normal vision participated in a computerized *wug* test. Jovanović used 125 pseudo-stems that followed Serbian ortho-phonotactic constraints. Each pseudo-stem was exactly five characters long, and had a fixed CVCVC structure. The final VC segment was controlled: all 25 Serbian consonants occurred five times as a final consonant, preceded once with each of the five vocals. For example, some of the pseudo-stems used in the experiment were: *bobaš, cogilj, gofić, nirib, salav* and so on. To implement the *wug* test, Jovanović downloaded 125 pictures from the *What is it?* website (http://puzzlephotos.blogspot.com). Each trial started with presentation of an unknown picture with its pseudo-word label in the nominative singular (for 2000 ms). Then, a grammatical Serbian sentence appeared with the critical pseudo-word in both

---

[3] Serbian has shallow orthography, and mapping from phonology to orthography is one to one. Hence, for the purpose of the present research, this difference can be disregarded.

instrumental singular allomorphs. One allomorph was positioned a row above and the other was positioned a row below blank space that was in line with the rest of the words forming a sentence (for example: *Motori se testiraju cogiljem/cogiljom* 'Engines are tested (by) cogiljem/cogiljom'. The participants' task was to choose one of the two forms by pressing a spatially corresponding button. There was no response time-out. It took approximately ten minutes for a participant to complete the task. Based on participants' choice, probability of each of the two allomorphic forms was estimated.

2.2. Simulation procedure

Implementation of the memory-based learning model started with the selection of an exemplar-storage that made up the "memory" of the model. For the present research, we used all 3481 masculine and neuter nouns from the *Frequency Dictionary of Contemporary Serbian* (Kostić 1999), which occurred in instrumental singular case. Neuter nouns were included because their instrumental singular can also attach both *-om* and *-em* suffix, depending on the final vowel (*-o* or *-e*). This inclusion gave additional noise in the exemplar-storage, thus making analogical learning more demanding.

In memory-based learning, the problem of predicting an allomorph is considered a simple classification problem: each pseudo-word needs to be classified as taking *-om* or *-em.* For this, the memory base was searched for the k nearest neighbors. For instance, in a model where the neighborhood size (k) equals 7, we would search the memory for the 7 stems that were most similar to the pseudo-word.[4] We could then look at how often the *-em* and *-om* suffixes occurred among these stems. The estimated probability of each suffix then was a simple ratio of the times it occured in the neighborhood to the total number of stems in that neighborhood. We tested models with different neighborhood sizes: we linearly increased k from 1 to 16, after which we used an exponential growth function of base 2 ($k = 32, 64, \ldots, 1024, 2048$), until finally k equalled the size of the lexicon (3481 items).

---

[4] In practice, the parameter k refers to nearest distances rather than nearest neighbors. When several exemplars occur at the same distance from the target, these exemplars are considered tied. In other words, a k-NN model looks for at least k exemplars. See Keuleers–Daelemans (2007) for a more detailed treatment of this issue.

In addition to the parameter k, memory-based learning models have two other crucial parameters: the **distance metric** used for computing the similarity between exemplars stored in memory and the pseudo-word to be classified, and the **decay function**, defining how a neighbor's weight in the classification decreases with distance from the target pseudo-word. We employed three well-known distance metrics: Jeffrey divergence, Levenshtein distance, and Hamming or Overlap distance. The Overlap metric is the coarsest: it simply counts the number of mismatching features. The Levenshtein distance is a generalized version of Overlap distance: it measures how many features must be inserted, deleted, or replaced to transform the stem into the pseudoword. Finally, Jeffrey divergence uses principles from information theory to give a weight to each feature, and operates as a weighted Overlap metric (for an in-depth presentation of these measures consult Rubner et al. 2000; Levenshtein 1966; Hamming 1950; for their application in linguistics see Daelemans–van den Bosch 2005). In addition to the distance metrics, we compared three decay functions: Zero Decay, where all neighbors have the same influence on classification, regardless of their distance to the pseudo-word; Inverse Distance Decay, where neighbors are weighted by the inverse of their distance; and Exponential Decay, where a neighbor's weight decreases exponentially with its distance. Since both the neighborhood size, the definition of similarity and its decay weighting affect the composition of the neighborhood, these parameters can interactively affect the outcome of a simulation.

