PSIHOLOGIJA, 2013, Vol. 46 (4), 397–420 UDC 159.9.07:81'373 ; 811.511.141'367.622

DOI: 10.2298/PSI1304397P

Entropy measures and predictive recognition as mirrored in gating and lexical decision over multimorphemic Hungarian noun forms[1]

Csaba Pléh[1], Kornél Németh[2], Dániel Varga[3], Judit Fazekas[2], and Klára Várhelyi[2]

[1] *Center for Cognition and Communication and Dept of Psychology, Eszterházy College, Eger, Hungary and Collegium de Lyon, ENS Lyon, France*
[2] *Budapest U of Technology and Economics, Department of Cognitive Science, Hungary*
[3] *Budapest U of Technology and Economics, Media Research Center at the Department of Sociology and Communications (MOKK), Hungary*

Our paper is an attempt to indicate the relevance of information theoretical accounts to understand word recognition and morphological processing in Hungarian, along with other studies using more traditional predictors like linear position and morphological composition. The first two experiments were gating studies. The effect of the decision points was only evident in frequent words. The correct recognition means for the recognition points differ from the means for one-before-recognition points, indicating that the recognition point follows a sudden drop of the entropy value. This shows how entropy measures can be used to predict word recognition in actual language performance. The next two experiments examined the word reconstruction effect. A clear bathtub effect (Aitchison, 1987) was obtained: reconstruction was highest in the cases where both the beginning and the end were correct. The last, lexical decision based study used four basic morphological types of markers (plural, second and first possessive) and three types of case *(-nak,-ban-,-ra* ‘DAT, INSIDE, ONTO). The main effect of the frequency and the error type was significant. Frequent words were judged faster but less accurately, suggesting a trade-off. The later the mistake is, the faster and easier its rejection was.

Keywords*: morphological processing, entropy, Hungarian, multimorphemic nouns, lexical decision, gating*

Corresponding author: pleh.csaba@ektf.hu

[1] The final version of the paper was much improved thanks to the suggestions of Petar Milin and two anonymous reviewers, and discussion with Jim Blevins. Thanks all of them. In preparing the final version of the paper Csaba Pléh was enjoying a scholarship at the Collegium de Lyon, ENS Lyon and a support within the framework of the TÁMOP-4.2.2.C-11/1/KONV-2012-0008 (Social Renewal Operative Program) project titled *The application of ICT in learning and knowledge acquisition: Research and Training Program Development in Human Performance Technology.* Said project was implemented by the support of the European Union and the co-financing of the European Social Fund.

THE PAST AND PRESENT OF INFORMATION THEORETICAL APPROACHES TO LANGUAGE

Information theory had a strong appeal for the nascent psycholinguistics in the early 1950s. George Miller (1951) in his textbook and in some empirical studies took up the approach represented by the Shannon-Weaver (1948, 1949) model of communication. In particular, he was trying to show that information value can be used as a possible predictor of word recognition difficulties (in fact, these efforts came up with negative results: our channels are limited not by information value, but rather by the number of chunks we are able to handle, as he himself (Miller, 1956) emphasized later. In another attempt he was trying to show that statistical approximations to English have a predictive value for sentence recognition. This was shown by using in fact rough subjective estimates of conditional probability. In the preparation of their materials subjects were given the first n words (e.g. 2 words) of a beginning sentence and were asked to put them in a sentence. In the next step with a different group of subjects the n + 1 words (the first 3 in the example) were provided as triggers for continuation. The strings prepared this way were used as stimulus materials (Miller & Selfridge, 1950). This statistical model of language, however, was very soon criticized on a theoretical basis by Chomsky (1957) who showed that information theoretical models are unable to deal with the inherent creativity of language. Soon, Miller and Chomsky (1963, Chomsky & Miller, 1963) also started to claim that information theory based approaches are not only unable to deal with grammatical competence but they are also unable to account for language performance. Models of language use have to entail a grammar as well (in particular, of course, at the given time, a transformational generative grammar). Since then, one of the crucial issues for a large portion of modern psycholinguistics has been the experimental study of how the rules of this grammar are implemented in models of language use.

As for our specific issue of morphological processing, László Antal (1962, 1964) in the 1960s made an interesting proposal. He started from the problems of segmentation in agglutinative languages and proposed that over words, the usual tendency is decreasing entropy. By this he basically meant the number of possible continuations at any given point that correspond to lexical density and grammatical structure. To take an English example, the string *boo* can continue as *boot, book, boor, boom* etc., having high uncertainty, while *prog* can only continue as *progr*, having no uncertainty at that point. Thus since Antal was lacking a frequency dictionary, he was using the entropy notion developed by Shannon-Weaver (1948) for equal probability outcomes where entropy is a function of the number of possible outcomes. Morphological boundaries break this monotonous decrease, and intuitively correspond to increasing entropy values.

If you take a word form like (1), the entropy value gradually decreases over the stem, and when the entropy value suddenly increases again you have a morpheme boundary. As the example below shows, there is indeed a gradual decrease of entropy over the word, now computed with not merely the options, but their frequencies also taken into account. Morpheme boundaries, however, correspond to plateaus rather than to increases.

(1) *igaz-ság-os-ak-at* ‘true-th-full-Plur-Accus’ “truesfulls”

Figure 1 shows the entropy values over the graphemes of this morphologically complex word.

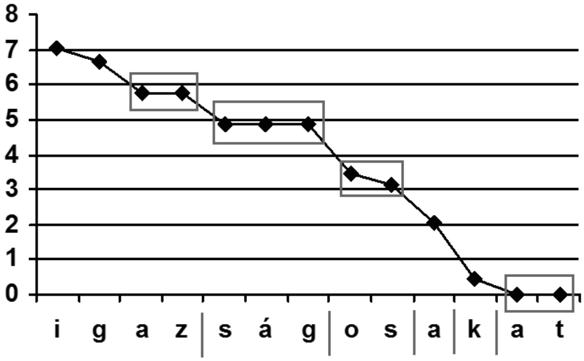

Figure 1. Entropy value changes over a multiple suffixed Hungarian noun. Entropy data based on the MOKK corpus (see below).

Entropy is measured here with taking frequency into account. The entropy of the observed corpus W conditioned on the prefix *x* is $H(W|x) = \sum_{w \in W} p(w|x) \log_2 p(w|x)$ $H(W|x) = \sum_{w \in W} p(w|x) \log_2 p(w|x)$, where $p(w|x)p(w|x$ is the probability of observing the word *w* in the corpus.

At the end of twentieth century the seemingly outdated information theoretical approaches became suddenly relevant and central again. One may suggest several reasons for this revival. One crucial factor, certainly, is the ease of computation, combined with the creation of large large lexical databases, often based on web sources (Baayen, 2005), and the remarkable development of statistical models of language processing. (For an overview with an eye on the psychology of language processing see Jurafsky & Martin, 2009). Information theory-based and Markov chain models were easy to criticize on a theoretical basis half a century ago, when neither the defenders nor the critics were able to execute actual computations over large corpora. Machine produced and machine analyzed texts – be them written or oral – provide an easy target to compute entropy values, transitional probabilities and the like. In this way, it became easy to compare the predictive power of grammar based and statistical predictions for language performance. Not surprisingly both in child language research (Saffran, Aslin, & Newport, 1996), and in language processing (Kostić, 1995, Moscoso del Prado Martín, Kostić, & Baayen, 2004) information theory related concepts reappear. There are even comparative entropy studies of printed text in different languages. Borgwaldt, Hellwig, and De Groot (2005) showed that the word initial letter to sound correspondences i.e., pronunciation entropy are different in different languages, with Hungarian being the most transparent one, using an orthography that is traditionally treated as having a high letter-to-sound correspondence. This corresponds to a letter-to-sound entropy value of 0.15 in Hungarian, while the English entropy is 0.65 and the French is 0.58.

