The Role of Letter Frequency on Eye Movements in Sentential Pseudoword Reading

Cengiz Acartürk
(acarturk@metu.edu.tr)
Middle East Technical University
Cognitive Science Program
Ankara, Turkey

Özkan Kılıç
(ozkankilic@ybu.edu.tr)
Yildirim Beyazit University, Computer Engineering
Middle East Technical University, Cognitive Science
Ankara, Turkey

Bilal Kırkıcı
(bkirkici@metu.edu.tr)
Middle East Technical University,
Foreign Language Education
Ankara, Turkey

Burcu Can
(burcucan@cs.hacettepe.edu.tr)
Hacettepe University,
Computer Engineering
Ankara, Turkey

Ayşegül Özkan
(oaysegul@metu.edu.tr)
Middle East Technical University,
Cognitive Science Program
Ankara, Turkey

Abstract

For a language learner, any new word is a pseudoword. A pseudoword is a string of of letters or phonemes that sounds like an existing word in a language, though it has no meaning in the lexicon. On the other hand, speakers are well aware of permissible phonemes, their frequencies and collocations in their language due to the phonotactics inherent in the language. For example, *saktal* is a pseudoword in Turkish, whereas *szyan* is not, due to Turkish phonotactics. This study investigates the relationship between pseudoword letter formation and eye movement characteristics in reading. In particular, we examine the role of Turkish vowel harmony, middle-word consonant collocation, and word-initial and word-final consonants on eye movements with adult native speakers reading sentences that involve predesigned Turkish pseudowords. The results of an experiment with 34 participants are indicative of the role of pseudoword formation on a set of eye movement parameters.

Keywords: Consonant collocations; Eye movements, Pseudowords; Reading; Turkish; Vowel harmony.

Introduction

If native speakers of English are asked to judge the acceptability of the words, *wug*, *toysion*, or *craphen*, they are likely to rate them as acceptable. Similarly, Turkish speakers may judge *talar* as a possible Turkish word though they hear it for the first time. These are all pseudowords, which sound like existing words in a language, without semantic content.

Pseudowords are commonly used as experiment stimuli in research in psychology, linguistics, neuroscience, and cognitive science. They are especially useful when researchers aim at overcoming likely effects of semantics in the experiments. Phonological well-formedness of words (Hammond, 2004), morphological productivity (Anshen & Aronoff, 1988), language development (Dabrowska, 2006), judgment of semantic similarity (MacDonald & Ramscar, 2001), vowel harmony (Pycha, Novak, Shosted & Shin, 2003), machine learning for orthography (Testolin, Stoianov, Sperduti & Zorfib, 2015), neuroimaging of reading (Mechelli, Gorno-Tempini & Price, 2003), and dyslexia (Grainger, Bouttevin, Truc, Bastien & Ziegler, 2003; Houpt, Sussman, Townsend & Newman, 2015) have been among the major topics that have been studied through the use of pseudowords as experimental stimuli. Pseudowords have been employed to test models of word and letter perception, such as the interactive activation model of context effects in letter perception (McClelland, & Rumelhart, 1981).

Previous studies have revealed that native speakers seem to make their judgments by using a probable combination of sounds (Hammond, 2004; Shademan, 2007), the co-occurrence of syllables or consonant collocations locally (Hay, Pierrehumbert & Beckman, 2004), non-locally (Finley, 2012; Frisch & Zawaydeh, 2001; Koo & Callahan; 2011), or through nucleus-coda combination probabilities (Treiman, Kessler, Knewasser, Tincoff & Bowman, 2000). Accordingly, the acceptability judgments of pseudowords are connected to orthographic neighborhood (e.g., Davis, 2010, see Rayner, Pollatsek, Ashby, Clifton Jr., 2012, for a review), which takes into account the relative positions of consonants and vowels (Perea & Lupker, 2004), as well as letter position coding through adjacent bigrams (Seidenberg & McClelland, 1989) and open bigrams of letters (Whitney, 2001).