3. Results

In the very first step of analysis we estimated the probability of producing *-om* and *-em* suffix for each noun based on the participants' responses in the *wug* test. These probabilities were then correlated with the outcomes of the memory-based learning simulations, where distance metric, decay weight and neighborhood size were systematically varied as factors. These results are presented in Figure 1.

As we can see from the plots, the similarity between human and computer results, expressed in terms of product-moment correlation coefficient, reached its maximum very rapidly. This means that, in most cases, a very small number of exemplars was sufficient for memory-based learning to make a correct analogy and to produce human-like output of suffix

allomorphy in Serbian instrumental singular pseudo-nouns. After including about ten nearest neighbors, nothing much could be gained, as the right-hand lines presenting exponential increase of neighbors show. Moreover, without decay weighting any further increase in number of neighbors was harmful for the similarity between human and computer-simulated behavior, while exponential and inverse decay weights just alleviated the cost of using large neighborhoods.
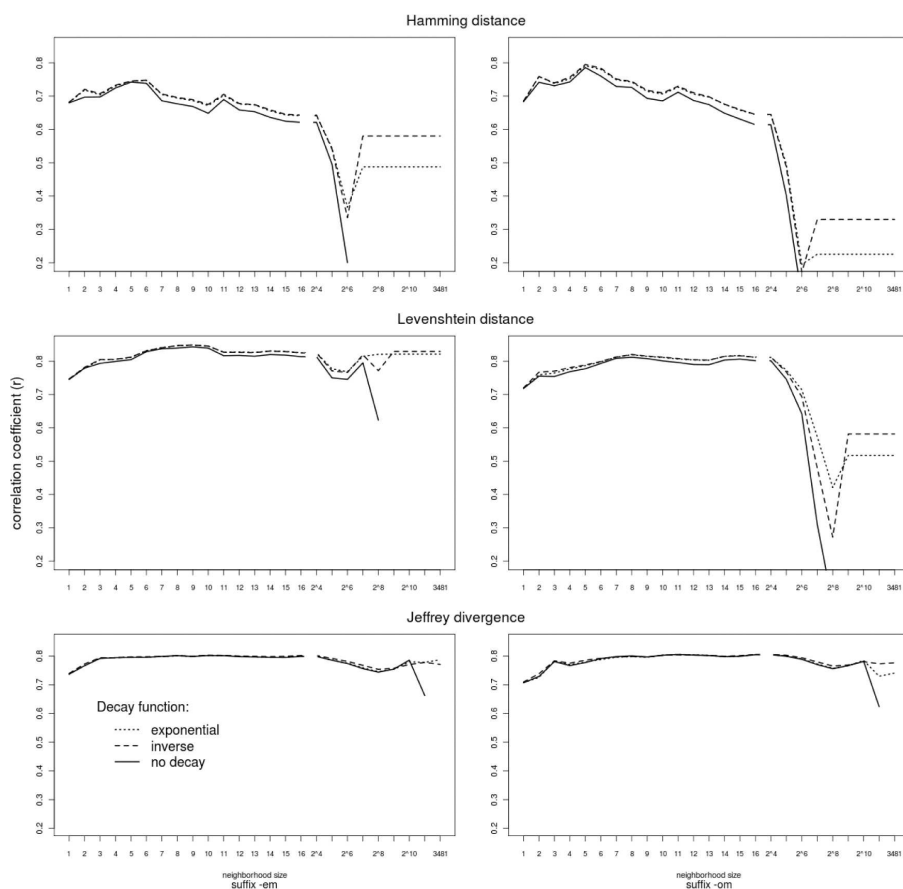

Fig. 1

Correlation coefficients between probabilities of producing *-om* and *-em* suffix allomorph, in behavioral experiment (*wug* test) and computer simulations. Line-breaks mark points where increase in neighborhood size changes from linear to exponential.