Our paper is an attempt to indicate the relevance of these notions for *word recognition and morphological processing in Hungarian.* Hungarian is an interesting example to study the relevance of information theoretical notions over word processing due to its extremely rich morphological paradigms. As Kornai (1992, 1994) pointed out, even if we take relatively frequent words such as *asztal* 'table' or *szék* 'chair' out of the several hundred possible word forms starting with these stems only a couple of dozens show up in the frequency dictionary (37 and 23 in the given cases). Thus to come through even once with all the forms of even a frequent word a much bigger corpus is needed than what we would experience in our entire life span. Kornai uses this example to argue for morphemic or analytic representation in a reasoning arguing for rules. We would like to show that in actual language performance, that is also an argument for the use of information values for segmentation purposes.

Results of 5 experiments shall be presented that were all done over nouns to reveal how the informational structure of the nouns shows up in word recognition and morphemic segmentation.

STRUCTURE OF HUNGARIAN NOMINAL FORMS: A FEW INTRODUCTORY REMARKS

Hungarian nouns can take many suffixed forms. Derivational suffixes, an inflectional marker of possession, plural suffixes, and case markers plus plural markers combined with the root stem in a rather systematic way illustrated in (2), (3), and (4). The overall structure is: STEM – DERIVATION – POSSESSION – POSSPLU-POSSPERS – POSSESSED – PLURAL – CASE. (See for the details in Papp, 1982; Kornai, 1994; Kiefer, 2000)

There are many interesting subregularities that we shall not go into. The basic morphotactic principle is that case marking is always the last one, the morpheme sequence goes from the most lexical to the most syntactic, with some obligatory constraints. The examples below show in an intuitive way how this is related to the sequential unfolding of the word forms of nouns.

(2) Inflectional suffixes

ház *house*
ház-at *house– Acc*
ház-ak *house-Plur*
ház-am *house-PossMe*
ház-ak-at *house-Plur-Acc*
ház-a-m-at *house-Poss–1stSing-Acc*

(3) *Derivational suffixes*

ház *house*
ház-as '*housey' i.e. married*
ház-as-ság '*house-y-ness' i.e. marriage*

(4) Both derivation and inflection

ház *house*
ház-as-ság *marriage*
ház-as-ság-a-i-m-ban *marriage-Poss-Plur–1stSing-IN 'in my marriages'*

There are several psycholinguistic issues related to any morphological system and in particular to Hungarian, which in this regard is a rich system. By richness we mean that each noun stem has several hundred possible word forms, even not counting derivational forms.

1. *Segmentation:* Allows for lexical access. Segmentation is also relevant for the second task, for the decisions about which units to integrate. Agglutinative languages may support these efforts by their structural features. In Hungarian the fixed first syllable stress and the word final position of case markers gives a cue concerning access of the given stem and a 'backward sign' for compilation. At the same time, whatever precedes the stress counts as one word (Kornai, 1992, 1994).
2. *Lexical access:* What is the left-to-right procedure to access stems? Are there any linear asymmetries related to this? What parts of the word token form are most transparent during processing?
3. *Formal combinatorics and decomposition:* Do we always analyze complex word forms into their morphemes or is holistic access also possible? The hybrid models proposed in the literature and interpreted for Hungarian are assuming a standard rule-based solution with holistic access related to derivation-inflection differences, to relative frequency, and they are as well suggesting that derived words would be processed as holistic units (for a review see Clark, 1991) especially in the case of non-transparent derivations (Marslen-Wilson et al., 1994), while inflection as a rule would be treated in an analytic manner. There are alternative views that start from considering individual schemata as starting points, and treat rule based processing as more secondary (see for a review Bybee, 2007, and for an implemented version of memory based learning Daelemans and van den Bosch, 2005).
4. *Semantic integration:* Along with integration of forms one also needs to integrate the meaning of multimorphemic words.
5. *Stem allomorphy:* Several Hungarian noun paradigms are characterized by multiple allomorphs. Are allomorphs always mapped onto the same citation form during access or are there different access files depending on frequency and phonetic motivation?

The studies presented here are not directly addressing these issues separately. We merely mention them to keep in mind that the results of the individual experiments using different methodologies on the final run map onto these basic issues of lexical access and morphological processing. In the present outline the studies are not directly testing individual hypotheses. Rather, they are presented as pointing towards possible processing procedures.

Gating studies that are relevant to lexical access were performed to look for the value of information values in processing word stems. For the analysis of morphological decomposition, scrambling studies and lexical decision studies were performed.

EXPERIMENT 1

A WEB-based gating and lexical access study in Hungarian

Gating is a traditional method to look for the temporal structure of word recognition. Since it was introduced by Grosjean (1980) a generation ago, it was used to show effects of frequency, word length, stress pattern, competing words (*try, cry, shy*), lexical uniqueness, morphological structure and sentential contexts (for a review see Grosjean, 1996).

Two studies of gating were performed with carefully selected rare and frequent Hungarian disyllabic nouns, in the first study ever of gating in Hungarian. The present study was administered over the Internet.

Method

Materials. Sixty words were used of the type *vizsga* ‘exam’, *japán* ’japanese’ *böllér* ’butcher’ and the like. All items were disyllabic nouns. Thirty of them were frequent, 30 rare, and within each group 15 were used with an early uniqueness point like *japán*, and 15 with a late uniqueness point like *cinke*. Selection was based on a manual look up of the MOKK (2006) corpus. The actual words are listed in the Appendix together with their frequency parameters and the entropy measures.

Subjects. A group of 51 healthy students with ages ranging from 18 to 25 years (31 female (20.41 year; *SD*=0.98; 20 male (21.11 year; *SD*=1.4)) of the BME (*Budapest University of Technology and Economics*, Budapest, Hungary) participated in this experiment as volunteers. Every subject had normal hearing. They were recruited through campus based email networks, gave informed consent before the actual experiment and they were tested individually at their own computer site, and received partial course credit for participating. None of the subjects had any prior experience with the experimental task.

Procedure. In this experiment a within subject design was used. Every participant heard all of the word fragments with consecutively longer gates (audio segments of 90, 120, 210, 300, 390 ms) but the order of the words was randomized. After every sound segment participants had to find out what the word was and then type the total word on the computer keyboard. After typing, at every guessing they needed to assign the certainty of the answer (1 – absolutely unsure, 2 – rather unsure, 3 – rather sure, 4 – absolutely sure). If the answer was correct – irrespective of the gate time and the confidence judgment – the program presented the first fragment of the next word.

Statistical analysis. The accuracy and the decision certainty were submitted to a three-way repeated-measures ANOVA with Frequency (2; rare, frequent), Uniqueness (2; early, late) and Gate length (5; 90 ms, 120ms, 210ms, 300ms, 390ms) as within-subject factors. All analyses involved Greenhouse-Geisser adjusted degrees of freedom for correction for non-sphericity. Post-hoc statistics were performed with Bonferroni tests.

Results

The basic results of the impact of frequency and uniqueness point are shown in Figure 2. Notice that in this off-line arrangement correct recognition is rather low across the board.