For Turkish, a language with shallow orthography, it has been shown that some of the properties that characterize pseudowords can be captured from a written corpus in parallel to native speakers' judgment (Kilic, 2014). Against this background, the present study first employs corpus-based frequencies for the formation of pseudowords according to a set of criteria, including the relative positions of consonants and vowels in pseudowords. Then we investigate pseudoword processing by locating the pseudowords in sentential contexts and having native speakers read them in sentential contexts. The underlying approach is that we capture eye movement characteristics as an indicator of word processing, which may enrich our knowledge about word processing by providing data that go beyond response time (cf. lexical decision tasks). In other words, we propose that *sentential pseudoword reading*, as we name it, has the potential to contribute to our understanding of word processing through the study of the relationship between subword characteristics of pseudowords and eye movement parameters.

Three dominant experimental factors that have been found to influence eye movement characteristics in reading are word length, frequency and sentential predictability (Rayner, Sereno & Raney, 1996; Rayner, 1998; Rayner, Pollatsek, Ashby, Clifton Jr., 2012). In the present study, we controlled word length by forming pseudowords of six letters. Using pseudowords allowed us to ignore sentential predictability and word-level frequency as factors for experimental control. Instead, we focused on subword frequencies that are specified by the combinations of the letters that form six-letter pseudowords. We studied three specific types of combinations by designing high-frequency and low-frequency letter bigrams for experimental purposes:

- Vowel combinations that follow vowel harmony
- Middle-word consonant collocation
- Word-beginning and word-end consonants

Silent reading is under the influence of inner speech. If a pseudoword is hard to speak out loud, it is also hard to read silently. In this study, we hypothesize that these three variables will affect pseudoword reading in Turkish as high frequency ensures easy pronunciation, whereas low frequency of these variables lead to difficult pronunciation. It is also expected that adjacent dependencies will have a strong influence on pseudoword reading. In the following section, we present an overview of these characteristics of word structure in Turkish.

The Turkish Language

Turkish is an agglutinating language, read from left-to-right, with a considerably shallow orthography using 8 vowels and 21 consonants derived from the Roman Alphabet (Göksel & Kerslake, 2005; Lewis, 2000). The description of Turkish word structure depends on morphophonological constraints. The continuation of a morpheme and the selection of the corresponding morph are determined by the preceding morph. The final vowel in the preceding morph affects the form of the vowel in the incoming morpheme (Turkish vowel harmony). Similarly, the final consonant in the most recent morph causes some changes on the first consonant of the next morpheme (Assimilation).

While a morpheme with a vowel is concatenated to a string in Turkish, its vowel is modified with respect to the roundedness and backness properties of the most recent vowel in the string as in (1).

(1)	*at-lar*	*kedi-ler*	*okul-lar*
	horse-Plu	cat-Plu	school-Plu
	horses	cats	schools

This is not an immediate dependency because diphthongs and consecutive vowel collocations are usually not allowed in Turkish. When the corpus frequencies are investigated, it is seen that the frequencies of words containing *a…a* or *e…i* as substrings are higher than the frequencies for the words with *ı…ü* or *o…e*. In other words, frequencies mimic vowel harmony. Similarly, some immediate consonant collocations are more frequent than others due to assimilation as shown in (2) for the ablative case marker -DAn.

(2)	*ev-den*	*et-ten*	*yatak-tan*
	house-Abl	meat-Abl	bed-Abl
	from house	from meat	from bed

Accordingly, some consonant collocations, such as *vd*, *tt*, and *kt*, are more frequent than *vt, td*, and *kd*. These are infrequent but not zero frequencies because there are exceptions to assimilation as well as vowel harmony.

