Relative entropy effects on the processing of spoken Romanian verbs

Nenadić, Filip; Milin, Petar; Tucker, Benjamin V.

Relative Entropy Effects on the Processing of Spoken Romanian Verbs

Filip Nenadić and Benjamin V. Tucker

University of Alberta

Petar Milin

University of Birmingham

Author Note

Abstract

A multitude of studies show the relevance of both inflectional paradigms (word form frequency distributions, i.e., inflectional entropy) and inflectional classes (whole class frequency distributions) for visual lexical processing. Their interplay has also been proven significant, measured as the difference between paradigm and class frequency distributions (relative entropy). Relative entropy effects have now been recorded in nouns, verbs, adjectives, and prepositional phrases. However, all of these studies used visual stimuli — either written words or picture-naming tasks. The goal of our study is to test whether the effects of relative entropy can be captured in the auditory modality as well. Forty young native speakers of Romanian (60% female) living in Serbia as part of the Romanian ethnic minority participated in an auditory lexical decision task. Stimuli were 168 Romanian verbs from two inflectional classes. Verbs were presented in four forms: present and imperfect 1st person singular, present 3rd person plural, and imperfect 2nd person plural. The results show that relative entropy influences both response accuracy and response latency. We discuss alternative operationalizations of relative entropy and how they can help us test hypotheses about the structure of the mental lexicon.

Keywords: relative entropy, inflectional paradigm, inflectional class, spoken word recognition, auditory lexical decision

Relative Entropy Effects on the Processing of Spoken Romanian Verbs

In this paper we continue a line of research that investigates how speakers of a language learn its probabilistic features, and how these probabilities predict language processing, i.a., responses to morphologically inflected word forms. Our study is grounded in the Word and Paradigm approach to morphology (see, e.g., Blevins, 2003; Hockett, 1954), and uses measures developed within information theory (Shannon, 1948) to formalize patterns in inflectional morphology that speakers learn. Previous studies used visual stimuli to show that responses in various experimental tasks are modulated not only by word frequencies or relative frequency distributions of words and their inflected forms, but also the prototypicality of the relative frequency distribution when compared to other words in its class (see, e.g., Filipović Đurđević & Milin, 2019; Milin, Filipović Đurđević, & Moscoso del Prado Martín, 2009). Our experiment investigates whether this phenomenon extends to Romanian and whether it is observable in spoken word recognition as well.

Inflectional and relative entropy

Highly inflecting languages, such as Serbian or Romanian, use form endings to create different meanings for a single word lemma. All form endings used in a particular lemma are referred to as the lemma's inflectional paradigm. Word lemmas sharing the same pattern of morphological transformations to create inflected forms (i.e., that have matching paradigms) are said to belong to the same inflectional class. A number of psycholinguistic studies have shown that listeners are senstive to characteristics of paradigms and classes by connecting these characteristics to performance in behavioral experiments (e.g., Baayen, Feldman, & Schreuder, 2006; Baayen & Moscoso del Prado Martín, 2005; Bertram, Laine, Baayen, Schreuder, & Hyönä, 2000; Colé, Beauvillain, & Segui, 1989; Hay & Baayen, 2005; Kostić, Marković, & Baucal, 2003; Moscoso del Prado Martín, Kostić, & Baayen, 2004; Taft, 1979).

One of the central findings in these studies is that the more balanced the occurence frequency of inflected forms for a word lemma, the faster the response times are for all forms of that lemma. This relative frequency distribution of a word lemma's inflectional paradigm is ordinarily quantified using informational entropy (H; Shannon, 1948), as described in Moscoso del Prado Martín et al. (2004) and presented in Equation 1. A word lemma's inflectional entropy is dependent on the ratio of frequencies of each of its inflected word forms $f(w_i)$ and the sum frequency of all the forms of the word lemma $f(w)$, i.e., stem frequency. In other words, inflectional entropy depends on the probability distribution of particular word forms $p(w_i)$. As a result, inflectional entropy corresponds to difficulty of form prediction and is high for paradigms with forms of similar relative frequencies; such words are responded to faster than words that have low inflectional entropy.

$$H(w) = -\sum_{i=1}^{n} \frac{f(w_i)}{f(w)} log_2 \frac{f(w_i)}{f(w)}$$
$$= -\sum_{i=1}^{n} p(w_i) log_2 p(w_i)$$

(1)

Milin et al. (2009) investigated the interplay between a noun's paradigm and its class's relative frequencies. The authors calculated how typical an inflectional paradigm is for the particular inflectional class the noun belongs to. The measure of prototypicality was relative entropy (RE) expressed as the Kullback-Leibler divergence. Equation 2 shows the formula used, where $p(i)$ refers to the distribution of a noun's paradigm, and $q(i)$ refers to the distribution of the entire class. In more detail, relative entropy depends on the ratio of frequencies of each of its inflected word forms $f(w_i)$ and the sum frequency of all the forms of the word lemma $f(w)$, but also the ratio of word form frequencies $f(e_i)$ and the sum frequency $f(e)$ calculated for the entire inflectional class ρ. Modeling the results of a visual lexical decision task conducted using Serbian nouns as stimuli showed that response latencies tend to become longer as relative entropy increases — if a particular noun's relative form frequency is less similar to the prototypical (mean) distribution derived from all the nouns in its class, participants have more difficulties processing it.

$$RE = \sum_{i\epsilon\rho}^{n} f(w_i)/f(w) log_2 \frac{f(w_i)/f(w))}{f(e_i)/f(e))}$$
$$= \sum_{i\epsilon\rho}^{n} p(i) log_2 (\frac{p(i)}{q(i)})$$

(2)

Baayen, Milin, Filipović Đurđević, Hendrix, and Marelli (2011) provided further evidence for the relevance of prototypicality in language processing. Besides replicating the findings from Milin et al. (2009), the authors also showed that relative entropy effects occured in sentential reading using a self-paced reading task combined with priming, again using Serbian nouns as targets. Furthermore, the prototypicality effect extends to the prototypicality of preposition-noun pairings in English, as nouns are processed with more difficulties if the relative frequency with which they are paired with prepositions is unusual.

Since then, relative entropy effects have also been recorded in Serbian verbs (Filipović Đurđević & Gatarić, 2018) and adjectives (Filipović Đurđević & Milin, 2019), again using the visual lexical decision task. Moreover, similar effects were recorded using using online measures as well, in the processing of prepositional phrases in English using a primed picture naming task and event related potentials (Hendrix, Bolger, & Baayen, 2017).