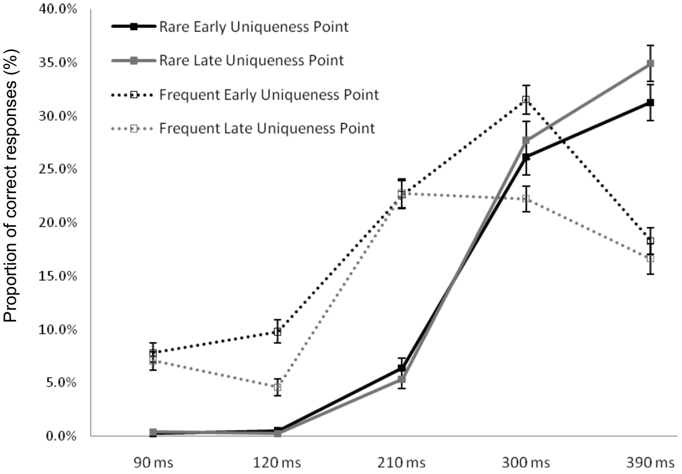

Figure 2. Proportion of correct responses by different gates.
The error bars show the standard error of mean.

Accuracy for frequent words was much higher when compared to rare words, $F(1,50)=84.917$, $p<0.00001$, $\eta^2=0.6294$. The main effect of Uniqueness shows that words with early uniqueness point were recognized with higher accuracy, $F(1,50)=35.708$, $p<0.00001$, $\eta^2=0.4166$. With the increase in presentation time (Gate length), the proportion of the correct responses increases as well, however, only between the 2nd, 3th and 4th gates (post-hoc $p<0.001$), $F(2.6169,130.8469)=280.066$, $p<0.00001$, $\eta^2=0.8485$. While the accuracy for words with early and late uniqueness point was identical (post-hoc $p=0.056$), the Frequency × Uniqueness interaction showed much higher accuracy for frequent words with early uniqueness point, $F(1,50)=97.264$, $p<0.00001$, $\eta^2=0.6605$). The Frequency × Gate length interaction has shown that the accuracy was higher for frequent words at 90ms, 120ms and 200ms, but lower at 390ms, $F(2.6929,134.6483)=77.488$, $p<0.00001$, $\eta^2=0.6078$.

The accuracy was higher for words with early uniqueness point, but reached significance (post-hoc $p=0.03169$) at 300ms only, $F(2.4027, 120.1352)=3.082$, $p<=0.017186$, $\eta^2=0.058$.

The Frequency × Uniqueness × Gate length interaction has shown that the accuracy, after reached the maximum at the 3nd gate, decreased more for words with late uniqueness point, $F(2.3283,116.4176)=3.826$, $p=0.005093$, $\eta^2=0.071$.

In the decision certainty, frequency also had a significant main effect, $F(1,51)=95.621$, $p<0.00001$, $\eta^2=0.6521$. The mean certainty value was 2.22 for frequent, and 1.99 for infrequent items. The certainty increases linearly as a function of the gate durations, $F(1.4762,75.2863)=266.69$, $p<0.00001$, $\eta^2=0.8394$. Also, the decision certainty was higher for words with late uniqueness point, $F(1,50)=31.448$, $p<0.00001$, $\eta^2=0.3814$.

As shown by the Uniqueness × Gate length interaction, the decision certainty differed in different gates, $F(2.893,147.5431)=31.2839$, $p<0.00001$, $\eta^2=0.3802$. It was higher for words with late uniqueness point, at 300ms and 390ms. The Frequency × Gate length significant interaction revealed that the decision certainty after 2nd gate increased more for frequent words (post-hoc $p<0.001$) than for rare ones $F(2.193,111.8793)=33.241$, $p<0.00001$, $\eta^2=0.3946$. The significant Frequency × Uniqueness interaction has shown that the decision certainty for frequent words is higher when the uniqueness point is late $F(1,50)=5.8563$, $p=0.0191$, $\eta^2=0.103$. Finally, the significant Frequency × Uniqueness × Gate length interaction revealed that the certainty of decision for rare words depends much less on the uniqueness point (a significant difference was found only at 390ms) and that it is much higher for words with late uniqueness point and for shorter gate durations, $F(3.11,158.6402)=8.84$, $p<0.00001$, $\eta^2=0.1477$.

EXPERIMENT 2

Gating and grammatical constraints

In the second gating experiment, laboratory based individual measurements were used.

Method

Materials and design. The entire design was very similar to the Experiment I with only a few differences. In Experiment II a mixed design was applied, with one between subjects factor. The between subjects factor was grammatical priming instructions. One group of participants was given the following information: “You will hear only two syllabic nouns without suffixes”, another group of the participants was not given this information. The grammatical constraint was introduced to look for possible top-down effect on recognition. Every participant heard all the word fragments shown in the Appendix with consecutively longer gates. In Experiment II, however, rather than using a stepwise increase over the same sequence, the sequence of words was randomized. Each subject heard all five length versions of each stimulus, but in a cross-stem randomized order.

Procedure. After every sound segment participants had to find out what the word was, and say out loudly the total word into a microphone that measured the reaction time of the answer. And thereafter at every guessing they needed to assign the certainty of the answer (1 – absolutely unsure, 2 – rather unsure, 3 – rather sure, 4 – absolutely sure). If the answer was correct – irrespective of the gate time and the confidence judgment-, the program presented the next item.

Subjects. Fourteen adults with ages ranging from 18 to 60 years (7 female (28.0 years, $SD=13.14$; 7 male (28.16 years, $SD=14.05$) participated in this experiment. Every subject had normal hearing. They were tested individually and they all gave informed consent before the experiment. None of the subjects had any prior experience with the experimental task.

Statistical analysis. The accuracy and the decision certainty were submitted separately to a three-way repeated-measures ANOVA with Frequency (2; rare, frequent), Uniqueness (2; early, late) and Gate length (5; 90ms, 120ms, 210ms, 300ms, 390ms) as within-subject factors. The grammatical constraint was used as between-subject factor. All analyses involved Greenhouse-Geisser adjusted degrees of freedom for correction for non-sphericity.

Results and discussion

In this study, as Figure 3 shows, in terms of accuracy there was an effect of Frequency, Uniqueness and of Gate length. Constraints resulted in 43 % correct recognition while without a constraint recognitions was less efficient (38 %), $F(1,12)=2.699$, $p=0.126$. The longer the gate, the higher was accuracy, $F(2.2823,27.388)=604.6026$, $p<0.000001$, $\eta^2=0.9805$. In addition, frequent words were recognized in 48 %, while rare words in 32 %, and this difference was also significant, $F(1,12)=141.2105$, $p<0.00001$, $\eta^2=0.9217$. Similarly, difference between early recognition point (42.47 %) versus late recognition point (37.96 %) was significant too, $F(1,12)=12.9465$, $p<0.001$, $\eta^2=0.6122$). The Frequency × Gate length interaction revealed that there is approximately 10–15% increase in performance for frequent words for successive gate durations, $F(4,48)=13.442$, $p<0.00001$, $\eta^2=0.5283$.