Row-wise comparisons of graphs by means of visual inspection already show that there were no substantial differences between the three distance metrics. However, in addition to visual inspection of graphs, we performed more detailed statistical comparisons of the three distance measures. Having two allomorphs crossed with three distance measures and three decay weights for each number of neighbors provided us with a total of 18 correlation coefficients per number of neighbors. In order to demonstrate that there were no significant differences between the 18 correlation coefficients within a given number of neighbors, we tested for the significance of the difference between the smallest and the largest correlation coefficient for each of the first sixteen neighborhood sizes, separately. If the difference between the smallest and the largest correlation coefficients was not significant, then we could deduce that none of the differences were. In other words, in this way we could demonstrate that all three distance measures using three different decay weighting methods performed equally well both for -*em* and for -*om* forms, for a given number of neighbors. The tests confirmed the null-hypothesis, thus proving that, in the range from one to sixteen nearest neighbors, with any of the three measures using any of the three decay weighting methods, we can achieve approximately the same success in simulating human production.

However, some variations were rather interesting and specific to a given measure. Using the simplest of the three measures, the Hamming distance (i.e., Overlap), gave somewhat lower correlations, but Jeffrey divergence, although the most sophisticated measure, did not perform better than Levenshtein distance. However, using Jeffrey divergence, the difference in similarity in producing each of the two allomorphic variants (ending with -*om* or with -*em*) was negligible. Levenshtein distance provided a better mach to the human responses for the -*em* allomorph, while the Hamming distance did exactly the opposite. Finally, larger neighborhoods were the least penalizing for Jeffrey divergence. It seems that this was the single point where some leverage from a more sophisticated measure was observed. This finding might be surprising, but can be simply explained by the fact that Jeffrey divergence is more fine-grained than the other measures. It expresses distances as real numbers. Therefore exemplars do not tie often at the same distance, while the Overlap and Levenshtein metric, which express distances in integer numbers, collapse many exemplars at the same distance. Jeffrey divergence reaches the same neighborhood size in absolute terms (the total number of exemplars) at a much later point than the other similarity metrics, thus having particular decay weighting as its intrinsic property.

In order to make our comparison of human and simulated behavior even more rigorous and conservative, we developed a specific statistical procedure which made use of logistic mixed-effect regression modeling (cf. Baayen et al. 2008; Jaeger 2008, etc.). Firstly, we ran our analysis for each distance metric and each decay weighting separately. Secondly, for a given distance metric (Jeffrey, Levenshtein, Hamming) with a given decay weighting (no decay, exponential decay, inverse decay), we iteratively applied linear-mixed modeling to test for the predictivity of a particular number of neighbors. We tested the range from one to sixteen neighbors. Each step in the procedure involved a comparison between two statistical models: a **more specific** model, which included probabilities from a simulation with k neighbors, with a **more general** model, which included those probabilities and the probabilities of the simulation with $k + 1$ neighbors residualized on the probabilities of the simulation with k neighbors. In other words, we included only **novel** variability in probabilities from the simulation with $k+1$ neighbors. The statistical models used binomially distributed participants' responses—selecting the *-om* or *-em* suffix allomorph, as a dependent variable, while items (pseudo-words) and participants were treated as random-effects. Two successive models were compared by applying likelihood-ratio tests, which produced Chi-squared values and corresponding *p*-values that are listed in Table 1.

The first line in the table can be read as follows: *Were the probabilities obtained from a memory-based learning model with a neighborhood size of one a significant predictor of the participants' choice of allomorph?* Given the very small *p*-values in the columns corresponding to each metric, the answer to the question ought to be that a memory-based learning model with a single nearest neighbor was a very significant predictor of the participants' choice of allomorph, regardless of the similarity metric and decay weighting used. The second line, and all subsequent lines, should be interpreted in the following way: *Did increasing the neighborhood size by one significantly improve the prediction of the memory-based learning model compared to the model with the previous neighborhood size?* The Hamming distance, which is the coarsest measure of all three, demonstrated somewhat shivering predictiveness; increasing the neighborhood size might increase or decrease model fit in the first few steps, with a sudden uprise for the model with a neighborhood size of eleven. The Levenshtein metric had a much more regular behavior. At each step from one to seven, the model significantly gained in fit. If we recall that the Lev-

