Using Corpus Statistics to Evaluate Nonce Words

Özkan Kılıç

Department of Psychology, Lehigh University, Bethlehem PA, USA
ozkan.kilic@lehigh.edu

Abstract. Nonce words are widely used in linguistic research to evaluate areas such as the acquisition of vowel harmony and consonant voicing, naturalness judgment of loanwords, and children's acquisition of morphemes. Researchers usually create lists of nonce words intuitively by considering the phonotactic features of the target languages. In this study, a corpus of Turkish orthographic representations is used to propose a measure for the nonce word appropriateness for linearly concatenative languages. The conditional probabilities of orthographic co-occurrences and pairwise vowel collocations within the same word boundaries are used to evaluate a list of nonce words in terms of whether they would be rejected, moderately accepted or fully accepted as novel words. A group of 50 Turkish native speakers was asked to judge the same list of nonce words on how native-like the words sound. Both the model and the participants displayed similar results.

Keywords: Nonce words, Orthographic representations, Conditional probabilities.

1 Introduction

Nonce words are frequently employed in linguistic studies to evaluate areas such as well-formedness [1], morphological productivity [2] and development [3], judgment of semantic similarity [4], and vowel harmony [5]. Nonce words are also used to understand the process of adopting loan words. The majority of loaned words undergo certain phonetic changes to more resemble the lexical entries of the language into which they will be adopted [6]. For example, *television* in Turkish becomes *televizyon /televızjon/* because */jon/* is more frequent than */ʒın/* in Turkish[1]. Similarly, *train* is adopted as *tren /tren/* because, similar to diphthongs, vowel-to-vowel co-occurrences are not usually allowed in Turkish non-compound words. This phenomenon shows that the speakers of a language are aware of the possible sound frequencies and collocations of their native languages, and they can make judgements on the naturalness of loan words, recently

[1] In the METU-Turkish Corpus, there are 181 occurrences with the segment */ʒın/* of which only 30 are at the terminating word boundaries. On the other hand, there are 5,945 occurrences with the segment */jon/* of which 3,190 are at the terminating word boundaries, excluding the word *televizyon.*

M. Colinet et al. (Eds.): ESSLLI 2012/2013, LNCS 8607, pp. 26–35, 2014.

invented words and nonce words by using their knowledge of the existing Turkish lexis. Alternatively, it can be claimed that when a loan word does not match statistical properties of a target language, the native speakers of that language either consciously change the word for a better alignment, or the speakers instead perceive the word in accordance with the sound patterns they are used to hearing in their language. It is also reported that known-word statistics is determinant in some linguistic processes [26, 27]; thus, the acceptability of nonce words can be a decision based on these statistics as in the current study.

The acceptability of nonce words can be studied by experimental investigations through phonotactic properties or factor-based analysis [7]. In the experimental investigations, it is observed that the participants accepted or rejected nonce words according to probable combinations of sounds [1, 8]. In factor-based analysis, the acceptability of nonce words is evaluated through the co-occurrences of syllables or consonant clusters locally [9] or non-locally [10–12] or through nucleus-coda combination probabilities [13].

In this study, the acceptability of nonce words was assessed using the conditional probabilities of the bigram co-occurrences of the orthographic representations locally and the pairwise collocations of the vowels within the same word boundaries. Similar models within the context of phonotactic modeling had been used for Finnish vowel harmony [14]. The model for Finnish language uses Boltzmann distribution. Yet the current study much simpler because the local bigram frequencies were used to evaluate Turkish nonce words. Two threshold values were set for the decision to reject, moderately accept and fully accept to judge how the words sound native-like. The threshold values were computed according to the length of each input string. For the evaluation of the conditional and collocation probabilities, the METU-Turkish Corpus containing about two million words was employed [15]. The list of nonce words was created intuitively by randomly combining frequent and infrequent syllables in Turkish. The same list of nonce words evaluated by the model was also given to 50 Turkish native speakers to judge the level of acceptability of each word. The 25 male and 25 female Turkish native speakers, had an average age of 31.26 ($s = 4.11$).The judgements from the native speakers and the model agreed on 82% of the words. In this paper, brief information about Turkish language and plausibility of conditional probabilities will be followed by details of the model and the results.