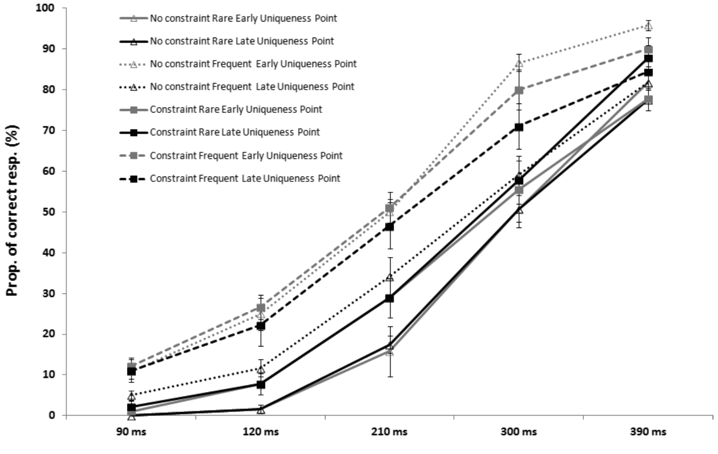

Figure 3. Effects of constraints, frequency, and uniqueness points on gating recognition. The error bars show the standard error of mean

The effect of the decision points was only evident in frequent words as Figure 4 shows. Frequent words with an early uniqueness point had a higher recognition rate. In the analysis of variance this is shown by a significant *Uniqueness × Frequency* interaction ($F(1,12)=11.645$, $p=0.0051$, $\eta^2=0.4925$).

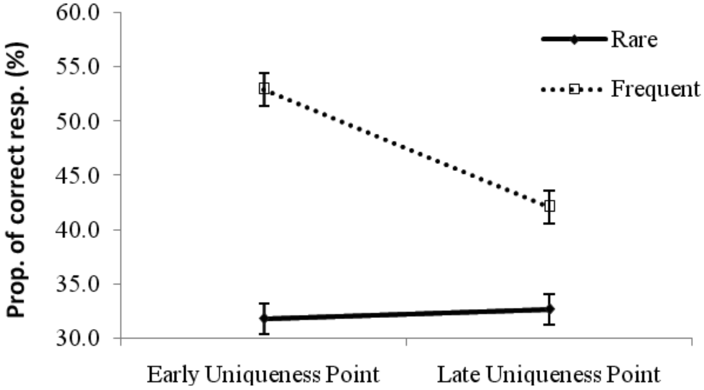

Figure 4. The effect of decisions points is only present in frequent words. The error bars show the standard error of mean.

The grammatical constraints showed a significant interaction with the decision points. Grammatical-phonotactic constraints only had an effect on word with a late decision point (constraint: 42 % vs. no constraint: 33 % of correct recognition), $F(1,12)=10.767$, $p=0.0065$, $\eta^2=0.4729$.

The decision certainty was higher for frequent words, $F(1,12)=8.2836$, $p=0.0138$, $\eta^2=0.4084$. As the length of the gate increased the decision certainty became higher and higher, $F(1.6412,19.694)=116.065$, $p<0.00001$, $\eta^2=0.9063$. The Frequency × Gate length interaction revealed that the self rated confidence of the responses (i.e., decision certainty) became higher for frequent words when compared to rare ones from the 3rd gate, $F(4,48)=4.5587$, $p=0.00335$, $\eta^2=0.2753$. The significant Uniqueness × Gate length interaction showed that the confidence was higher for words with late uniqueness point at longer gates, $F(4,48)=3.3959$, $p=0.01591$, $\eta^2=0.2206$. As the Uniqueness × Gate length × Constraint interaction shows, the decision certainty was higher for words with late uniqueness point when there was no grammatical constraint, $F(4,48)=4.3961$, $p=0.00415$, $\eta^2=0.2681$.

Reanalysis of the gating result using entropy measures

With the entropy measurements we wanted to give a more robust generalization of the intuitive notion of uniqueness point. In a way, an exploratory study was run along the lines initiated by Moscoso del Prado Martín, Kostić, and Baayen (2004) and by Wurm, Ernestus, Shreuder, and Baayen (2006) who had shown that using an information gain predictor which is combining the entropy of the stem with the conditional entropy of the bound morphemes, better prediction can be obtained for word recognition data then merely using frequencies as predictors.

To obtain corpus based interpretations of uniqueness points we worked with the MOKK corpus and frequency dictionary of Hungarian *(*http://mokk.bme.hu/

resources/webcorpus/index_html*)*. For a description of the corpus see Kornai et al. (2006). We used the stratum of the corpus that has a 96% correspondence to regular Hungarian orthography. That is roughly equivalent to everyday written Hungarian. This means 589 million word tokens and 7.2 million word types. For the gating experiments, we worked with the non-inflected nouns of the corpus. The corpus is part of speech tagged, and is passed through an automatized morphological disambiguator.

Each of the audio segments the subjects had to work with was manually transcribed. These transcripts are called ‘prefixes’ in the following. Thus prefixes are not grammatical prefixes, but the strings preceding a given point.

As an example here are the transcribed prefixes for the word

ablak

90ms: *a,* 120ms: *a,* 210ms: *ab,* 300ms: *abla,* 390ms: *abla*

Information metrics used

Several metrics were defined over prefixes. They all measure how a word beginning constrains the set of possible continuations.

Beginning types occurrences log: This metric is simply the number of word forms thus tokens in the corpus starting with the given beginning. This is much skewed variable, so its base 2 logarithm is used as a normalized variable.

Beginning frequency log: This is the number of TOKENS in our corpus starting with the given beginning. One can consider this the weighted version of the former, weighted by word frequencies. Here again, the logarithm is used.

Entropy: Entropy of the corpus, conditioned on the given beginning. Informally, this means the expected number of questions we need in a “20 Questions” game, if we must guess a randomly chosen word (by the frequency distribution in the corpus) with the given beginning. More formally, the entropy of the observed corpus W conditioned on the beginning *x* is

$$H(W|x) = \sum_{w \in W} p(w|x) \log_2 p(w|x) H(W|x) = \sum_{w \in W} p(w|x) \log_2 p(w|x),$$

where $p(w|x)p(w|x)$ is the probability of observing the word *w* in the corpus, conditioned on the fact that *w* starts with the beginning *x*. Typically, although not necessarily, the entropy monotonically decreases when calculated for longer and longer beginnings of a given word.

Entropy change: The decrease in entropy when compared to the previous gate. This is defined as $H(W|x_0) - H(W|x_{-1})H(W|x_0) - H(W|x_{-1})$ where $x_0 x_0$ is the beginning at the current gate, and $x_{-1}x_{-1}$ is the beginning at the previous gate. The value is not defined for the first gate. Also, the entropy change is measured over a fix 90 ms time interval, except for the first two gates.

Entropy values of the word items

As an item based post hoc test of the relevance of frequency and uniqueness, a two way analysis of variance was performed over the entropy values for the fourth letter position shown in the Appendix, with frequency and uniqueness point as the two factors. Both factors were significant. For frequency $F(1,56)=19.213$, $p<0.001$, rare items had a much lower entropy at the fourth letter than frequent ones (0.844 vs 1.782), which indicates that there is less variability with the decrease of frequency. At the same time, the selection variable of late versus early decision point also had a significant effect ($F(1,56)=27.478$, $p<0.001$). The early uniqueness point items had much lower entropy (0.752) at the fourth letter than late decision point items (1.874). This can be taken as a *post hoc* support of our classification. Comparison of the eta squares as an estimate of variance explained shows that decision point had a stronger effect (0.329) than frequency (0.255). While there was a significant interaction, its impact is very weak in explained variance (0.09). Figure 5 shows the overall differences. These estimates have to be carefully interpreted. They certainly give a post hoc support for our selection of items, but one has to bear in mind that the items were manually selected.