Another salient aspect of Turkish word structure is word boundary. Some letters, e.g., *k, g, d, z* and *y*, are observed more frequently than, e.g., *c*, *g*, *r*, *v*, and *f* in word-initial or word-final position. Even Turkish words in root forms mostly follow the regularities briefly exemplified above. In this study, word-initial and word-final boundaries are treated as a single variable (either low or high frequency letters at both boundaries) to keep the stimuli size manageable and the experiment duration reasonable.

The present study examines the effects of the aforementioned three major aspect of word formation in Turkish, namely the frequency of vowel combinations that follow vowel harmony (henceforth, vowel harmony collocation), the frequency of middle-word consonant collocations (henceforth, consonant collocation), and the frequency of word-beginning and word-end consonants (henceforth, word boundary collocation) on eye movement characteristics in sentential reading. These three variables are assumed to be binary variables with either high frequency or low frequency in a 2 x 2 x 2 design. A total of 80 pseudowords were created as experimental stimuli. The letter bigram frequencies were obtained from the METU Turkish Corpus (Say, Zeyrek, Oflazer & Ozge, 2002). All pseudowords had the same template of vowels (V) and consonants (C) shown in (3).

(3) $C_1V_1C_2C_3V_2C_4$

Low-frequency or high-frequency vowel collocations, i.e. V_1V_2, were selected, where high-frequency pairs of V_1V_2 followed the rules of Turkish vowel harmony. Similarly, high-frequency and low-frequency pairs were selected for the consonant collocation C_2C_3, and for the word boundary collocation C_1C_4. Below, we present the experiment, its methods, materials and the results obtained.

Experiment

Thirty-four university students at Middle East Technical University (METU), Turkey, participated in the experiment (18 female and 16 male university students, mean age: 23.59, SD: 5.04). The participants were asked to read silently 80 single-line sentences displayed separately on a computer screen while their eye-movements were recorded by an EyeLink 1000Hz desktop eye tracker. Each sentence included one pseudoword, which was located around the middle of the display. The stimuli were shown in random order. Simple true/false questions were asked randomly after some screens to ensure active engagement of the reader with the task. 20 random and simple true/false questions, which were about words other than the pseudowords in the sentences, were asked as well in order to ensure that the participants had been actively reading. The experiment was conducted in single sessions. Each session took approximately 40 minutes. The participants were paid 25 TL (approximately 6$) as an incentive for participation.

Material

A total of 80 pseudowords of the form $C_1V_1C_2C_3V_2C_4$ were formed for the eight categories (cf. 2 x 2 x 2 design of high- and low-frequency bigrams) by using orthographic frequencies obtained from the METU Turkish Corpus. Frequency was specified as a binary variable, with either high or low values. For example, if the frequency of a consonant bigram is above the mean of the frequencies of all possible bigrams, it was assumed to be a high consonant collocation frequency for selecting C_2C_3. Otherwise, it was assumed to be a low frequency consonant collocation. Similarly, high-frequency and low-frequency bigrams were selected for the word boundary collocations C_1C_4, and for the vowel harmony collocations V_1V_2. We did not distinguish between onset and offset frequencies while studying the bigrams in the corpus to keep the experiment design simple. Below, we present more detail about the three types of collocations.

Turkish Vowel Harmony Collocation (V_1V_2). Since Turkish has eight vowels, 8 x 8 vowel bigrams were produced and their frequencies were calculated from the METU corpus. The vowels might have zero or more intervening characters in between. For example, while *a…a* is a very frequent vowel substring, *ı…ü* is an unlikely one. We calculated all available vowel bigrams. The high-frequency bigrams represented Turkish vowel harmony, whereas the ones with low frequency were due to the words that were exceptions to vowel harmony.

Consonant Collocation (C_2C_3). Turkish has 21 consonants. Therefore, 21 x 21 bigrams from *bb* to *zz* were produced by calculating their bigram frequencies from the corpus. For example, *ml* is much above the average of all possible bigram frequencies while *fv* is very rare in the corpus. Accordingly, high-frequency consonant collocations and low-frequency consonant collocations were identified for forming the pseudoword consonant collocations of the form C_2C_3. In contrast to vowel harmony collocations, only adjacent bigrams were calculated for consonant collocation since C_2C_3 is an adjacent-bigram collocation.