2 Turkish Language and Conditional Probability

Turkish has 8 vowels and 21 consonants, and it is agglutinative with a considerably complex morphology [16, 17]. While communicating, the word internal structure in Turkish is required to be segmented because Turkish morphosyntax plays a central role in semantic analysis. For example, although Turkish is considered as an SOV language, the sentences are usually in a free order. Thus, the subject and object of a verb can only be determined by the morphological markers as in (1) rather than the word order.

(1)	*Köpek adam-ı ısırdı.*	*Köpeğ-i adam ısırdı.*
	Dog man-Acc bit	Dog-ACC man bit
	The dog bit the man.	The man bit the dog.

The description of Turkish word structure depends heavily on morphophonological constraints and morphotactics. In Turkish morphotactics, the continuation of a morpheme is determined by the preceding morpheme or by the stem as in (2).

(2)	*ev-de-ki*	**ev-ki-de*
	house-Loc-Rel	
	The one in the house	

These morphotactic constraints in Turkish are captured by statistical models based on conditional probabilities [18, 19]. In addition to morphotactics, the morphophonology of Turkish needs a brief explanation because nonce words have to mimic this morphophonology.

Vowel harmony is dominantly effective in Turkish morphophonology in order to preserve the roundedness and the frontness of vowels within the same word boundaries. While a morpheme with a vowel is concatenated to a string, its vowel is modified with respect to the roundedness and frontness properties of the most recent vowel in the string as in (3).

(3)	*ev-ler*	*oda-lar*	*bil-di*	*duy-du*
	house - Plu	room - Plu	know - Past	hear - Past
	houses	rooms	knew	heard

Another important phenomenon in Turkish morphophonology is voicing. If some of the strings terminating with the voiceless consonant, *p, t, k, ç,* are followed by the suffixes starting with vowels, then the consonants are voiced as *b, d, ğ, c* as in (4).

(4)	*sonuç*	*sonuc-um*	*kanat*	*kanad-ı*
	result	result -1S.Poss	wing	wing - Acc
		my result		he wing

Consonant assimilation is also important in Turkish morphophonology. The initial consonants of some morphemes undergo an assimilation operation if they are attached to the strings terminating in the voiceless consonants, *p, t, k, ç, f, s, ş, h, g,*. For example, the surface forms of of the Turkish past tense *-DI* in (5) start with a *-t* because of the terminal sounds *-t* and *-ş* of the roots.

(5)	*at-tı*	*konuş-tu*
	throw - Past	speak - Past
	threw	spoke

The final Turkish morphophonological phenomena that need to be briefly mentioned are deletion and epenthesis. Some of the loanwords as in (6) either lose their final vowel (deletion) or receives an additional copy of their final consonant (epenthesis).

(6) *hak*	*hakk-ım*	*isim*	*ism-im*
right	right - 1S.Poss	name	name - 1S.Poss
	my right		my name

The Turkish morphophonological phenomena described above occur in the co-occurrences of the orthographic representations in the concatenating positions except in vowel harmony and the deletion. This results in high conditional probabilities evaluated using the frequencies of the pairs of immediately consecutive orthographic representations. Since the vowel harmony and deletion take place after or before the concatenation positions, their pairwise collocations within the same word boundaries are also required to be utilized in the statistical model.

The transition probability between A and B is simply based on the conditional probability statistics as in Formula 1.

$$P(B|A) = \textit{(frequency of AB) / (frequency of A)} \tag{1}$$

Infants are reported to successfully discriminate speech segments using transitional probabilities of syllable pairs [20, 21]. Adults also make use of transitional probabilities between word classes to acquire syntactic rules [22, 28]. Similarly, transition probabilities are dominantly used in unsupervised morphological segmentation and disambiguation [18, 19], [23–25].

Statistical approaches to linguistics support the empiricist view, which states that knowledge comes only or primarily from sensory experience instead of being genetically encoded. Such approaches provide an explanatory account of some linguistic phenomena such as the one in the current study. Considering the properties of the Turkish language, using the conditional probabilities of orthographic representations and the collocations of vowels within the same word boundaries is a plausible model to decide whether nonce words or loan words will be *rejected*, *moderately accepted* or *accepted*. In the current study, it is assumed that native speakers judge nonce words mainly based on their morphotactic, morphophonological and phonotactic properties. These properties can be captured by constraints on orthographic collocations by the model explained in the next section.