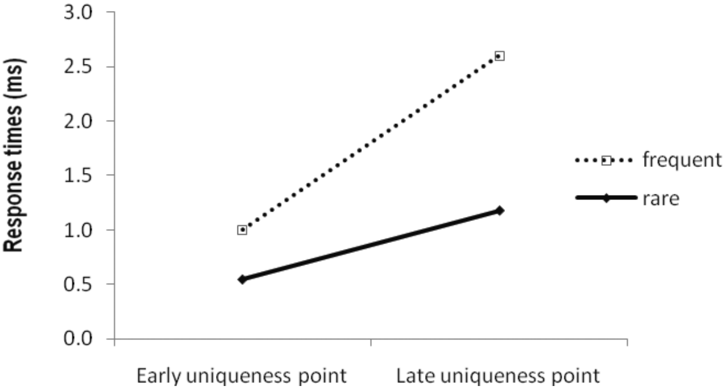

Figure 5. The entropy values at point four of the different items

Recognition and entropy

Based on the data matrix in Experiment II, for each event of successful recognition, the beginning at the recognition point was paired with the beginning at the immediately preceding gate (we call this one-before-recognition point). Figure 6 shows the means on the different measurements of the beginning metrics, for both the recognition points (dashed) and one-before-recognition points (solid). The data for each gate were aggregated (the x axis of the graphs).

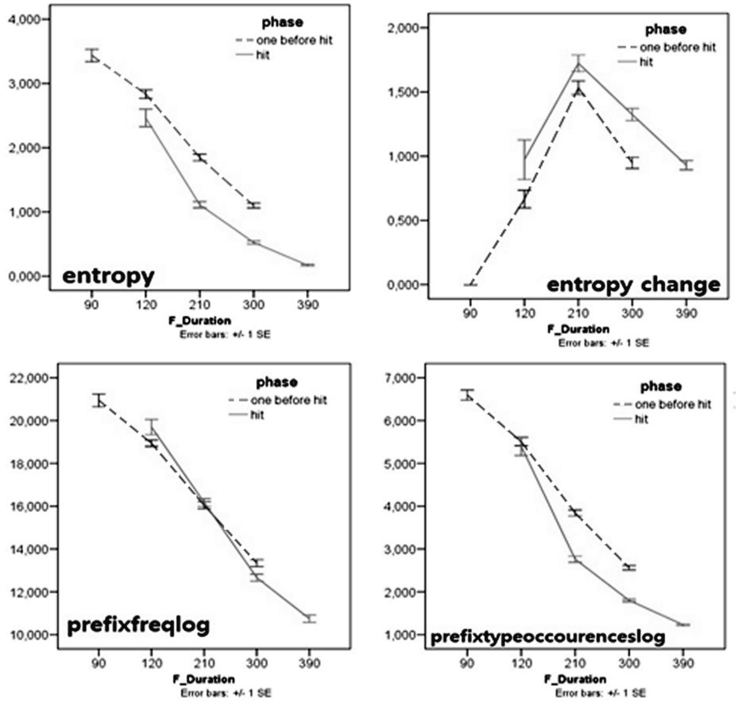

Figure 6. *Different entropy measures and recognition in the gating study*

In Experiment II, by applying the Mann-Whitney-Wilcoxon test, we verified that the means for the recognition points differ from the means for one-before-recognition points at a high significance level. This is true for each of the metrics we analyzed here. We note that this result is the least obvious for the entropy change measure, as it is a highly non-monotonous function of beginning-length. Intuitively, this means that **the recognition point follows a sudden drop of the entropy value**, which is the hypothesis we started from. On the other hand, the confidence of the subjects in their prediction (standardized on a subject level) has no statistically significant correlation with the entropy change of the gate.

Going further, we used linear regression to rank the above metrics according to their capacity to predict our observations. Table 1 shows the Akaike Information Criterion values (Akaike, 1974). Smaller values mean better fitting models.

Table 1. Information criterion values for the different predictors

Measure	Value
entropy	4939
beginning type occurences log	4998
entropy change	5018
beginning frequency log	5035

As we can see, the entropy measure proved to be the best linear predictor of recognition points. This result is in line with the model of Moscoso del Prado Martín, Kostić, and Baayen (2004). Namely, what is treated traditionally as a uniqueness point in gating research is actually a decrease in entropy. Entropy based measures proved to be better here than merely frequency, similar to the more complex model of Moscoso del Prado Martín et al. (2004).

Scrambling words and reconstruction

In the last decade a lot has been written on the transposed letter effect. Essentially, people are able to reconstruct letter transpositions while reading, and the beginnings and ends are especially crucial in this process as already shown by Bruner and O'Dowd (1958) with tachistoscopic presentations (e.g., Chambers, 1979, Perea, Duñabeitia, & Carreiras, 2008), such as reading *porblem* as *problem*. The essential issue of the different models is how to account for the broken coalition between letters and their positions in given words.

In looking for different possible informational factors in morphological processing, rather than merely stem words, two experiments were done in Hungarian to look for the reconstruction effects depending on the position of the misplaced letters, thus arguing for a positional effect in word access in an agglutinative language.

EXPERIMENT 3

The relative importance of error position in word recognition

Method

Subjects and procedure. Volunteer university students (n=40) in an off-line experiment administered over the Internet. They were asked to reconstruct misspelled words in writing.

Materials. Stem words and suffixed words were both used. In the list of 33 forms in one subgroup of items the beginning and the end of the word were correct, but the middle was misspelled (*szi**tim**kus* for *szim**pati**kus).* In another group of forms only the beginning was correct (*söt**bténe** for söt**étben*** 'in the darkness'), and in the third group only the ending was correct (***mlé**ény for **mell**ény* 'vest'). One third of the letters were reordered in every word in a way such that no letter stayed in its original place. Each group contained 11 words. The word length was between 5 and 15 letters.

Results

As Figure 7 shows, reconstruction was highest in the case when both the beginning and the end were correct, in line with the English classical data of Bruner and O'Dowd (1958) and Chambers (1979). Thus, in compensating for letter misplacement the beginnings that are crucial for lexical access, as well as the ends that are crucial for sentential role assignment are most attended to (both contrasts were significant $t(39)=6.95$ and $t(39)=7.50$, respectively, both $p<0.001$).

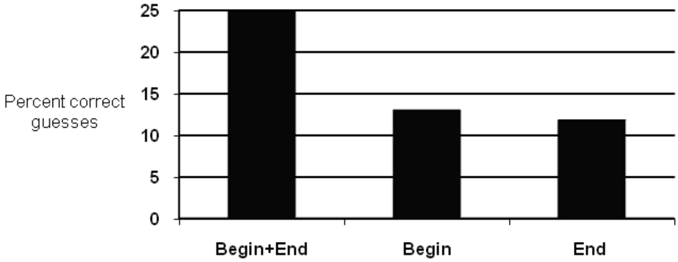

Figure 7. Words are better reconstructed if the beginning and end are both correctly spelled

EXPERIMENT IV.

A systematic study of letter misplacement (Várhelyi, 2010)

In a further study, more carefully selected stimulus materials were used, and the changes only involved two letters.

Method

Materials. Hungarian words of 7 and 8 letters with frequencies between 1000 and 100,000 in the MOKK corpus were used with no vowels in the critical to be exchanged positions.

Let us see the examples with the target word *probléma* ‘problem’.