Table 1

The likelihood-ratio test *p*-values for a series of successive mixed-effect models with k and $k + 1$ residual probabilities as covariate predictors of participants' binary response. Significant *p*-value ($p < 0.05$) reads as significant contribution of $k + 1$ residuals ("ND" = no decay, "ED" = exponential decay, "ID" = inverse decay)

	Hamming distance			Levenshtein distance			Jeffrey divergence		
k	ND	ED	ID	ND	ED	ID	ND	ED	ID
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.003	0.001	0.001	0.000	0.001	0.000	0.002	0.002	0.001
2	0.141	0.310	0.312	0.036	0.004	0.005	0.001	0.001	0.001
3	0.008	0.009	0.009	0.036	0.108	0.109	0.157	0.236	0.363
4	0.015	0.030	0.039	0.044	0.034	0.034	0.639	0.446	0.366
5	0.044	0.033	0.038	0.001	0.002	0.003	0.661	0.781	0.714
6	0.468	0.683	0.727	0.027	0.030	0.032	0.393	0.501	0.719
7	0.769	0.919	0.949	0.062	0.035	0.023	0.146	0.155	0.138
8	0.193	0.262	0.243	0.143	0.210	0.249	0.755	0.676	0.836
9	0.770	0.737	0.734	0.788	0.948	0.976	0.128	0.131	0.107
10	0.003	0.009	0.009	0.221	0.205	0.218	0.863	0.767	0.947
11	0.646	0.635	0.600	0.365	0.399	0.556	0.378	0.346	0.384
12	0.219	0.204	0.254	0.949	0.960	0.738	0.651	0.679	0.743
13	0.462	0.643	0.565	0.086	0.113	0.101	0.865	0.688	0.853
14	0.344	0.472	0.477	0.871	0.734	0.687	0.757	0.512	0.309
15	0.678	0.698	0.736	0.785	0.908	0.822	0.125	0.125	0.167

enshtein distance uses only the simple operations of insertion, deletion and substitution of feature values when expressing the distance between two exemplars, then it is striking that using this metric made for better predictions than using Jeffrey's divergence, a more complex measure with a higher information load. With Jeffrey divergence, increasing the neighborhood size improved the fit at each step, until a model with a neighborhood size of three was reached. However, Jeffrey's metric with three neighbors and without decay weighting gave a slightly worse fit than the model with a Levenshtein metric with seven neighbors and no decay weighting. This is indicated both by smaller beta estimate and *z*-statistic ($\beta = 3.267$, $z = 11.82$, $p < 0.001$ and $\beta = 4.041$, $z = 13.74$, $p < 0.001$), and measures of goodness-of-fit (greater AIC: 4612 vs. 4590; and smaller log-likelihood value: -2302 vs. -2291). All this leads to the conclusion

that Levenshtein distance is not just a simpler, but also a better predictor of human responses. It only needs a larger neighborhood for proper analogical inferences.

After closer examination of the candidate model (using the Levenshtein distance, no decay weighting and with seven nearest neighbors), another interesting pattern of results was revealed. Our critical predictor covariate—probabilities from simulation using Levenshtein distances with $k = 7$ nearest neighbors—entered into significant interaction with the random effect of participants. In other words, in addition to the by-participant adjustment for the intercept, we needed to let loose the slope of the predictor allowing it to vary across participants. Furthermore, results showed significant correlation between by-participant random intercepts and by-participant random slopes for simulated probabilities ($r = -.60$, after removing spurious residuals). Although this might seem discouraging for interpretation and understanding, it actually uncovered a detailed interplay between probabilities obtained from computer simulation and that from participants' responses. Firstly, the adjustments for intercept show that participants differ significantly in their "readiness" to produce an *-em*-ending variant. Secondly, although, as expected, we observed a significant positive correlation between simulated and human probabilities in producing the variant with the suffix *-em*, this correlation varied across participants: changes in probabilities for the *-em* suffix variant matched changes obtained in computer simulations more tightly for some participants than others. Finally, strong negative correlation between by-participant random intercepts and by-participant random slopes for simulated probabilities told us that the higher the base probability for producing an *-em* variant for a given participant, the less tightly she/he matched changes in computer simulated probabilities. However, one should keep in mind that the observed variation did not affect the overall predictability of simulation outcomes, but only revealed additional peculiarities in participants' behavior.