3 The Model

Let s be a string such that $s = u_1u_2 \ldots u_n$, where u_i is a letter in the Turkish alphabet. The string s is unified with the empty strings σ and ε such that $s = \sigma u_1 u_2 u_n \varepsilon$, where σ denotes the initial word boundary and ε denotes the terminal word boundary. Word boundaries are essential in the judgement process. For example, although the sound $\breve{g}$ is moderately frequent in Turkish, it never occurs as an initial sound but it is rarely the terminal letter. The overall transition probability of the string s is evaluated from the METU-Turkish Corpus using Formula 2, which is actually the product of Formula 1.

$$P_t(s) = \prod_{1}^{n+1} P(u_i|u_{i-1}) \tag{2}$$

For example, using the Formula 2, $P(a|\sigma)$ gives the probability of the strings starting with the letter a, and $P(b|a)$ estimates the probability of the substring ab in the corpus. Now let v be a subset of the string s such that $v = u_{i,1}u_{j,2}\dots u_{k,m}$ where $u_{k,m}$ is the m^{th} vowel in the k^{th} location of the string s. The overall vowel collocations of the string s are estimated from the substring of vowels v using Formula 3.

$$P_c(v) = \prod_{2}^{m} \frac{g(v_{i-1}v_i)}{f(v_{i-1})} \qquad if\ |v| > 1$$

$$P_c(v) = \frac{f(v_i)}{CorpusSize} \qquad if\ |v| = 1 \tag{3}$$

In the Formula 3, the function $f(v_i)$ gives the frequency of the words that contain the vowel v_i as a substring in the corpus. The function $g(v_{i-1}v_i)$ gives the frequency of words in which the vowels v_{i-1} and v_i are collocating not necessarily in immediately consecutive positions but within the same word boundaries. This frequency is divided by $f(v_{i-1})$ because some Turkish words may violate Turkish vowel harmony. The division provides the model with the obedience or violation of the vowel harmony in a probabilistic manner with respect to v_{i-1}. The acceptability probability of the string s is calculated by $P_a(s) = P_t(s)P_c(v)$. The acceptability decision of the string s in the model is made by using the Formula 4.

$$\begin{array}{lll} Accept & if & P_a(s) \geq 10^{-(t+v)} \\ Moderately\ accept & if & 10^{-(t+v+1)} \leq P_a(s) < 10^{-(t+v)} \\ Reject & if & 10^{-(t+v+1)} > P_a(s) \end{array} \tag{4}$$

where t is the number of transitions (which is *the length of the string* + 1) and v is the number of the vowel collocations (which is *the number of the vowels* - 1) in the string. If the string s has only one vowel, then $v = 1$. The threshold values are chosen to best fit the participants responses. Thus, they are changeable values depending on the size size of the corpus.

The model was applied to the list of nonce words given in the following section. The same list was also given to the 50 Turkish native speakers to evaluate the acceptability of each item. The comparison of the results from the model and the native speakers is given below.

Table 1. The results of the model and the results of the participants (Bold text indicates that the model predicted the majority of participants' responses)

		Responses of the Participants		
Nonce Words	Results of the Model	Reject	Moderately Accept	Accept
öğtar	**Reject**	96%	4%	
söykıl	**Reject**	96%	4%	
talar	**Accept**			100%
telüti	**Reject**	64%	28%	8%
prelüs	**Reject**	84%	14%	2%
katutak	**ModeratelyAccept**	8%	50%	42%
par	**Accept**		14%	86%
öçgöş	**Reject**	100%		
jeklürt	**Reject**	100%		
böşems	**Reject**	88%	12%	
trüğat	**Reject**	96%	4%	
cakeyas	**Reject**	92%	8%	
çörottu	**Reject**	74%	16%	10%
döyyal	**Reject**	78%	22%	
efföl	**Reject**	92%	8%	
aznı	Reject	32%	60%	8%
fretanit	**Reject**	64%	30%	6%
erttiçe	**ModeratelyAccept**	36%	64%	
goytar	Reject	38%	52%	10%
hekkürük	Reject	41%	47%	12%
henatiya	**ModeratelyAccept**	36%	64%	
taberarul	**Reject**	84%	16%	
gövük	Reject	30%	44%	26%
sör	**ModeratelyAccept**		78%	22%
perolus	**Reject**	84%	16%	
kletird	**Reject**	98%	2%	
ojuçı	**Reject**	100%		
ürtanig	**Reject**	94%	6%	
lezğaji	**Reject**	100%		
lamafi	**ModeratelyAccept**		64%	36%
nort	Reject	38%	42%	20%
netik	**Accept**		18%	82%
meşipir	ModeratelyAccept		24%	76%
oblan	**ModeratelyAccept**		58%	42%
öftik	**Reject**	62%	34%	4%
özola	**ModeratelyAccept**	32%	60%	8%
ayora	Accept		72%	28%
sengri	**ModeratelyAccept**	32%	68%	
sakkütan	**Reject**	58%	34%	8%
şepilt	**Reject**	78%	22%	
şür	**ModeratelyAccept**		78%	22%
puhaptı	**ModeratelyAccept**	38%	44%	18%
upapık	**Reject**	54%	28%	18%
ülü	Reject	28%	52%	20%
yukta	**ModeratelyAccept**		74%	26%
zerafip	**Reject**	54%	34%	12%
upgur	**Reject**	70%	16%	14%
kujmat	**Reject**	90%	10%	
lertic	**Reject**	94%	6%	
düleri	Accept		64%	36%