Word Boundary Collocation (C_1C_4). The frequency values of the 21x21 bigrams, from *b…b* to *z…z*, were produced from the 21 consonants in Turkish and their word boundary frequencies were calculated from the corpus. Since C_1C_4 is not an adjacent-bigram collocation, we allowed intervening characters between the first character of words and the last character when calculating the bigram frequencies. For example, *k* is a frequent word-initial boundary while *z* is a frequent word-final boundary in Turkish, making *k...z* a frequent word boundary pattern, whereas *f…b* is a very rare word boundary co-occurrence in Turkish.

Since each of the three independent variables can be either *high* or *low* frequency, eight (2 x 2 x 2) groups, each of which had ten representative pseudowords, were formed. Eight samples from the 80-pseudoword set are shown in Table 1.

Table 1: Pseudoword groups and representative samples

	V_1V_2: Low		**V_1V_2: High**	
	C_2C_3: Low	**C_2C_3: High**	**C_2C_3: Low**	**C_2C_3: High**
C_1C_4: Low	*vöfvac*	*nındüd*	*lagşav*	*remliv*
C_1C_4: High	*töjkır*	*köndız*	*hupşar*	*kadraz*

V_1V_2: Vowel harmony frequency
C_2C_3: Consonant collocation frequency
C_1C_4: Word boundary frequency

The pseudowords were located around the center of 80 meaningful and different sentences of similar length (63 to 65 characters). For avoiding likely effects of the orthographic features of the neighbor words, the pre-target word was fixed in the sentences: The word "aslında" (English 'actually'), which is a very frequent Turkish adverb was consistently used as the single pre-target word. For preventing parafoveal information intake from the post-target word, we used the frequent Turkish bigram *bi-* such that all the post-target words started with this same bigram (cf. Lima & Inhoff, 1985). A sample sentence is shown in (4), where GEN is the genitive marker, PLU is the plural marker, NEG is the negative marker, PASS is the passive marker, and PAST is the past tense marker in Turkish. The target pseudoword is underlined.

(4) Siz-in mektuplar aslında köndız bitmez dendiği için
Your-GEN letter-PLU indeed köndız end-NEG say-PASS for
dağıtılmamış.
distribute-PAST-NEG.

In (4), the target word is the pseudoword *köndız*. It has a low-frequency vowel harmony collocation V_1V_2, in this case *öı*; a high-frequency consonant collocation C_2C_3 *nd*; and a high-frequency boundary collocation C_1C_4 *kz*. (sample sentences are provided in the Appendix).

Results & Discussion

We calculated a set of eye movement measures, as listed below.

- *First fixation duration* is the duration of the first fixation on the first-pass reading of the target pseudoword.
- *First pass gaze duration* is the sum of individual fixation durations in the first-pass reading of the target pseudoword. In other words, the entire word is an area-of-interest and this value represents the total time spent by a participant before his/her gaze left the word for the first time.
- *First pass fixation count* is the sum of individual fixations in the first-pass reading of the target pseudoword. The first-pass reading covers all the fixations on the word without leaving it. If the eye shifts from a pseudoword to another word on its right, the first pass is over. If the eye shifts from a pseudoword to another word on its left, the first pass is over, as well.
- *Regression in count* is the number of regression fixations that return to the target pseudoword after the first-pass reading. If the gaze re-fixates on a previously fixated word, but this time from its right, this is called Regression-in.

The mean values for those eye movement parameters, for the eight pseudoword categories are shown in Table 2.