The current study

We have seen from the brief overview of the literature that speakers are not only sensitive to patterns of morphological form, but also that they are sensitive to discrepancies of a particular item's pattern in comparison to the rest of its class. These findings arguably indicate that the structure and/or access to the mental lexicon is shaped by learned patterns observable in paradigms and in groups of paradigms (classes). However, all previous studies used visual stimuli, most often written words or sentences, and sometimes imagery in a picture-naming task.

The principal novelty in our study is the change of the sensory modality used, as we investigate spoken word recognition rather than reading words or naming pictures. As Goldinger (1996, pp. 564) stated, "generalising across modalities must be ventured cautiously" — and there are many reasons why we should investigate whether any effects recorded in reading experiments are present in the auditory modality as well. A person's experience with spoken language may be different to their experience with written language. Listening to spoken words has a more pronounced temporal dimension than reading the same text: a single word is seen in one or two fixations of the eye (Rayner, Chace, Slattery, & Ashby, 2006), while most current models of spoken word recognition place an emphasis on the speech stimulus unfolding in time (Magnuson, Mirman, & Harris, 2012; Weber & Scharenborg, 2012). Furthermore, visual material can be accessed again via longer fixations or a regression, while words presented in the auditory modality ordinarily cannot be replayed. Finally, certain effects are not the same in the visual and the auditory modality. For example, the effects of phonological neighborhood density are opposite in visual versus auditory modality (see Vitevitch & Luce, 1998; Yates, Locker, & Simpson, 2004), and some semantic richness effects are noted in the visual modality only (Goh, Yap, Lau, Ng, & Tan, 2016).

Additionally, instead of investigating the phenomenon in English or Serbian language, which are predominant in the literature about relative entropy effects, we test the existence of the effect in Romanian. In fact, to the best of our knowledge, this study is the first instantiation of an auditory lexical decision experiment presented in Romanian. The Romanian language belongs to the group of Indo-European languages, and is similar to other languages in the Romance group such as Italian or French. Similarly to those languages, but also similarly to Serbian, Romanian has a relatively complex inflection system.

In this study, we focus on effects of relative entropy in verb processing, for which a visual lexical decision was conducted in Serbian (Filipović Đurđević & Gatarić, 2018). Romanian regular verbs are traditionally separated into four conjugations, distinguished by their respective verb endings in infinitive form: verbs ending in -a (e.g., *a vindeca*, to heal), verbs ending in -ea (*a sedea*, to sit), verbs ending in -e (*a scrie*, to write), and verbs ending in -i or -î (*a veni*, to come). However, a newer classfication recognizes 11 different verb conjugations (Gönczöl, 2007), while an even larger number of

existing conjugation models is described on https://dexonline.ro/modele-flexiune/V and in the resources provided by RACAI's Linguistic Web Services (Tufiş, Ion, Ceauşu, & Ştefănescu, 2008). We use four tense forms of verbs ending in -a and -i and belonging to models of conjugation (inflectional classes) marked as *201* and *401*, described more thoroughly in the Methods section. These inflectional classes are some of the more frequent ones in Romanian, providing a larger number of verbs to calculate relative entropy, i.e., the difference between probability distribution of verb paradigms and the probability distribution of their respective inflectional classes. We presented these verbs together with verb-like Romanian pseudowords to a group of Romanian-Serbian bilinguals, and find that effects of relative entropy can be captured in the auditory modality as well.

Method

Participants

Participants in this study were 40 young speakers of Romanian (60% female) that live in Serbia as part of the Romanian ethnic minority. All participants were bilinguals speaking Romanian in a dialect characteristic to the Serbian Banat region and Serbian. Participants finished elementary school and high-school in Romanian and lived in Romanian-speaking communities. Thirty-eight participants were tested during their last week in high-school (18 years old), while two participants were older and already students at the department of Romanian Language and Literature, University of Novi Sad, Serbia. There was no compensation offered for participating in the experiment.

Stimuli

Stimuli were verbs retrieved from an online Romanian dictionary available at https://dexonline.ro. We used the list of officially admitted Romanian words for the game *Scrabble*. We selected four forms for 168 verb lemmas from conjugational models (inflectional classes) *201*, ending in -a, and *401*, ending in -i. The forms we selected (present first person singular, present third person plural, imperfect first person singular, and imperfect second person plural) overlapped with other word forms of the same verb, as presented in Table 1. For simplicity, we will refer to particular word forms as combinations of the verb ending (-i or -a) and one of these four forms that were selected for the experiment. Given that in the inflectional class *401* present first person singular and present third person plural also overlap, this yields a total of seven different combinations: -a.P1S, -a.P3P, -a.I1S, -a.I2P, -i.P1S/P3P, -i.I1S, -i.I2P. Other information about verbs, such as their inflectional class, part of speech tags, and frequency, were obtained from the Romanian Balanced Corpus (ROMBAC), a corpus provided by the RACAI's Linguistic Web Services (Tufiş et al., 2008). ROMBAC is a corpus of 41 million words that contains an equal representation of five genres (journalism, legalese, fiction,

medicine, and biographical data from Romanian literary personalities).

The selection of specific verb classes was made based on how pervasive they were in the Scrabble list and ROMBAC, as we chose the most frequent classes. Verb lemmas were selected for having at least three of the verb forms we were interested in accounted for in the ROMBAC. Some additional selection of verb lemmas was made based on recording quality and then finally at random, to reduce the number of lemmas in order to make an experiment of manageable length. Specific verb forms were selected to cover two simple tenses as using a complex tense requires presenting the participant with two words (note that the perfect tense in Romanian can be constructed as a simple tense, but that the complex form is far more prevalent). Within the present and the perfect tense, we selected specific forms to cover both singular and plural forms and first, second, and third person.

– Table 1 should be placed approximately here –

Since the participants are presented with isolated words in the auditory lexical decision experiment and cannot know which of the syncretic forms of the word is intended, we calculated word frequencies as a sum of all overlapping word forms (e.g., summing the frequencies of *abonam* as imperfect first person singular and as imperfect first person plural). Word frequency ranged from 0 to 11,654 (Q1 = 1, Q3 = 33).