1. mispositioning at the beginning messing up the first two letters RPOBLÉMA
2. mispositioning at the beginning messing but involving the second and third letters PORBLÉMA
3. mispositioning in the middle PROLBÉMA
4. mispositioning at the end but keeping the last letter intact PROBLMÉA
5. mispositioning at the end involving the last and the penultimate letter PROBLÉAM

In the actual lists the experimental material was interspersed with 120 pseudowords, also manipulated with scrambling. The use of pseudowords made the lexical decision task a relevant decision.

Procedure. A priming method was used, rather than active reconstruction. In each pair the prime was the misspelled word, and the target the correct one. Thus in the given example RPOBLÉMA as a prime was followed by PROBLÉMA as the target. Subjects had to decide if the target was a word. A 100 ms prime was followed by the target, with a 100 ms exposure time.

Subjects. Twenty volunteer university students took part in the experiment.

Figure 8 shows the basic position effects on reaction times. Smaller reaction times indicate more priming. A clear bathtub effect (Aitchison, 1987) was obtained: the distortions in the middle had the least effects on priming towards the target word. Indirectly this indicates that if the beginning and the end is correct in the prime, there is more facilitation towards the access of the target word.

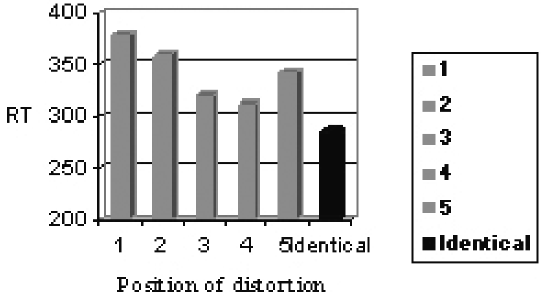

Figure 8. Priming effects of transposed letters depend on the position

Information uncertainty, morphological structure and lexical decision times.

The morphological structure and the linear composition of words has a clear effect on the recognition of complex words. Earlier studies in Hungarian used lexical decision times to look for the relative importance of different linear and grammatical parts of the word in lexical access and morphological analysis. Pléh and Juhász (1995) employed examples like (4) in a visual lexical decision study to look for the relevance of grammatical prefixes, stems, derivational and inflectional suffixes.

(4) Correct word: *el-intéz-és-é-ben Pref-deal-ing-his-in*
'in his dealing, Perf'
Prefix wrong combination ***meg**– intéz-és-é-ben*
Non existing prefix ***mag**-intéz-és-é-ben*
Stem wrong *el-**untéz**-és-é-ben*
Possessive wrong *el-intézés-é-ben*
Harmony violation in case *el-intézés-é-**ban***
Misspelled case *el-intézés-é-**beg***

Figure 9 shows the reaction times for the different locations of errors. A clear bathtub effect (Aitchison, 1987) was observed here as well. Errors in the middle were recognized much more slowly. This was interpreted as implying that in reading Hungarian people pay attention to the beginning for lexical access and to the end for grammatical role assignment.

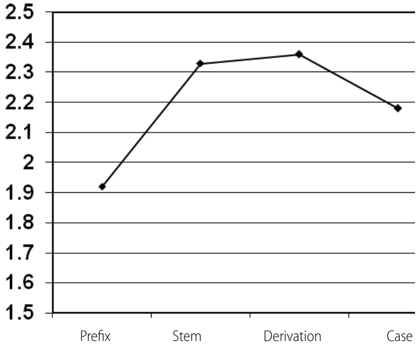

Figure 9. A clear bathtub effect in lexical decisions over multimorphemic Hungarian nouns (Pléh and Juhász, 1995)

EXPERIMENT 5

A systematic study of morphemic structure and linear position in lexical decisions

In this study an effort was made to obtain a more systematic analysis of the overall effects found by Pléh and Juhász (1995). In particular, we wanted to use the same stems as used in the gating studies, but in morphologically more complex word forms, and with a clearer control for possible frequency effects. Our further hope is to use more sophisticated entropy based predictions to reanalyze the results later.

Method

Materials and design. In the basic arrangement of the experimental material – prepared by Judit Fazekas – four basic morphological types were used:

- stem *böllér* 'sticker'
- plural *böllér-ek* 'stickers'
- case *böllér-nek* 'stickerDAT'
- Plur + case *böllér-ek-nek* 'stickersDAT'

Three types of markers (plural, second and first possessive) and three types of case *(– nak,-ban-,-ra* 'Dat, INSIDE, ONTO). Suffixes were used that all take the citation allomorph, since we did not want to complicate matters with issues of fusion and other morphological phenomena.

An entirely within subject design was used, where each subject saw each combination but the items were randomized between subjects. The general outline was: frequency (2 levels) x morphological structure (3 levels) in placing the errors (stem, marker, case). Within the suffix errors there were two subtypes: vowel harmony errors, and non-existing stems. Each subject had to decide about 360 words half of which were correct, and half of which were wrong. In the entire study 4700 items were tested.

Subjects. A group of 72 healthy students with ages ranging from 18 to 34 years (44 female (21.41 year; *SD*=2.44; 28 male (21.75 year; *SD*=1.86)) of the BME (*Budapest University of Technology and Economics*, Budapest, Hungary) participated in this experiment as volunteers. Every subject had normal or corrected-to-normal vision. They were tested individually and they all gave informed consent before the experiment and received partial credit for participating. None of the subjects had any prior experience with the experimental task.

Procedure. The task was a simple word/nonword decision task. The subjects watched the screen and had to decide about pop up words (that appeared on the center of the screen) that was a correct Hungarian word (press button 'I') or not (press button 'R'). Before the word/nonword appearance there was a fixation mark (duration was random between 600–1400 ms) on the centre of the screen.

Results

The main effect of frequency was significant ($F(1,71)=235.275$, $MSE=0.003$, $p<0.001$, $eta^2=0.768$). The error type had a main effect ($F(5,355)=136.644$, $MSE=0.004$, $eta^2=0.658$). Also, there was a significant interaction between frequency and error type ($F(5,355)=127.326$, $MSE=0.03$, $p<0.001$, $eta^2=0.642$).

Figure 10 shows the overall results for correctness of the results.

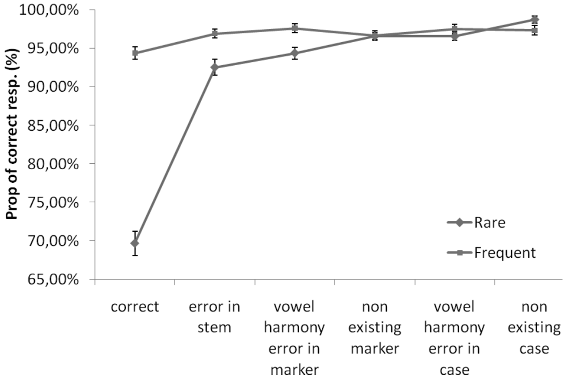

Figure 10. Correctness of acceptance-rejection as a function of word structure. The error bars show the standard error of mean.

Paired comparisons revealed a structure [1>2>3]> [4>5>6]. Basically, errors at the end of the word were recognized easier, more correctly.

Figure 11. shows the overall results for reaction times.