By-participant variations in the intercept and the slope for simulated probabilities are presented in Figure 2. As we can ascertain from the left panel, there is a balanced number of participants that produce *-em*, as well as those who produce *-om* with the higher odds. Similarly, adjustments for the slope, represented on the right panel, are scattered on the both sides as well. It turned out that computer simulated probabilities appear as a **prototypical** or average participant, while real participants vary around. And this was to be expected too: analogical learning had taken

place from the *Frequency Dictionary of Contemporary Serbian* (Kostić 1999), as exemplar-storage, where allomorphic variants for masculine and neuter instrumental singular were averaged across many native speakers. Thus, exemplars from a multitude became wide and middling, leading to the analogical inferences of a typical native speaker of the Serbian language.
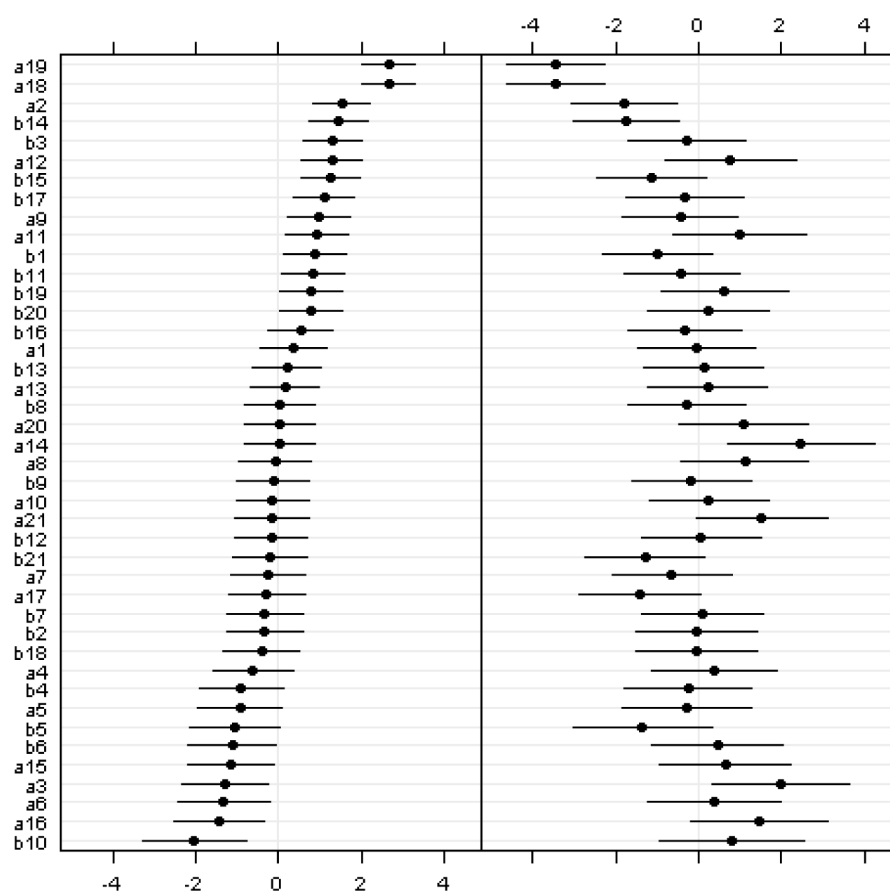

Fig. 2

Visualization of the by-participants adjustments for the intercept (left-hand panel) and slope (right-hand panel) of the probabilities obtained in a computer simulation using Levenshtein distance with $k = 7$ nearest neighbors. Both panels are centered to the grand intercept and grand slope. Thus, values greater than zero correspond to higher intercept value and steeper slope for a given participant, and vice versa.