Table 2: Mean values for eye movement parameters on target pseudowords

	V_1V_2: Low		**V_1V_2: High**	
	C_2C_3: Low	**C_2C_3: High**	**C_2C_3: Low**	**C_2C_3: High**
	First Fixation Duration (ms)			
C_1C_4: Low	265.97	253.67	259.58	253.72
C_1C_4: High	259.74	247.01	274.45	258.06
	First Pass Gaze Duration (ms)			
C_1C_4: Low	467.97	435.13	454.60	390.80
C_1C_4: High	426.58	360.19	431.32	331.75
	First Pass Fixation Count			
C_1C_4: Low	1.86	1.86	1.84	1.66
C_1C_4: High	1.76	1.57	1.71	1.40
	Regression in Count			
C_1C_4: Low	1.27	1.29	1.18	1.22
C_1C_4: High	1.23	1.28	1.28	1.34

We also calculated the averages for eye movement measures for the six-letter, legitimate words of the form CVCCVC in the sentences (excluding the sentence-initial word, the sentence-end words, and post-target words). We found nine words as such, in the stimuli sentence set. Table 3 shows the mean values, which can be taken as a baseline for comparison with the pseudoword values.

Table 3: Mean values for eye movement parameters on legitimate words

First Fixation Duration (ms)	201.65
First Pass Gaze Duration (ms)	237.33
First Pass Fixation Count	1.20
Regression in Count	1.20

The data in Table 3 are not sufficiently representative due to the limited number of words. On the other hand, the values in Table 3 are close to the average values obtained for six-letter words in another study with a richer data set (Acartürk, et al., in preparation). Therefore, we believe that these values can be conceived as close to representative mean values.

Analyses

Four three-way ANOVAs were run on the data from 34 participants to examine the role of vowel harmony collocation, consonant collocation and word boundary collocation on the following eye movement measures in sentence reading: First Fixation Duration, First Pass Dwelling Time, First Pass Fixation Count, and Regression in Count. No significant three-way interaction was observed between the variables and these measures ($F(1,33)=.266$, $p=.61$, $\eta^2=.001$; $F(1,33)=.002$, $p=.97$, $\eta^2=.000$; $F(1,33)=.208$, $p=.65$, $\eta^2=.006$; and $F(1,33)=.000$, $p=.99$, $\eta^2=.000$ respectively). However, there were significant two-way interactions and single effects, as presented below.

First Fixation Duration. Consonant collocation frequency had a significant effect on the first fixation duration, $F(1,33)=8.428$, $p<.05$, $\eta^2=.20$. As the frequency of the collocated consonant bigram decreased, the average duration of the first fixation on the pseudoword significantly increased, and vice versa.

First Pass Gaze Duration. A statistically significant interaction was found between consonant collocation frequency and word boundary collocation frequency, $F(1,33)=4.608$, $p<.05$, $\eta^2=.12$. In other words, the effect of consonant collocation is greater in the high-frequency word-boundary collocation condition than in the low-frequency word-boundary collocation condition, and the effect of word-boundary collocation is greater in the high-frequency consonant collocation condition than in the low-frequency consonant collocation condition.

First Pass Fixation Count. There was a statistically significant interaction between consonant collocation frequency and word boundary collocation frequency, $F(1,33)=4.341$, $p<.05$, $\eta^2=.12$, similar to the finding obtained for the gaze duration. This means that if the frequency of consonant collocation is high, then word boundary collocation frequency has an effect on the first pass fixation count, such that if the word boundary collocation frequency changes from high to low, first pass fixation count increases. If both the consonant collocation frequency and the word boundary collocation frequency change, the effect is stronger.

Regression in Count. A statistically significant interaction was obtained between vowel harmony collocation frequency and word boundary collocation frequency, $F(1,34)=5.002$, $p<.05$, $\eta 2=.13$, without an interaction with consonant collocation frequency. In particular, when the consonant collocation frequency is high, pseudowords with high word boundary collocation frequency gets more re-fixations from its right if the vowel harmony collocation frequency is also high. This effect disappears if the dominant variable, the consonant collocation, becomes infrequent.