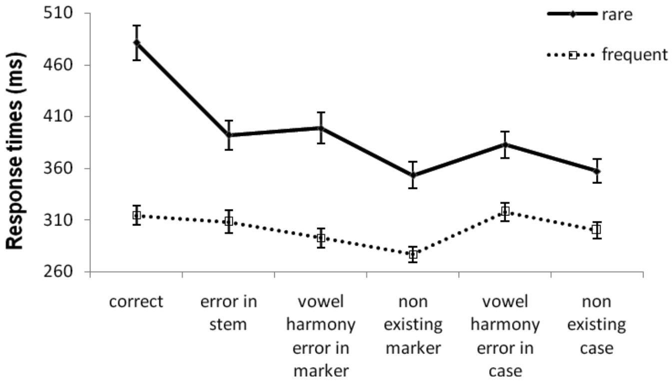

Figure 11. Reaction time data in the lexical decision task. The error bars show the standard error of mean.

The main effect of frequency was significant (*F*(1,71)=293.496, *MSE*=6307.829, *p*<0.001, eta^2=0.805). The error type had a main effect, but with much less explanatory power (F*(*5,355)=31.583, MSE=3613.921. eta^2=0.308). There was a significant interaction between frequency and error type (*F*(5,355)=29.119, *MSE*=1983.079, *p*<0.001, eta^2=0.291).

In paired comparisons the correct forms were slower to be accepted. Within the errors, non-existing stems were slower to be rejected than non-existing markers (i.e. word middle errors) or case markers. There was no clear bathtub effect in the reaction times. If anything, word end case marker errors were recognized slower.

The two vowel harmony errors were not different between each other. As expected, however, vowel harmony errors were slower to be rejected than non-existing phonemes at the given position. This is similar to the general observation introduced by the famous prefix-stripping effect: if you have to make decisions about an existing but misplaced morpheme, rejection takes longer, since it involves two steps (Taft, 1979).

Separate analyses were also made of frequent and rare stem items. Basically for correct performance the significant effects are obtained for both sets of errors. However, in frequent stems the word middle distortion towards a non-existing form (non-existing marker) leads to surprisingly low performance as shown on Figure 12. That also resulted in being no significant paired difference here between stem errors and word middle errors.

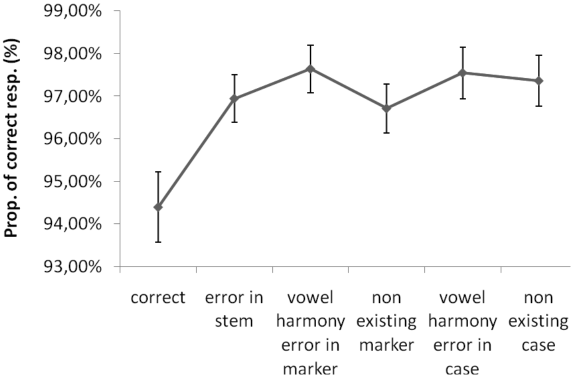

Figure 12. In frequent items word middle distortions lower performance. The error bars show the standard error of mean.

At the same time, reaction times with these items were rather fast in both groups as Figure 13 shows. Thus, there might be a trade-off here.

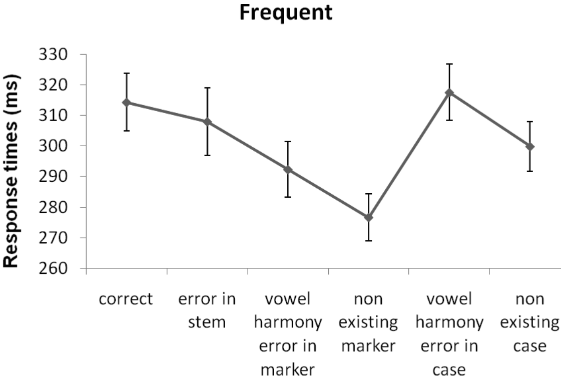

Figure 13. Reaction times in both frequent and rare items are fast in word middle non-existing form errors. The error bars show the standard error of mean.

GENERAL DISCUSSION

The series of experiments have provided some relevant data on lexical access and morphological processing in a strongly agglutinative language, Hungarian. Out of the five issues outlined in the introduction, the data have implications regarding three of them.

Regarding segmentation and lexical access both the results of the gating experiments (Experiment I and II) and those of the scrambling studies (Experiments IV and V) are relevant. In accord with data supporting the cohort theory of Marslen-Wilson (1975, 1984) the onset of words is very crucial in lexical access and processing of root morphemes. According to the gating results a syllable length segment can lead to 50 % of correct recognition. This is related to the uniqueness point issue as well, i.e., competing words have a top down inhibitory effect on recognition. The more rivals are there to a word, the later the recognition point. Therefore, early uniqueness words are recognized earlier. This effect, however, disappears in rare words. Rare words seem to be recognized in a more strictly bottom up manner. A more specific effect found was the effect of grammatical constraints, which, given the instructions, can be seen as a specific top down effect. The scrambling studies indicate a strong word onset primacy in Hungarian as well, similar to effects found over many other languages.

The entropy effects are by far not trivial. The low entropy connected to early uniqueness points shows that the uniqueness points used so many times in fact are related to a more objective measure, the entropy at the given point. Thus, we suggest entropy to be used in further studies instead of the uniqueness point. The uniqueness point is only sensitive to the absolute values, while entropy and its change is sensitive to a frequency distribution. In a nutshell, entropy change is important in explaining neighbourhood effects.

The relationships between entropy measures and recognition performance should be seen with a careful epistemological attitude. The entropy values are derived from a large impersonal data set. It is far from trivial that these objective, external characterizations have a determining power over individual recognition performance. In the details of this, determining power entropy values had more explanatory power than mere frequency. Thus, the studies that still have to be continued strongly support the original insights of Kostić (1995) regarding the use of information metrics in studying the underlying mechanisms both of lexical access and morphological decomposition.

REFERENCES

Aitchison, J. (1987). *Words in the mind.* London: Blackwell.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEE Transactions on Automatic Control. 19,* 716–723.

Antal, L. (1962). A megnyilatkozások tagolása morfémák szerint [Segmentation of utterances into 'morphemes']. *Magyar Nyelvőr, 86*, 189–202.

Antal, L. (1964). *A formális nyelvi elemzés* [Formal linguistic analysis]. Budapest: Gondolat.

Baayen, R. H. (2005). Data Mining at the Intersection of Psychology and Linguistics. In A. Cutler (Ed.), *Twenty-Fist Century Psycholinguistics: Four Cornerstones* (pp.69–83). Mahlah, N.J: Erlbaum.

Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. (2005). Onset entropy matters – letter-to-phoneme mappings in seven langauges. *Reading and Writing, 18,* 211–229.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bruner, J.S. & O'Dowd, D. (1958). A Note on the Informativeness of Parts of Words. *Language and Speech, 1,* 98–101.

Chambers, S. (1979). Letter and order information in lexical access. *Journal of Verbal Learning and Verbal Behavior, 18,* 225–241.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of languages. *Handbook of mathematical psychology. Vol. II.* New York: Wiley.

Chomsky, N. (1957). *Syntactic structures.* Hága: Mouton.

Daelemans, W., & van den Bosch, A. (2005) *Memory-Based Language Processing.* Cambridge: Cambridge University Press.

Fazekas, J., Fehér K., & Szabó, E. (2008). *Some off-line sentence understanding demonstrations.* Seminar paper, Dept of Cognitive Science, BME, Budapest.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*, 267–283.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11*, 597–604.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (Prentice Hall Series in Artificial Intelligence) (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.

Kiefer, F. (Ed.). (2000). *Strukturális magyar nyelvtan. 3. Alaktan* [Structural Hungarian grammar. 3. Morphology]. Budapest: Akadémiai Kiadó.