4. Discussion

We aimed at demonstrating that the production of allomorphs of the instrumental singular of Serbian masculine nouns can be accounted for by memory-based learning. We started by implementing a memory-based learning model of the allomorphy in question and comparing probabilities assigned by the model with probabilities of each of the allomorphs being produced by native speakers. Our analyses showed that outcomes of the model closely resembled native speakers' behavior. We make no attempt in claiming that the model architecture mirrors the organization and processing performed by the cognitive system. However, we wish to state that, in principle, the patterns observed in the behavior of native speakers can be accounted for by a very simple learning principle. This would argue against application of rules in describing linguistic phenomena. Interestingly, predictions of the model were equally successful for *-em* and *-om* suffixes, allowing our model to argue against "default" accounts of language.

In the model we applied, the probability of a given allomorphic form was a result of the analogical inference based on a simple orthographic similarity between the stem and a small number of exemplars stored in memory. We compared predictions derived from three measures of orthographic similarity. Our results showed that the three measures were highly similar in predicting human responses. For each of the measures, strong resemblance to native speakers' behavior could be achieved even when analogical inference was based on only one exemplar (the one that is most similar to the stem in question). As expected, resemblance to native speakers' behavior increased with an increase in the number of exemplars evoked from the memory, but only up to a certain neighborhood size. Moreover, at a certain point, taking more exemplars into account degraded the performance of the model. The more exemplars are taken into account, the more a model resembles a frequency based approach: if all exemplars in the model's memory are taken into account, the output of the model is simply the the ratio of *-em* vs. *-om* suffix frequencies in the memory. This suggests that making decisions independent of similarity would be a bad strategy. Although the three similarity measures were equally successful as predictors, the speed of the observed degradation (produced by an increase in neighborhood size) differed. Additionally, introduction of exponential or inverse decay weighting appeared as beneficial in larger and large neighborhoods, diminishing overall degradation.

Detailed analyses that took into account both simplicity of similarity measures and the speed of degradation that followed increase in neighborhood size, demonstrated that the optimal solution uses a simple measure and a neighborhood of seven exemplars. From this conclusion follows the almost indecent question of whether this number could be "magical", not only for working memory load (Miller 1956, but for language processing, as well. Unfortunately, present findings from one language and concerning one specific phenomenon cannot provide the answer.

The observed results add to a growing body of research showing that memory-based learning models make relevant predictions about the cognitive processes involved in various linguistic phenomena, such as the formation of Dutch plurals (Keuleers et al. 2007; Keuleers–Daelemans 2007), Dutch diminutives (Daelemans et al. 1997), English past tense (Keuleers 2008), German plurals (Hahn–Nakisa 2000), Spanish gender assignment (Eddington 2002b) and so on. On a more general level, these findings fit well with the **gradient** view of various linguistic phenomena. This framework solicits for continual as opposed to discrete transitions between linguistic categories (cf. Albright–Hayes 2003; Hay–Baayen 2005; Baayen et al. 2006; Bybee 2007; Keuleers et al. 2007; Milin et al. 2009, etc.).

Allomorphy seems to be a prime example not only of gradience and continuity of language phenomena, but also of the analogical nature of morphological production. On the one hand, allomorphic variations appertain to the **degree** of one realization or another, not to crisp, clear-cut categories. On the other hand, in humans, analogical inferences can naturally produce form variation. Based on analogy, forms can be generated in fine-grained varieties, without the need for categories (cf. Albright–Hayes 2003).

References

Albright, Adam – Bruce Hayes 2003. Rules vs. analogy in English past tenses: A computational/experimental study. In: Cognition 90 : 119–61.

Allen, Mark – William Badecker 1999. Stem homograph inhibition and stem allomorphy: Representing and processing inflected forms in a multilevel lexical system. In: Journal of Memory and Language 41 : 105–23.

Baayen, R. Harald 2003. Probabilistic approaches to morphology. In: Rens Bod – Jennifer Hay – Stefanie Jannedy (eds): Probabilistic linguistics, 229–87. MIT Press, Cambridge MA.

Baayen, R. Harald – Doug J. Davidson – Douglas M. Bates 2008. Mixed-effects modeling with crossed random effects for subjects and items. In: Journal of Memory and Language 59 : 390–412.