4 Results

The nonce word *talar* is evaluated as in (7)

$$\begin{aligned}(7)\qquad P_a(talar) &= P_t(\sigma talar\varepsilon)P_c(aa)\\ &= P(t|\sigma)P(a|t)P(l|a)P(a|l)P(r|a)P(\varepsilon|r)P_c(aa)\\ &= 7.66e-06P_c(aa) = 7.66e-06*4.75e-01 = 3.63e-06\end{aligned}$$

Since $P_a(talar) \geq 10^{-(6+1)}$, in which 6 conditional probability estimations and 1 vowel collocation are evaluated, the nonce word *talar* is accepted.

The word list is evaluated by the 50 Turkish speakers. The participants are composed of 25 males and 25 females with at least undergraduate degrees. They are given the words written on a paper with a 3-level scale (A: Accept, M: Moderately accept, R: Reject), and instructed that these words need to be evaluated by native speakers because the words are going to be used as novel words to name some recently invented colors, objects and actions in Turkey. The distribution of the native speaker responses and the results of the model are given in Table 1.

For 82% of the words the Turkish native speakers' responses are in agreement with the results from the model. The model failed to simulate the responses from the participants in 18% of the results.

The nonce word *ülü* was rejected by the model but accepted by the participants. A possible reason might be that the nonce word *ülü* sounds similar to an existing Turkish word *ölü* 'death'. Similarly, the responses for the nonce word *nort* were in disagreement. This nonce word has a similar pronunciation to an English word *north* and the most of the participants also knew English as a foreign language. Therefore, the participants might also make use of their foreign language knowledge to evaluate nonce words.

5 Discussion and Conclusion

The acceptability of loan words and nonce words is mainly determined by the phonological properties of the target language and the current approaches are syllable-based [7–13]. Since there are no lexical entries for nonce words, the model in this study tries to estimate the acceptability of the words using the bigram conditional probabilities and collocations of the orthographic representations within the word boundaries, which is a simplified way of inducing Turkish morphophonology.

Although the model does not assume to utilize any property of Turkish phonology and it does not implement any phonologic filtering mechanism, it is able to mimic, in a remarkable way, a large number of the responses from the participants. Indeed, this study does not propose that acceptability is based on raw orthographic representations rather than syllables and phonemes. Instead, it underlines that simple pairwise conditional properties and vowel collocations from a corpus can give an estimation of the acceptability of a list of Turkish nonce words. This can be used by researchers that need an evaluation for the nonce words for their studies when no phonologically annotated corpus with syllables exists.

The argument in this study could be extended to grammaticality judgement in a way that the speaker does not need to store explicit rules about which rules are grammatical. Sensitivity to statistical properties of observed combinations can be enough to account for the speaker's grammaticality judgement behaviour. Yet, in this case, it is necessary to include additional steps in the model to represent how a speaker *smooths* a novel grammatical construction with zero probability by using the frequency information from known constructions in order to reject, moderately accept, or accept the novel construction.