Kornai, A. (1992). Frequency in morphology. In I. Kenesei & Cs. Pléh (Eds.), *Approaches to Hungarian. Vol. 4.* (pp. 247–268). Szeged: JATE.

Kornai, A. (1994). *On Hungarian morphology.* Budapest: MTA Nyelvtudományi Intézete.

Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., & Varga, D. (2006). Web-based frequency dictionaries for medium density languages. In A. Kilgarriff & M. Baroni (Eds.), *Proceedings of the 2nd International Workshop on Web as Corpus* (pp. 1–9). ACL–06.

Kostić, A. (1995). Information load constraints on processing inflected morphology. In Feldman, L.B. (Ed.), *Morphological aspects of language processing* (pp. 317–344). Mahlah, NJ: Erlbaum.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science, 189,* 226–228.

Marslen-Wilson, W. D. (1984). Spoken word recognition: a tutorial review. In H. Bouma & D. G. Bouhwis (Eds.), *Attention and performance X: Control of language processes* (pp. 125–150). Hove, UK: Lawrence Erlbaum Associates.

Marslen-Wilson, W. D., Komisarjevsky Tyler, L., Waksler, R., & Older, R. (1994). Morphology and Meaning in the English Mental Lexicon, *Psychological Review*, *101*(4), 3–33.

Milin, P., Filipović Đurdević, D., &, Moscoso del Prado Martín, F. (2009). The Simultaneous Effects of Inflectional Paradigms and Classes on Lexical Recognition: Evidence from Serbian. *Journal of Memory and Language, 60,* 50–64.

Miller, G. A. (1951). *Language and communication.* New York: McGraw Hill.

Miller, G. A. (1956). The magical number seven, plus or minus two. *Psychological Review, 63*, 81–97.

Miller, G. A., & Selfridge, J. A. (1950). Verbal Context and the Recall of Meaningful Material. *American Journal of Psychology, 63,* 176–185.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In D. R. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (*vol. ii*). New York: John Wiley.

MOKK (2006). A WEB based Hungarian frequency dictionary: http://szotar.mokk.bme.hu/szoszablya/searchq.php

Moscoso del Prado Martín F., Bertram, R., Shreuder, R., & Baayen, R. H. (2004). Morphological Size in a Morphologically Rich Language: The Case of Finnish Compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1271–1278.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the Bits Together: An Information Theoretical Perspective on Morphological Processing. *Cognition, 94,* 1–18.

Niemi, J., Laine, M., & Tuominen, J. (1994). Cognitive Morphology in Finnish: Foundations ofa New Model, *Language and Cognitive Processes*, *9(3),* 423–446.

Papp, F. (1982). *A magyar főnév paradigmatikus rendszere* [The paradigmatic system of Hungarian nouns]. Budapest: Akadémiai Kiadó.

Perea, M., Duñabeitia, J. A., & Carreiras, M. (2008). Transposed-letter priming effects for close vs. distant transpositions. *Experimental Psychology, 55*(6), 384–93.

Pléh, Cs., & Juhász, L. (1995). Processing of multimorphemic words in Hungarian. *Acta Linguistica Hungarica, 43*, 211–230.

Saffran, J. R, Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379–423. http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

Shannon, C. E., & Weaver, W. (1949). *A Mathematical Model of Communication.* Urbana, IL: University of Illinois Press.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition, 7*, 263–272.

Várhelyi, K. (2010). *TLE and position. A comparison of the SOLAR, open-bigram and SERIOL models.* Seminar paper, Dept of Cognitive Science, BME, Budapest.

Wurm, L. H., Ernestus, M., Shreuder, R., & Baayen, R. H. (2006). Dynamics of the Auditory Comprehension of Prefixed Words: Cohort Entropies and Conditional Root Uniqueness Points. *The Mental Lexicon, 1,* 125–146.

Appendix

The nouns with their parameters from the lexical database together with the entropy measures. The absolute frequencies are orom a corpus of 589 million word tokens. Source: *the* MOKK *corpus* (http://szotar.mokk.bme.hu/szoszablya/searchq.php)

Table 1. Rare words

Category	**Word**	**frequency**	**lemma-frequency**	**uniqueness point**	**First 4**	**entropy at point 4**
early uniqueness point (mean 71.26 ± 3,10)	böllér	86	191	3	böll	0,603096
	cécó	112	196	3	cécó	0
	csöbör	63	151	3	csöb	0
	dunna	77	186	3	dunn	0,652352
	dzsúdó	128	173	3	dzsú	1,139827
	güzü	168	172	3	güzü	0,583207
	gyűszű	57	166	3	gyűs	1,472454
	kégli	66	181	3	kégl	0
	lucsok	66	156	3	lucs	0,777575
	nábob	121	174	3	nábo	0
	pőcsik	6	7	3	pőcs	0
	rücsök	8	45	3	rücs	0,49999
	sasszé	34	128	3	sass	1,849419
	user	14	34	3	üzér	0,443576
	zsepi	63	151	3	zsep	0
late uniqueness point (mean 63,20 ± 3,12)	bögöly	90	161	5	bögö	0
	cinke	3	185	5	cink	1,948223
	dublőr	21	82	5	dubl	1,594457
	krokett	46	160	5	krok	1,383661
	pányva	28	114	5	pány	1,231779
	pincsi	15	108	5	pinc	2,014432
	polka	100	168	5	polk	0,289986
	poorly	43	164	5	pörö	1,022923
	rozmár	80	188	5	rozm	0,560543
	stóla	62	187	5	stól	0
	trojka	93	159	5	troj	0
	tartar	48	80	6	tart	3,096681
	stangli	17	67	7	stan	1,35673
	strázsa	152	350	7	strá	1,783065
	svindli	150	198	7	svin	1,001279

Table 2. Frequent words

Category	**Word**	**frequency**	**lemma-frequency**	**uniqueness point**	**First 4**	**entropy at point 4**
early uniqueness point (mean 31194,07 ±2662,91)	asszony	50211	87569	2	assz	1,540912
	forum	167634	416906	3	fóru	0
	szoftver	31974	71251	3	szof	0.076625
	utca	50381	141877	3	utca	1,019915
	papa	25598	34252	3	pápa	1,069351
	szféra	11602	26148	3	szfé	0.045589
	tonna	11982	18232	3	tonn	0,597695
	üveg	12906	28887	3	üveg	3,004682
	kenyér	10883	33280	3	keny	1,081985
	típus	13921	52944	3	típu	1,068367
	műsor	17263	76473	3	műso	1,100964
	ablak	19222	70943	3	abla	0,837288
	dollar	13876	38662	3	doll	0.028648
	ünnep	12579	50454	2	ünne	2,838832
	japán	17879	36339	3	japá	0,213908
late uniqueness point (mean 80346,3333 ±2557,32)	család	71077	218999	5	csal	2,471073
	kérdés	133771	573154	5	kérd	1,426039
	oldal	104786	487728	5	olda	0,752174
	személy	103517	292109	5	szem	4,187367
	tanár	103452	207561	5	taná	2,713847
	válasz	71364	210854	5	vála	3,116823
	város	171466	337930	5	váro	1,767857
	tanács	95137	182986	5	taná	2,713847
	termék	52327	203634	5	term	3,705582
	csapat	52767	137308	5	csap	2,467515
	verseny	59557	147910	6	vers	3,023900
	század	70603	148126	5	száz	2,258290
	osztály	39483	131650	5	oszt	3,421070
	nemzet	42976	81750	6	nemz	2,381595
	vizsga	32912	93346	5	vizs	2,540940