The experiment also contained 168 pseudowords generated using LINGUA (Westbury, Hollis, & Shaoul, 2007). The pseudowords were in accordance with Romanian orthography and matched the word stimuli in the number of phonemes. Each selected pseudoword ended in one of the seven verb form endings present in the word stimuli, but neither the "pseudostem" nor the entire pseudoword had meaning in Romanian.

Stimuli were recorded in a recording studio using a *Neumann U87* microphone. The recording was conducted by a professional radio technician, and words and pseudowords were read by a professional female radio announcer. The announcer first read the word list, and afterwards the pseudoword list. Pauses were made during reading, and stimuli read with an error were re-recorded. The order of the stimuli in both lists was randomized.

Four experimental lists were created, each containing the same 168 pseudowords and each verb lemma in one of its four verb forms. Due to an error made in the counterbalancing procedure, each lemma ocurred only once per list, but not all lists had the same number of occurences of particular verb endings. This, in addition to an already existing overlap between present first person singular and present third person plural in the conjugational model *401* noted in Table 1, led to the experimental lists having a degree of imbalance in the frequencies of each of the verb form endings. However, statistical analyses showed that this imbalance did not influence participant performance, as no significant differences were found between experimental groups (see below).

Procedure

The participants were tested in groups of three or fewer in a spacious classroom, seated far apart from one another and facing different directions. Before the beginning of the experiment, the participants were given sixty seconds to complete as much as they could in a variant of a chain test (Ek, Fellenius, & Jacobson, 2003; Lindgren & Laine, 2011). The task in the chain test is to draw boundary marks between words in a concatenated text without any punctuation marks. In our experiment, we used the first paragraph of a children's story titled *Zâna munților* by Petre Ispirescu. Results on the chain test were intended to be used as a measure of language proficiency and as a criterion for potential participant exclusion.

After completing the chain test, each participant was assigned to one of the four experimental lists. The stimuli were presented using E-Prime experimental software (Schneider, Eschman, & Zuccolotto, 2002) and headphones. Each trial started with a visual fixation point which lasted 700 ms. The fixation point was followed by the audio stimulus at which point the participants made a lexical decision by pressing one of the two designated keyboard buttons. Responses could be made during the audio stimulus presentation. If the participant failed to respond within 3000 ms, the experiment would progress to the next fixation point and stimulus. The main block was preceded by a practice block with five stimuli. The participants were instructed to respond as quickly and as accurately as possible. The entire experimental session lasted approximately 15 minutes.

Data analysis

All analyses were conducted in the statistical software environment R (R Core Team, 2018), utilizing packages *mgcv* (Wood, 2011) and *itsadug* (van Rij, Wieling, Baayen, & van Rijn, 2017) for the main analyses. Word frequencies increased by 1 were log-transformed (we will refer to this variable simply as frequency). We registered the number of consecutive word stimuli including the one when the participant responded (run length), trial number, and stimulus duration in ms. Chain test score was calculated as the difference between the number of correctly and incorrectly made word partitions. All of the previously mentioned variables were standardized, centered on zero.

Participant response accuracy and response latency were analyzed separately. In the response latency analysis, only correct responses were considered. We analyzed response accuracy using generalized additive mixed-effects logistic regression and response latency using generalized additive mixed-effects regression. The predictor structure was the same in both cases. The models included the variables mentioned above (frequency, run length, trial number, stimulus duration in ms, chain test score) and inflectional entropy (H) and relative entropy (RE). The models also included interactions between RE and frequency, between RE and stimulus duration in ms, word Form (i.e., the seven combinations of conjugational model and part of speech described previously), and a factor controlling

for the experimental session (Group). Random intercepts for stimuli and random smooths of trial number per participant were included in the models as well. Besides observing whether a predictor is significant, we also used the *compareML* function to assess whether an iteration of the model is better than its predecessor.

We acknowledge that the inflectional classes and word forms used in this study do not represent the entire span of classes and forms in Romanian verbs. Therefore, there is an argument to be made towards using the Form variable as a random, rather than a fixed effect. We decided to present the Form variable as a fixed effect, but we explored both options in our analyses and found no qualitative differences in the results.

Results

The participants answered correctly to 74.25% of the trials (word accuracy 78.42%, pseudoword accuracy 70.07%). We excluded three participants that had less than 60% correct responses to all stimuli, as shown in Figure 1. All participants had an accuracy rate higher than 55% for words. We address this somewhat lower response accuracy in the Discussion.

– Figure 1 should be placed approximately here –

We continued our analysis on responses to words only, following the guidelines given in Baayen and Milin (2010). We removed unrealistic responses shorter than 100 ms and approximated a normal distribution of response latency by using Box-Cox transformation (Box & Cox, 1964) with -0.4 set as the exponent. After the transformation, overly long responses were excluded as well. In total, 2.97% of data was excluded. We examined the distribution of response latency per participant using quantile-quantile plots and tested their normality using Shapiro-Wilk tests, finding all of them to be satisfactory.

We then created separate datasets to analyse response accuracy and response latency using generalized mixed effects (logistic) regression. However, our models showed suppression effects between inflectional entropy (H) and relative entropy (RE), as exclusion of H or RE from the model made the other variable highly significant (whereas their effect was suppressed when included together). We determined that the issue comes from a small number of 12 lemmas with RE values higher than 1.9 which created a long tail in the skewed distribution of RE. The lemmas in this outstretching region of RE with very few data points also happened to have low H, falsely increasing the correlation between H and RE. At the same time, due to the small number of data points, the error bands in model estimates were very wide towards this edge of RE values.

At this point, we decided to exclude these 12 lemmas from further analysis (7.73% of remaining data removed). The following subsections present the results of modeling response accuracy and response latency without these lemmas, i.e., are based on the data from 37 participants responding to

156 verb lemmas. This exclusion somewhat limits the range of RE values that we investigate in the present analysis; as we mention in the Discussion, larger datasets might be needed to reliably encompass a broader range of values of inflectional and relative entropy.

The dataset that contained both correct and incorrect responses to test the effects of predictors on response accuracy included 5,604 rows total with 156 verb lemmas and 589 different verb forms. The dataset that tested the effects of predictors on response latency excluded all incorrect responses (19.31% of the remaining data) and consisted of 4,522 total rows, with 156 verb lemmas and 587 verb forms (note that for two verb forms only incorrect responses were made).