6 Limitations and Future Work

The model needs to be tested with larger word lists to improve the results. The model is successful because there is a close correspondence between phonotactic and orthotactic in Turkish. If one wants to test the model in different languages, it requires improvements in terms of the morphophonological properties of the target languages. The model uses exact orthographic representations. Thus, it requires an additional phonological similarity measure for the representations to increase the success rate because it seems that the native speakers also make use of phonologic similarities among sounds, such as accepting the nonce word *ülü* since *ü resembles ö* in the real word *ölü* in terms of roundedness and backness.

The threshold values for the acceptability decisions depend on word lengths. They also need to be improved with respect to the target languages. The model also needs to be tested in and adapted for different languages with ablaut or umlaut phenomena such as English and German, and the templatic languages such as Arabic and Hebrew, because they are not linearly concatenative and immediate sound co-occurrences are not powerful enough to capture their morphophonological properties.

References

[1] Hammond, M.: Gradience, phonotactics, and the lexicon in English phonology. Int. J. of English Studies 4, 1–24 (2004)

[2] Anshen, F., Aronoff, M.: Producing morphologically complex words. Linguistics 26, 641–655 (1988)

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

[4] MacDonald, S., Ramscar, M.: Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In: Proc. of the 23rd Annual Conference of the Cognitive Science Society, University of Edinburgh (2001)

[5] Pycha, A., Novak, P., Shosted, R., Shin, E.: Phonological rule-learning and its implications for a theory of vowel harmony. In: Garding, G., Tsujimura, M. (eds.) Proc. of WCCFL, vol. 22, pp. 423–435 (2003)

[6] Kawahara, S.: OCP is active in loanwords and nonce words: Evidence from naturalness judgment studies. Lingua (to appear)

[7] Albright, A.: From clusters to words: Grammatical models of nonce word acceptability. Handout of talk presented at 82nd LSA, Chicago (January 3, 2008)

[8] Shademan, S.: From clusters to words: Grammatical models of nonce word acceptability. Grammar and Analogy in Phonotactic Well-formedness Judgments. Ph. D. thesis, University of California, Los Angeles (2007)

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

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

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

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

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

[14] Goldsmith, J., Riggle, J.: Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. Natural Language & Linguistic Theory (to appear)

[15] Say, B., Zeyrek, D., Oflazer, K., Özge, U.: Development of a corpus and a treebank for present-day written Turkish. In: Proc. of the Eleventh International Conference of Turkish Linguistics (2002)

[16] Göksel, A., Kerslake, C.: Turkish: A Comprehensive Grammar. Routledge, London (2005)

[17] Lewis, G.: Turkish Grammar, 2nd edn. University Press, Oxford (2000)

[18] Kılıç, Ö., Bozşahin, C.: Semi-supervised morpheme segmentation without morphological analysis. In: Pro. of the LREC 2012 Workshop on Language Resources and Technologies for Turkic Languages, Istanbul, Turkey (2012)

[19] Yatbaz, M.A., Yuret, D.: Unsupervised morphological disambiguation using statistical language models. In: Pro. of the NIPS 2009 Workshop on Grammar Induction, Representation of Language and Language Learning, Whistler, Canada (2009)

[20] Aslin, R.N., Saffran, J.R., Newport, E.L.: Computation of conditional probability statistics by human infants. Psychological Science 9, 321–324 (1998)

[21] Gomez, R.L.: Variability and detection of invariant structure. Psychological Science 13, 431–436 (2002)

[22] Kaschak, M.P., Saffran, J.R.: Idiomatic syntactic constructions and language learning. Cognitive Science 30, 43–63 (2006)

[23] Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Tran. on Speech and Language Processing 4(1) (2007)

[24] Bernhard, D.: Unsupervised morphological segmentation based on segment predictability and word segments alignment. In: Proc. of 2nd Pascal Challenges Workshop, pp. 19–24 (2006)

[25] Demberg, V.: A language-independent unsupervised model for morphological segmentation. Ann. Meet. of Assoc. for Computational Linguistics 45(1), 920–927 (2007)

[26] Debrowska, E.: The effects of frequency and neighbourhood density on adult native spakers' productivity with Polish case inflections: An empirical test of usafe-based approaches to morphology. Memory and Language 58, 931–951 (2008)

[27] Baayen, R.H., Dijkstra, T., Schreuder, R.: Singulars and plurals in Dutch: Evidence for a parallel dual route model. Memory and Language 37, 94–117 (1997)

[28] Reeder, P.A., Newport, E.L., Aslin, R.N.: From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. Cognitive Psychology 66, 30–54 (2013)