Discussion and Conclusion

Our findings revealed mixed results. This indicates the need for further research on pseudoword reading in Turkish. In particular, the findings for the first fixation durations showed that the consonant collocation frequency is the dominant aspect that influences eye movements in reading, since low frequency consonant collocations result in longer fixation durations and vice versa. The frequency of consonant collocations also significantly interacts with the frequency of word boundary collocations, according to the findings obtained for the first pass fixation count and first pass gaze duration, showing that when they both have low frequency, a higher number of first pass fixation count and a longer first pass gaze duration is observed. Finally, the regression-in-count findings suggest that when the consonant collocation frequency is already high, regression-in counts increase when the vowel harmony collocation frequency and the frequency of word boundary collocations are also high. It is likely that this is because pseudowords of this type look like real words. This might have caused the participants to assume misreading a known word, which was followed by re-fixations after the first pass.

In reading research, the first fixation duration on a word is usually conceived as the most valuable indicator of word recognition, since it can be seen as a reflection of the initial-stage processes in word recognition (Rayner, 1998). The second fixation on a word (viz. refixation), further fixations on the same word and regression fixations reveal more complex processes than word recognition, such as syntactic and semantic processes at a sentential-level. Accordingly, at this stage, our findings indicate the frequency of the middle-word consonant allocation as a dominant factor that influences pseudoword reading in Turkish. We also believe that the *sentential pseudoword reading* paradigm has the potential to enrich our knowledge about word processing with higher ecological validity compared to alternative approaches, such as lexical decision tasks and naming tasks.

Acknowledgments

This study has been partially supported by TUBITAK (The Scientific and Technological Research Council of Turkey) 113K723, EU H2020 ISCH COST Action IS1404 Evolution of reading in the age of digitisation (E-READ), and EU FP7 Marie Curie Actions IRIS, 610986, FP7-PEOPLE-2013-IAPP. The authors thank Zuhal Ormanoğlu and Irmak Olcaysoy Ökten for their contribution.

References

Acartürk, C., Kırkıcı, B., Beken, F. (in preparation). Eye movement characteristics in Turkish reading: A corpus-analytic study.

Anshen, F., & Aronoff, M. (1988). Producing morphologically complex words. *Linguistics*, 26, 641–655.

Dabrowska, E. (2006). Low-level schemas or general rules? The role of diminutives in the acquisition of Polish case inflections. *Language Sciences*, 28, 120–135.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*(3), 713.

Finley, S. (2012). Testing the limits of long-distance learning: learning beyond a three-segment window. *Cognitive Science*, *36*, 740–756.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language, 77*, 91–106.

Göksel, A., & Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. Routledge: London and New York.

Grainger, J., Bouttevin, S., Truc, C., Bastien, M., & Ziegler, J. (2003), Word superiority, pseudoword superiority, and learning to read: A comparison of dyslexic and normal readers. *Brain and Language, 87*, 432–440.

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies*, 4, 1–24.

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In: J. Local, R. Ogden, and R. Temple (Eds.), *Phonetic Interpretation: Papersbin Laboratory Phonology VI*. Cambridge: Cambridge University Press.

Houpt, J. W., Sussman, B. L., Townsend, J. T., & Newman, S. D. (2015). Dyslexia and configural perception of character sequences. *Frontiers in Psychology, 6*, 482.

Kilic, O. (2014). Using corpus statistics to evaluate nonce words. In M. Colinet, S. Katrenko, R. K. Rendsvig (Ed.s), *Pristine Perspectives on Logic, Language, and Computation: ESSLLI 2012 and ESSLLI 2013 Student Sessions, Selected Papers* (pp. 26-35). Berlin, Heidelberg: Springer Berlin Heidelberg.

Koo, H., & Callahan, L. (2011). Tier-adjacency is not a necessary condition for learning phonotactic dependencies. *Language and Cognitive Processes, 77*, 1–8.