Baayen, R. Harald – Laurie Beth Feldman – Robert Schreuder 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. In: Journal of Memory and Language 53 : 496–512.

Berko, Jean 1958. The child's learning of English morphology. In: Word 14 : 150–77.

Blevins, James P. 2004. Inflectional classes and economy. In: Lutz Gunkel – Gereon Müller – Gisela Zifonun (eds): Explorations in nominal inflection, 41–85. Mouton de Gruyter, Berlin & New York.

Bloomfield, Leonard 1933. Language. Holt, Rinehart and Winston, New York.

Boudelaa, Sami – William D. Marslen-Wilson 2001. Morphological units in the Arabic mental lexicon. In: Cognition 81 : 65–92.

Boudelaa, Sami – William D. Marslen-Wilson 2004. Allomorphic variation in Arabic: Implications for lexical processing and representation. In: Brain and Language 90 : 106–16.

Bybee, Joan L. 2007. Frequency of use and the organization of language. Oxford University Press, Oxford.

Daelemans, Walter – Peter Berck – Steven Gillis 1997. Data mining as a method for linguistic analysis: Dutch diminutives. In: Folia Linguistica (Acta Societatis Linguistica Eurpaeae) 31 : 57–75.

Daelemans, Walter – Antal van den Bosch 2005. Memory-based language processing. Cambridge University Press, Cambridge.

Eddington, David 2002a. Dissociation in Italian conjugations: A single-route account. In: Brain and Language 81 : 291–302.

Eddington, David 2002b. Spanish gender assignment in an analogical framework. In: Journal of Quantitative Linguistics 9 : 49–75.

Estes, William K. 1994. Classification and cognition (Oxford Psychology 22). Oxford University Press, New York.

Hahn, Ulrike – Ramin Charles Nakisa 2000. German inflection: Single route or dual route? In: Cognitive Psychology 41 : 313–60.

Hamming, Richard W. 1950. Error detecting and error correcting codes. In: Bell System Technical Journal 29 : 147–60.

Harris, Zellig S. 1951. Methods in structural linguistics. University of Chicago Press, Chicago.

Harris, Zellig S. 1957. Co-occurrence and transformation in linguistic structure. In: Language 33 : 283–340.

Hay, Jennifer – R. Harald Baayen 2005. Shifting paradigms: Gradient structure in morphology. In: Trends in Cognitive Sciences 9 : 342–8.

Hayes, Bruce – Zsuzsa Cziráky Londe 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. In: Phonology 23 : 59–104.

Ivić, Pavle 1990. O jeziku nekadašnjem i sadašnjem [On past and contemporary language]. Bigz-Jednistvo, Belgrade.

Jaeger, Florian 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. In: Journal of Memory and Language 59 : 434–46.

Järvikivi, Juhani – Raymond Bertram – Jussi Niemi 2006. Affixal salience and the processing of derivational morphology: The role of suffix allomorphy. In: Language and Cognitive Processes 21 : 394–431.

Järvikivi, Juhani – Jussi Niemi 2002a. Allomorphs as paradigm indices: On-line experiments with Finnish free and bound stems. In: SKY Journal of Linguistics 15 : 119–43.

Järvikivi, Juhani – Jussi Niemi 2002b. Form-based representation in the mental lexicon: Priming (with) bound stem allomorphs in Finnish. In: Brain and Language 81 : 412–23.

Jovanović, Tamara 2008. Ispitivanje prirode alomorfije: grafo-fonoloski korelati alomorfije u srpskom jeziku i efekat frekvence sufiksa [Examining allomorphy: Graphophonological correlates of allomorphy in Serbian and suffix frequency effect]. MA thesis, University of Novi Sad.

Jovanović, Tamara – Dušica Filipović Đurđević – Petar Milin 2008. Kognitivna obrada alomorfije u srpskom jeziku [The cognitive processing of the allomorphy in Serbian]. In: Psihologija 41 : 87–101.

Kertész, Zsuzsa 2003. Vowel harmony and the stratified lexicon of Hungarian. In: The Odd Yearbook (ELTE SEAS Undergraduate Papers in Linguistics) 7 : 62–77.