Response accuracy

The initial model iterations showed that the particular experiment Group, run length, chain test score, trial number, stimulus duration in ms, and the interaction between RE and frequency are insignificant predictors of response accuracy. The final model, presented in Table 2, included the random effects in addition to word Form (i.e., the combination of inflectional class and verb form), frequency, H, RE, and the interaction between RE and stimulus duration in ms as significant predictors of response accuracy. Note that the intercept in the section of the table presenting parametric effects represents the -i.I2P form (imperfect second person plural in the -i inflection).

– Table 2 should be placed approximately here –

The parametric effect of Form was noted (Figure 2). Wald tests (Table A1) showed that the most difficult was the imperfect second person singular in the -i inflection (-i.I2P). Only the present third person plural in the -a inflection (-a.P3P) was not significantly easier than this Form level. In turn, -a.P3P was more difficult than only -a.P1S and -a.I1S forms. Finally, -a.I1S was also significantly easier than -a.I2P. When a more conservative Bonferroni correction is applied, only two contrasts remain significant — when -i.I2P is compared to -a.P1S and -a.I1S.

– Figure 2 should be placed approximately here –

Although included as smoothed terms, frequency and H had a linear relationship with response probability. Higher frequency words were more often responded to correctly. Inversely, higher H values were connected to lower accuracy (left side of Figure 3). For RE, a similar declining trend is visible, although the relationship is not fully linear, as presented on the right side of Figure 3. There is a higher probability that words with higher RE will be misclassified as pseudowords. This effect also seems larger than the effect of H.

– Figure 3 should be placed approximately here –

Finally, the interaction between duration in milliseconds and RE indicated that the effects of RE are larger in shorter words, and become less pronounced for longer words. In Figure 4, this is represented by the topographical lines becoming more and more diverging as word duration increases

— creating a larger zone of equality between words of different RE.

– Figure 4 should be placed approximately here –

Response latency

The initial model predicting response latency showed that the particular experiment Group, word run length, H, and the interactions between RE and duration, and RE and frequency, should be excluded. The final model included the random effects in addition to the word Form, chain test score, trial number, stimulus duration in ms, frequency, and RE (Table 3). The intercept in the section of the table presenting parametric effects again represents the -i.I2P form.

– Table 3 should be placed approximately here –

The effect of word Form is presented in Figure 5. Participants responded to -i.I2P and -a.P3P more slowly than to -I.P1S/P3P, -a.I2P, and -a.P1S (Table A2). Other combinations of inflectional class and word form did not differ significantly to one another. When the more conservative Bonferroni correction is applied, the only contrast that remains significant is the one comparing -a.P3P and -a.I2P.

– Figure 5 should be placed approximately here –

The effects of trial number and stimulus duration in ms were linear. Later trials and shorter stimuli were associated with faster responses. Higher frequency was also associated with faster responses, but this effect is reduced with the most highly frequent words. Higher scores on the chain test were also connected to shorter response latency, but this relationship existed only for average to high scores on the test. Participants with the lowest scores on the chain test did not respond slower than their peers with lower-average scores on the same test. Crucially, we also see that RE is a significant predictor of response latency — an increase in RE leads to longer response latencies (Figure 6).

– Figure 6 should be placed approximately here –

Discussion

The results of our experiment indicate that the task was quite difficult for our participants. Error rates for both word and pseudoword stimuli were higher not only in comparison to other studies investigating relative entropy in verb recognition using written stimuli (Filipović Đurđević & Gatarić, 2018), but also in comparison to large databases of auditory lexical decision tasks in other languages (e.g., Ernestus & Cutler, 2015; Ferrand et al., 2018; Tucker et al., 2019). A number of factors could have played a role in causing the decrease in accuracy. Primarily, these relate to participant characteristics and difficult stimuli.

Most of the participants were in their last days of high school and were therefore younger in comparison to University students ordinarily tested in similar experiments. It may be expected that

University students are selected to have a certain, high degree of language competencies, whereas there is no clear indication of how developed these competencies were in our participants, nor whether they would continue to pursue higher education. Participants may have also been less interested or motivated than a University student participating in an experiment. Additionally, in comparison to University students, our participants had little if any previous experience with experimental studies. Finally, studies show that bilingualism can also affect performance in lexical decision tasks, as lexical access has been claimed not to be language selective (see Lagrou, Hartsuiker, & Duyck, 2011).

Where item characteristics are concerned, the main difficulty may lie in certain low-frequency lemmas used in our experiment. When combined with particular less frequent form endings (such as the imperfect second person plural form), some potentially very rarely encountered word forms have been included in the experiment. Verb form endings were also not informative to the word/pseudoword distinction — our experiment featured stimuli that all ended in a limited number of verb form endings, which made the experimental session even more demanding than an ordinary lexical decision session. We also note that more errors were registered in responses to pseudowords as opposed to words. Although the trend of lower accuracy to pseudoword than to word stimuli is noted in most auditory lexical decision experiments, the extent of it is an indication of very difficult pseudoword stimuli. Having acceptable verb form endings made our pseudowords morphologically complex and morphologically complex pseudowords may be more difficult to discard than simplex pseudowords (Morris, Porter, Grainger, & Holcomb, 2011). At the same time, the participants completed only a single experimental session which was relatively short in comparison to sessions in large-scale auditory lexical decision studies (Ernestus & Cutler, 2015; Ferrand et al., 2018; Tucker et al., 2019), eliminating any beneficial practice effects.

Besides participant and stimuli characteristics, additional circumstantial factors could have impacted participant performance. The experiment was conducted in a classroom with more noise than the standard sound-attenuated booth and participants, although separated, were tested in groups of up to three. Uncontrolled and unsystematic distractions reduce word recognition accuracy (Lorentzen, Nenadić, Kelley, & Tucker, 2019).

However, the score on the chain test is not one of the factors that explains lower accuracy, as this measure of individual language capabilities did not improve model fit. Higher chain test score was only connected to shorter response latency. Given the form of the task, this finding is not entirely unexpected — the task in the chain test was to separate words as quickly as possible, while the words in the text were commonly known, since they came from sentences at the beginning of a fairy tale. The chain test required speed rather than word knowledge. We take the lack of correlation between the chain test score and accuracy in the auditory lexical decision task as further evidence for differences in (language) skills required for particular language tasks. This is an additional argument in favor of

testing effects in both auditory and visual modalities and with a variety of experimental tasks.