Lewis, G. (2000). *Turkish Grammar, Second edition*. Oxford: University Press (2000)

Lima, S. D., & Inhoff, A. W. (1985). Lexical access during eye fixations in reading: effects of word-initial letter sequence. *Journal of Experimental Psychology: Human Perception and Performance, 11*(3), 272.

MacDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. Proceedings of the 23rd Annual Conference of the Cognitive Science Society (611-616), University of Edinburgh.

Machelli, A., Gorno-Tempini, M. L., & Price, C. J. (2003). Neuroimaging studies of word and pseudoword reading: Consistencies, inconsistencies, and limitations. *Journal of Cognitive Neuroscience, 15*(2), 260-271.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375.

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory and Language, 51*(2), 231-246.

Pycha, A., Novak, P., Shosted, R., & Shin, E. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In G. Garding and M. Tsujimura (Eds.), *Proceedings of the WCCFL 22*, (423–435).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of reading* (2^{nd} ed.) Psychology Press.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1188–1200.

Say, B., Zeyrek, D., Oflazer, K., & Ozge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. *Proceedings of the Eleventh International Conference of Turkish Linguistics*.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523.

Shademan, S. (2007). From clusters to words: Grammatical models of nonce word acceptability. Grammar and analogy in phonotactic well-formedness Judgments. *PhD Dissertation*, University of California, Los Angeles (2007)

Testolin, A., Stoianov, I., Sperduti, A., & Zorfib, M. (2015). Learning orthographic structure with sequential generative neural networks. *Cognitive Science*, 1-28.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In: M. B. Broe and J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon* (pp. 269–282). Cambridge: Cambridge University Press.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review, 8*(2), 221-243.

Appendix

Below we present sample stimuli for each combination of vowel harmony collocation V_1V_2, middle-word consonant collocation C_2C_3, and boundary word collocation C_1C_4, respectively. H is used for high frequency, and L is used for low frequency. Pseudowords are underlined for demonstration.

V_1V_2 is high, C_2C_3 is high, C_1C_4 is high

- Kaldırmaya çabalasan aslında biylen birden yere düşüp kırılmazdı.
- Mirası harcamasaydım aslında sirden binası şimdi çoktan bitmişti.

V_1V_2 is high, C_2C_3 is high, C_1C_4 is low

- Hasan cebindeki parayı aslında niyled birimi olarak düşünüyordu.
- Son yılın modası aslında lirdev birası ile karides pişirmekmiş.

V_1V_2 is low, C_2C_3 is high, C_1C_4 is high

- Sizin mektuplar aslında köndız bitmez dendiği için dağıtılmamış.
- Dün Emine'nin nişanlısı aslında bındül binayı bulamadı gerçekten.

V_1V_2 is low, C_2C_3 is high, C_1C_4 is low

- Radyonun sesini açarsan aslında nındüd bizden sonuçları öğrenir.
- Şu kediler sokakta aslında lülrıv birini bulup sürtünüyorlarmış.

V_1V_2 is high, C_2C_3 is low, C_1C_4 is high

- Yarın deniz kenarında aslında bıgvıl birini izleyerek eğlenecek.
- Baharda sigara içmek aslında sagşan binası çevresinde yasaktır.

V_1V_2 is high, C_2C_3 is low, C_1C_4 is low

- Bu bölgedeki kuşların aslında lagşav birini korkuttuğu söylenir.
- Fizik dersi haricinde aslında revşev bilimi konusundan bahsetti.

V_1V_2 is low, C_2C_3 is low, C_1C_4 is high

- İstatistik dersinde aslında söcşun birden bütün verileri bozmuş.
- Hastaneye sabah gelenler aslında töjkır birini görmek istiyorlar.

V_1V_2 is low, C_2C_3 is low, C_1C_4 is low

- Doğa resimlerinde aslında löcşuv binası betimlemelerini kullanır.
- Cuma gecesi televizyonda aslında şobçüc birini izleyerek uyudum.