Keuleers, Emmanuel 2008. Memory-based learning of inflectional morphology. Doctoral dissertation, University of Antwerp.

Keuleers, Emmanuel – Walter Daelemans 2007. Memory-based learning models of inflectional morphology: A methodological case study. In: Lingue e Linguaggio 6 : 151–74.

Keuleers, Emmanuel – Dominiek Sandra – Walter Daelemans – Steven Gillis – Gert Durieux – Evelyn Martens 2007. Dutch plural inflection: The exception that proves the analogy. In: Cognitive Psychology 54 : 283–318.

Kostić, Đorđe 1999. Frekvencijski rečnik savremenog srpskog jezika [Frequency dictionary of contemporary Serbian]. Institute for Experimental Phonetics and Speech Pathology & Laboratory of Experimental Psychology, University of Belgrade, Belgrade. (www.serbiancorpus.edu.rs)

Krott, Andrea – Robert Schreuder – R. Harald Baayen – Wolfgang U. Dressler 2007. Analogical effects on linking elements in German compounds. In: Language and Cognitive Processes 22 : 25–57.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. In: Cybernetics and Control Theory 10 : 707–10.

Lieber, Rochelle 1982. Allomorphy. In: Linguistic Analysis 10 : 27–52.

Lyons, John 1968. Introduction to theoretical linguistics. Cambridge University Press, Cambridge.

Marr, David 1982. Vision: A computational investigation into the human representation and processing of visual information. W. H. Freeman, San Francisco.

Milin, Petar – Dušica Filipović Đurđević – Fermin Moscoso del Prado Martín 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. In: Journal of Memory and Language 50 : 50–64.

Miller, George A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. In: Psychological Review 63 : 81–97.

Mirković, Jelena – Mark S. Seidenberg – Marc F. Joanisse 2011. Probabilistic nature of inflectional structure: Insights from a highly inflected language. In: Cognitive Science 35. In press.

Norris, Dennis 2005. How do computational models help us build better theories? In: Anne Cutler (ed.): Twenty-first century psycholinguistics: Four cornerstones, 331–46. Lawrence Erlbaum, Mahwah NJ.

Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. In: Journal of Experimental Psychology 15 : 39–57.

Pléh, Csaba 1989. The development of sentence interpretation in Hungarian. In: Elizabeth Bates – Brian MacWhinney (eds): The crosslinguistic study of sentence processing, 158–84. Cambridge University Press, Cambridge.

Pléh, Csaba – Ágnes Lukács – Mihály Racsmány 2002. Morphological patterns in Hungarian children with Williams syndrome and the rule debates. In: Brain and Language 86 : 377–83.

Ratcliffe, Robert R. 1998. The "broken" plural problem in Arabic and comparative Semitic: Allomorphy and analogy in non-concatenative morphology. John Benjamins, Amsterdam & Philadelphia.

Rubner, Yossi – Carlo Tomasi – Leonidas J. Guibas 2000. The Earth Mover's distance as a metric for image retrieval. In: International Journal of Computer Vision 40 : 99–121.

Saussure, Ferdinand de [1916] 1966. Course in general linguistics. McGraw-Hill, New York.

Schreuder, Robert – R. Harald Baayen 1995. Modeling morphological processing. In: Laurie Beth Feldman (ed.): Morphological aspects of language processing, 131–54. Lawrence Erlbaum, Hillsdale, New Jersey.

Skousen, Royal 2002. An overview of analogical modeling. In: Royal Skousen – Daryle Lonsdale – Dilworth B. Parkinson (eds): Analogical modeling: An exemplar-based approach to language, 11–26. John Benjamins, Amsterdam & Philadelphia.

Smith, Edward E. – Douglas L. Medin 1981. Categories and concepts. Harvard University Press, Cambridge MA.

Spencer, Andrew 2001. Morphology. In: Mark Aronoff – Janie Rees-Miller (eds): The handbook of linguistics, 213–38. Blackwell, Oxford & Malden MA.

Zec, Draga 2006. Phonology within morphology in South Slavic: The case of OV augmentation. University of Nova Gorica, Nova Gorica.