Although we do note a decrease in response speed and accuracy, this does not indicate that the relationship between the predictors and the criterion should change in some manner. In the present study, the standard predictors of trial number, stimulus duration, and word frequency have the expected relationship with accuracy and/or response latency. Other studies with lower participant accuracy due to suboptimal conditions also recorded standard predictor effects. For example, another auditory lexical decision experiment performed in the field showed that response accuracy and speed increase as the participants reach their mid-twenties in terms of age (note that our participants were younger) (Lorentzen et al., 2019). More importantly, as a consequence of surrounding noise and distractions, overall accuracy in this field experiment was 79%, which is more than 8% lower in comparison to the laboratory study complementing it (Tucker et al., 2019). Participant response latency suffered as well, as it increased by more than 200 ms on average. However, the general shapes of predictor effects still matched those from the more controlled, laboratory setting. A similar finding was noted by Mandera, Keuleers, and Brysbaert (2020) in a crowd-sourced online visual lexical decision experiment with 78% overall accuracy in responses to word stimuli. In other words, despite the quantitative drop in accuracy or speed in comparison to laboratory studies, the quality of the process of spoken word recognition should remain the same, therefore maintaining the same relationship between relevant predictors and dependent variables.

The interaction between the inflectional class and part of speech (what we referred to as word Form) indicated that certain combinations were more difficult for the participants. The most curious is the case of the imperfect second person plural in the -i inflection (-i.I2P), especially when response accuracy is considered. Participants responded to these word stimuli almost at a chance level, uncertain whether the stimulus is an actual word of Romanian or not. According to our informants working in the field of language and literature in București, Romania, the imperfect tense is still used in conversational speech, although not too often, and the tense is more prominent in reading and writing. Perhaps our participants are not exposed to many instances of imperfect tense being used in their everyday communication in Romanian. But why doesn't this drop in accuracy happen for other imperfect forms we had in the experiment? The I2P in -a inflection overlaps with the present tense second person plural, making it much easier for the participants. The imperfect first person singular/plural in both inflections (they overlap) could be more present in the media or in the books our participants read, as they are a more probable point of view for the narrator. Therefore, we explain the specific case of I2P in the -i inflection as well as other variation in accuracy and speed as a consequence of frequency of exposure, which could arguably be different between our participants and the data from ROMBAC (Tufiş et al., 2008). Another difficult word Form was the present third person plural (-a.P3P), which arguably should not be too foreign to our participants, but was responded to

relatively slowly. One potential explanation for this finding could be the overlap between -a.P3P and the imperative form, which may have surprised the participants. More encompassing future studies are required to test these hypotheses and provide a detailed account of verb use and processing in Romanian. Particular attention could be given to dialectal differences or differences in effects of word frequency on bilingual speakers.

The effects of relative entropy also match those recorded in previous studies using visual stimuli (e.g., Filipović Đurđević & Gatarić, 2018; Filipović Đurđević & Milin, 2019; Milin et al., 2009). We found that high relative entropy was connected to a higher probability of an incorrect response. If a correct response was made, then the response latency increased when participants responded to words with higher relative entropy. We note that in the previous studies the effects of inflectional entropy became insignificant with inclusion of relative entropy (see Milin et al., 2009), while we find that higher inflectional entropy, alongside higher relative entropy, predicted more errors in responses. In the one study that used verbs as stimuli in the visual lexical decision task (Filipović Đurđević & Gatarić, 2018) the error rate was very small (less than 10%), so statistical analyses focused solely on response latency, and even in that analysis inflectional entropy was not considered. Therefore, the inhibitory effect of inflectional entropy is opposite to the one we would expect from Milin et al. (2009), but it was never before tested in the auditory modality nor in verbs. Another possibility is that the larger number of errors in our study enabled capturing a trend in response accuracy that could not be extracted from the data presented in Milin et al. (2009). Since relative entropy did seem to have a stronger effect than inflectional entropy in our study as well and given that this is the first time this effect has been recorded, we argue that future studies are needed to confirm whether the inhibitory effect of inflectional entropy on response accuracy for (auditory) verb stimuli is stable or not.

Another novel finding in our study is the interaction between word duration and relative entropy, as the effects of relative entropy on response latency seem to be smaller for longer words. Relative to word offset, participants respond to longer words faster (see for example Tucker et al., 2019) as they have more time to accumulate evidence supporting the "word" decision and the number of plausible word competitors diminishes. In such circumstances, perhaps any sort of suffixal morphological information becomes less useful (or impeding), as the participants at least may have had sufficient time to decide on the word lemma, which may not be the case in shorter words. Still, our study is the first to use the auditory lexical decision task, and the interaction between duration and relative entropy was barely significant; additional experiments in other languages are needed to support this finding.

Taken together, the results of our study show that statistical information about form distribution is connected to processing of isolated words in the auditory modality as well as the visual — matching with the prototype facilitates inflected verb processing. The process of lexical access is shaped by the discrepancy in the frequency distribution of the word's paradigm and its class. The effect of relative

entropy is yet another in a long list of various phonological/orthographic, morphological, frequency, and even semantic effects (Goh et al., 2016) that are predictive of human performance in spoken word recognition, present even in the case of isolated word recognition. These results are against a full feed-forward approach to spoken word recognition, where words are recognized solely based on bottom-up acoustic information (see, e.g., Norris & McQueen, 2008; Norris, McQueen, & Cutler, 2000). Although it is conceivable that in perfect circumstances with unambiguous input and a system that makes no errors while perceiving it one would not need anything besides the acoustics, humans do not operate in such conditions in their daily lives. Instead, these results indicate that a listener of a certain language is sensitive to patterns found in that language, learns them, and adapts language processing to best fit previous experience. In other words, the listener musters all available information to aid language processing. The function of this learning is to prepare the listener to decide what is most likely being said in cases of uncertainty, but also to allow a degree of laxness or reduction in communication (see, e.g., Ernestus & Warner, 2011; Kemps, Ernestus, Schreuder, & Baayen, 2004; Lindblom, 1990; Podlubny, Nearey, Kondrak, & Tucker, 2018) which is counterbalanced by predictability.

Specifically, in a highly inflected language, distributional characteristics of word forms are stored and used to guide expectations and therefore facilitate (or, if outside the ordinary, inhibit) processing. Put plainly, morphological information in the form of word paradigms and classes and their interplay are cognitively relevant and it seems that the information-theoretic approach offers a way to define the relationship between patterns found in a language and human language processing (Moscoso del Prado Martín et al., 2004). Importantly, however, how we define (i.e., calculate) this relationship is also a statement of how we assume the process unfolds in human language users. Given that the effect of relative entropy has now been recorded across multiple languages, parts of speech, experimental tasks, and modalities, we argue that it is worthwhile discussing how different operationalizations of this measure of prototypicality may guide our conceptualizations of the mental lexicon and the process of accessing units stored within.

First, we could investigate how word-to-paradigm relative form frequency distributions are learned, and consequently how the information about a particular lemma's relative entropy is updated. In the current way of calulating relative entropy, all lemmas are considered to have equal contribution to their paradigm's relative word form frequency distribution. However, certain lemmas are more frequent than others and listeners encounter them more often, perhaps modulating their relative word form frequency distribution upon every such encounter. We could investigate whether more frequent lemmas also have more of an impact in shaping their paradigm's relative frequency distribution. This can be tested by weighing the contribution of each word lemma based on lemma frequency and testing whether relative entropy calculated in this way predicts response latencies better than the operationalization used in our study and in previous studies. Simply put, we would be answering the

question whether the mental lexicon updates the information on the lemma level (making every lemma an equal contributor to the form frequency distribution in its paradigm) or on the encounter level (making some lemmas more impactful for creating expectations due to their higher frequency of occurrence and therefore relevance).

Second, we also wonder about the form in which relative frequency information is stored (or shaped). Currently, the relative frequency distribution for a particular word's paradigm is compared to a similar distribution calculated based on all the lemmas in its class. However, a single measure of central tendency calculated on a sample of values can often be misleading. We suggest that classes of lemmas should be statistically investigated in search for potential groups of lemmas within the class (or, perhaps, even shared between classes) that have similar "profiles" of word form relative frequency distributions. If such salient groups of lemmas can be formed, we could calculate relative entropy of a particular lemma as discrepancy in comparison to the group it most closely resembles, rather than the class as a whole. This would in turn provide evidence for "natural groups" within grammatical classes, based predominantly on patterns in information gained from language use.

It is only fair to note that our study shows a limited view of the effects of relative entropy as we sample a small number of verb lemmas from only two inflectional classes and exclude lemmas with extremely high relative entropy (due to the small number of data points we had in that region). Further, as recognized by an anonymous reviewer, our study and studies before ours use regular verbs (nouns, adjectives) and do not investigate what impact the presence of irregularity or stem allomorphy may have on inflectional/relative entropy effects. With regard to both of these limitations, we look towards large databases of auditory lexical decision task as a way to test whether findings from smaller targeted experiments such as ours can be generalized to a larger set of words (see, e.g., Keuleers & Balota, 2015). The recently published MEGALEX database could be particularly interesting for relative entropy research since it provides lexical decision responses to multiple word classes in both the auditory and the visual modality for the moderately inflected French language (Ferrand et al., 2018).

Conclusion

We tested whether the effects of relative entropy, a measure of word prototypicality, extend to the Romanian language and to the auditory modality in an auditory lexical decision task. We find effects of relative entropy to both response accuracy and response latency. We argue that the way in which measures of prototypicallity are related to participant processing speed and accuracy can inform our models of the development and organization of information in the mental lexicon.

References

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438.

Baayen, R. H., & Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three germanic languages. *Language*, *81*(3), 666–698.

Bertram, R., Laine, M., Baayen, R. H., Schreuder, R., & Hyönä, J. (2000). Affixal homonymy triggers full-form storage, even with inflected words, even in a morphologically rich language. *Cognition*, *74*(2), B13–B25.

Blevins, J. P. (2003). Stems and paradigms. *Language*, *79*(4), 737–767.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211–252.

Colé, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language*, *28*(1), 1–13.

Ek, U., Fellenius, K., & Jacobson, L. (2003). Reading acquisition, cognitive and visual development, and self-esteem in four children with cerebral visual impairment. *Journal of Visual Impairment & Blindness*, *97*(12), 741-754.

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1469–1488.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*(SI), 253–260.

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., . . . Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition.

Behavior Research Methods, *50*(3), 1285–1307.

Filipović Đurđević, D., & Gatarić, I. (2018). Simultaneous effects of inflectional paradigms and classes in processing of Serbian verbs. *Psihologija*, *51*(3), 259–288.

Filipović Đurđević, D., & Milin, P. (2019). Information and learning in processing adjective inflection. *Cortex*, *116*, 209–227.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M., & Tan, L.-C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorization megastudy. *Frontiers in Psychology*, *7*, 976.

Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes*, *11*(6), 559–568.

Gönczöl, R. (2007). *Romanian: An essential grammar.* London: Routledge.

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, *9*(7), 342–348.

Hendrix, P., Bolger, P., & Baayen, H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(1), 128.

Hockett, C. F. (1954). Two models of grammatical description. *Word*, *10*(2-3), 210–234.

Kemps, R., Ernestus, M., Schreuder, R., & Baayen, H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, *90*(1-3), 117–127.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468.

Kostić, A., Marković, T., & Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (p. 1-44). Walter de Gruyter, Berlin.

Lagrou, E., Hartsuiker, R. J., & Duyck, W. (2011). Knowledge of a second language influences auditory word recognition in the native language. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *37*(4), 952.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Lindgren, S.-A., & Laine, M. (2011). Multilingual dyslexia in university students: Reading and writing patterns in three languages. *Clinical Linguistics & Phonetics*, *25*(9), 753–766.

Lorentzen, P., Nenadić, F., Kelley, M. C., & Tucker, B. V. (2019). Massive auditory lexical decision: Investigating performance in noisy environments. In *Proceedings of the 11th International Conference on the Mental Lexicon* (pp. 1–4). doi: https://doi.org/10.7939/r3-hrnv-sn79

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, M. Joanisse, & KenMcRae (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 76–103). Cambridge University Press Cambridge, UK.

Mandera, P., Keuleers, E., & Brysbaert, M. (2020). Recognition times for 62 thousand english words: Data from the English Crowdsourcing Project. *Behavior Research Methods*, *52*, 741-760.

Milin, P., Filipović Đurđević, 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*(1), 50–64.

Morris, J., Porter, J. H., Grainger, J., & Holcomb, P. J. (2011). Effects of lexical status and morphological complexity in masked priming: An ERP study. *Language and Cognitive Processes*, *26*(4-6), 558–599.

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), 1–18.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech

recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

Podlubny, R. G., Nearey, T. M., Kondrak, G., & Tucker, B. V. (2018). Assessing the importance of several acoustic properties to the perception of spontaneous speech. *The Journal of the Acoustical Society of America*, *143*(4), 2255–2268.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide.* Psychology Software Incorporated.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The massive auditory lexical decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204.

Tufiş, D., Ion, R., Ceauşu, A., & Ştefănescu, D. (2008). RACAI's linguistic web services. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC* (pp. 28–30).

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using gamms.* (R package version 2.3)

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401.

Westbury, C., Hollis, G., & Shaoul, C. (2007). LINGUA: The language-independent neighbourhood generator of the University of Alberta. *The Mental Lexicon*, *2*(2), 271–284.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36.

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*(3), 452–457.

Table 1

Verb classes and verb forms used in the experiment with examples

Inflectional class	Part of speech	Example	Overlap
-a, 201	P1S	abonez	cP1S
	P3P	abonează	P3S, IMP2S
	I1S	abonam	I1P
	I2P	abonați	P2P, cP2P, IMP2P
-i, 401	P1S	amintesc	P3P, cP1S
	P3P	amintesc	P1S, cP1S
	I1S	aminteam	I1P
	I2P	aminteați	none

Notes. Part of speech codes: the first portion refers to present (P), imperfect (I), conjuctive present (cP), or imperative (IMP); the number refers to either first (1), second (2), or third (3) person; the final letter refers to either singular (S) or plural (P).

Table 2

Summary of generalized additive mixed-effects logistic regression predicting participant accuracy

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1.0384	0.3045	3.4097	0.0007
Form-a.P3P	0.5393	0.3713	1.4524	0.1464
Form-a.I2P	0.7495	0.3325	2.2540	0.0242
Form-i.P1S/P3P	0.9405	0.4211	2.2333	0.0255
Form-i.I1S	1.0584	0.3911	2.7063	0.0068
Form-a.P1S	1.0612	0.3125	3.3963	0.0007
Form-a.I1S	1.3082	0.3021	4.3302	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(zTrial,Subject)	48.0588	332.0000	275.3136	< 0.0001
s(Stimulus)	322.5750	579.0000	910.4929	< 0.0001
s(zlogF)	1.0000	1.0000	39.0427	< 0.0001
s(H)	1.0001	1.0001	4.0289	0.0447
s(RE)	2.1273	2.3207	65.5433	< 0.0001
ti(zDuration,RE)	2.4153	2.8262	7.9090	0.0424

Table 3

Summary of generalized additive mixed-effects regression predicting participant response latency

| A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	0.0586	0.0007	85.5302	< 0.0001
Form-a.P3P	0.0003	0.0007	0.4586	0.6466
Form-a.I1S	0.0009	0.0006	1.5917	0.1115
Form-i.I1S	0.0010	0.0007	1.4173	0.1565
Form-a.P1S	0.0013	0.0006	2.0922	0.0365
Form-a.I2P	0.0014	0.0006	2.1535	0.0313
Form-i.P1S/P3P	0.0016	0.0007	2.2415	0.0250
B. smooth terms	edf	Ref.df	F-value	p-value
s(zTrial,Subject)	122.6173	331.0000	3.9330	< 0.0001
s(Stimulus)	262.8445	578.0000	0.8771	< 0.0001
s(zChainT)	2.1100	2.1462	4.4349	0.0136
s(zTrial)	1.0001	1.0001	6.4445	0.0112
s(zDuration)	1.0001	1.0002	154.4913	< 0.0001
s(zlogF)	2.0742	2.3589	8.9445	0.0001
s(RE)	1.2468	1.3578	9.5211	0.0014
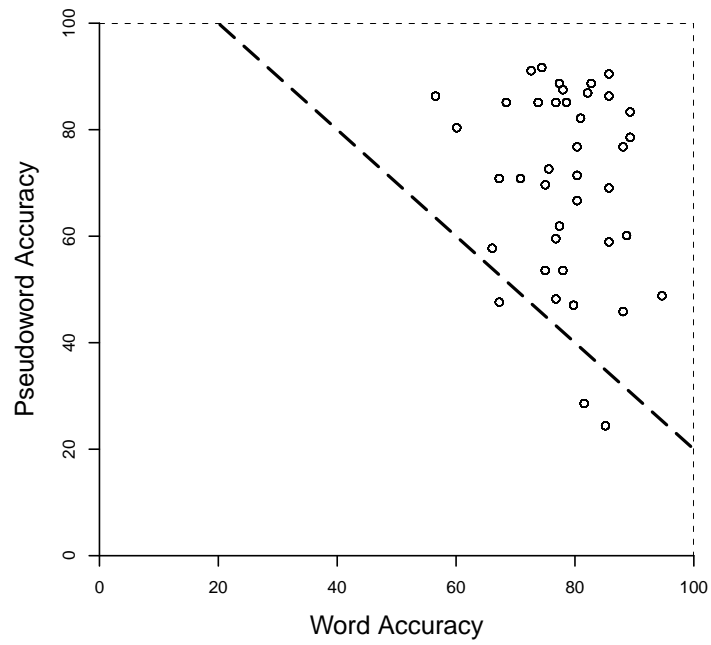
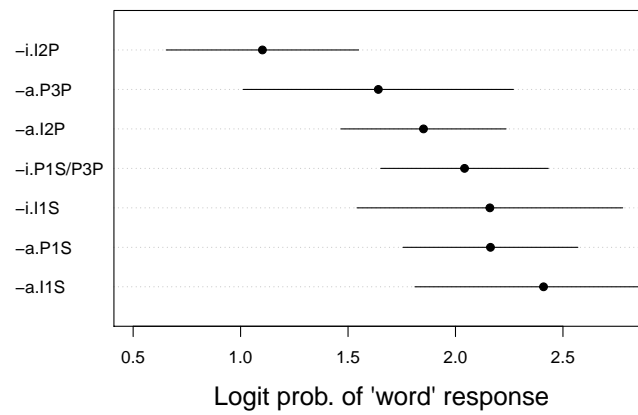*Figure 1*. Participant response accuracy to word (x-axis) and pseudoword (y-axis) stimuli.

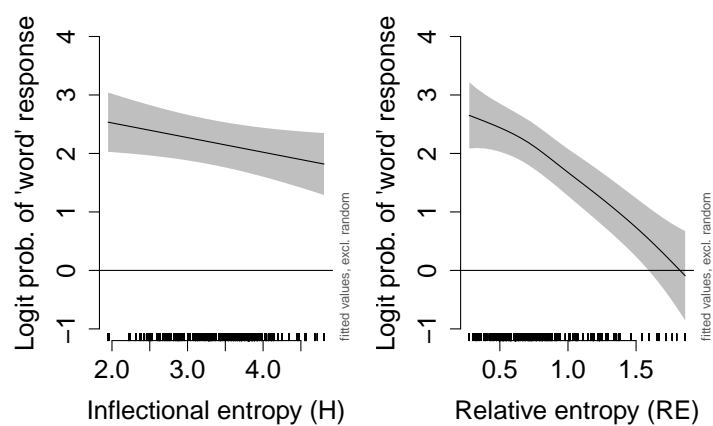*Figure 2*. Effect of word Form in the model predicting response accuracy.

Figure 3. Inflectional and relative entropy show a negative relationship with response accuracy, but the effect of relative entropy seems somewhat larger.
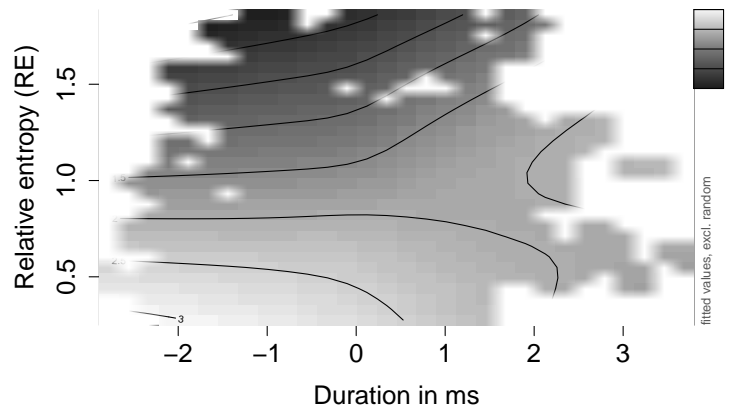
Figure 4. Additive effects interaction of relative entropy and stimulus duration in milliseconds on response probability shows that the effect of relative entropy decreases for longer words.
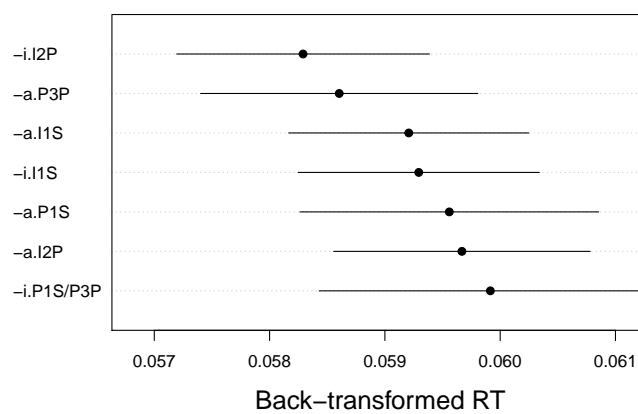

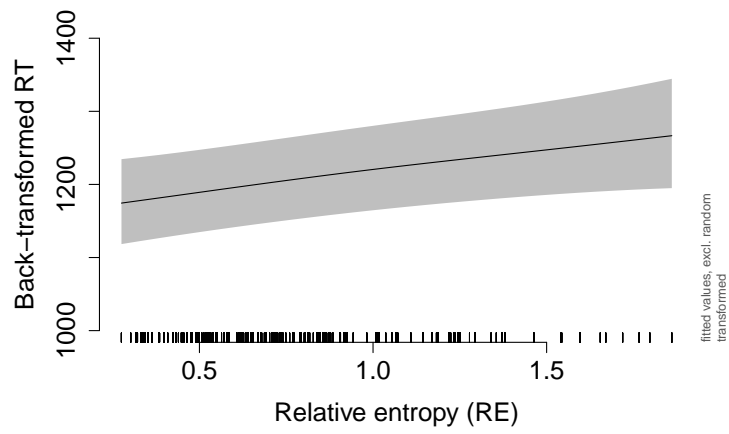*Figure 5*. Effect of word Form in the model predicting response latency.

Figure 6. Higher relative entropy is related to longer response latencies.

Appendix

Parametric effects contrasts

Table A1

Chi-square values for contrasts between levels of word Form in the model predicting response accuracy

	-i.I2P	-a.P3P	-a.I2P	-i.P1S/P3P	-i.I1S	-a.P1S	-a.I1S
-i.I2P	—	2.11	5.08*	4.99*	7.32*	11.53†	18.75†
-a.P3P	2.11	—	0.83	1.55	2.16	4.22*	7.42*
-a.I2P	5.08*	0.83	—	0.35	0.87	1.98	5.49*
-i.P1S/P3P	4.99*	1.55	0.35	—	0.08	0.13	1.05
-i.I1S	7.32*	2.16	0.87	0.08	—	0.00	0.57
-a.P1S	11.53†	4.22*	1.98	0.13	0.00	—	1.29
-a.I1S	18.75†	7.42*	5.49*	1.05	0.57	1.29	—

*Statistically significant contrasts ($p < .05$) based on *wald_gam()* function from the *itsadug* R package.

†Contrasts that remain significantly different ($p < .05$) after applying the Bonferroni correction.

Table A2

Chi-square values for contrasts between levels of word Form in the model predicting response latency

	-i.I2P	-a.P3P	-a.I1S	-i.I1S	-a.P1S	-a.I2P	-i.P1S/P3P
-i.I2P	—	0.21	2.53	2.01	4.38*	4.64*	5.02*
-a.P3P	0.21	—	1.65	1.44	5.98*	9.42†	8.53*
-a.I1S	2.53	1.65	—	0.03	0.95	1.27	1.85
-i.I1S	2.01	1.44	0.03	—	0.27	0.49	1.04
-a.P1S	4.38*	5.98*	0.95	0.27	—	0.10	0.58
-a.I2P	4.64*	9.42†	1.27	0.49	0.10	—	0.31
-i.P1S/P3P	5.02*	8.53*	1.85	1.04	0.58	0.31	—

*Statistically significant contrasts ($p < .05$) based on *wald_gam()* function from the *itsadug* R package.

†Contrasts that remain significantly different ($p < .05$) after applying the Bonferroni correction.