Turkish Resources for Visual Word Recognition

Begüm Erten, Cem Bozsahin, Deniz Zeyrek

Cognitive Science Department
Informatics Institute, Middle East Technical University (METU), Ankara Turkey
B.Erten.Uyumaz@tue.nl, {bozsahin,dezeyrek}@metu.edu.tr

Abstract

We report two tools to conduct psycholinguistic experiments on Turkish words. `KelimetriK` allows experimenters to choose words based on desired orthographic scores of word frequency, bigram and trigram frequency, ON, OLD20, ATL and subset/superset similarity. Turkish version of `Wuggy` generates pseudowords from one or more template words using an efficient method. The syllabified version of the words are used as the input, which are decomposed into their sub-syllabic components. The bigram frequency chains are constructed by the entire words' onset, nucleus and coda patterns. Lexical statistics of stems and their syllabification are compiled by us from BOUN corpus of 490 million words. Use of these tools in some experiments is shown.

Keywords: morphology, lexical statistics, pseudoword generation

1. Introduction

In this paper we report two Turkish language resources (r1/r2) and two tools (t1/t2), namely a stem list of publicly available BOUN corpus of (Sak et al., 2011), with its stem frequencies (r1), which is approximately 490 million words, its syllable patterns (r2), and two tools to experiment with them. The statistics are assessed for the use of the data and the tools in psycholinguistic tasks, namely pseudoword generation and lexical decision tasks. We provide as a public resource our own KelimetriK (t1) software development and a Wuggy project add-on (t2). Wuggy is a commonly used tool in psycholinguistic experiments. To the best of our knowledge, the Turkish resources and the tools we provide for these purposes are first of their kind in terms of size and functionality.

2. Background

Visual word recognition tasks are ideal experimental paradigms for investigating mental lexical processing which begins with the recognition of graphemes, then proceeds to recognition of the word, and finally accessing its semantic information (Mainy et al., 2007). Overall, lexical processing consist of four dimensions: phonological, lexical, grammatical and semantic (Cibelli, 2012).

Pseudowords, or word-like non-words, are useful tools for manipulating the dimensions of lexical processing because they look like real words except lacking a semantic representation. They are used in visual word recognition tasks where the subjects are expected to respond to visually presented verbal items in a short time.

The lexical decision task (LDT) is a visual word recognition task, and it provides a good comparison among words and pseudowords in terms of visual lexical processing. As its name suggests, subjects are expected to decide whether a presented verbal item is a real word or a pseudoword.

The verbal items in a visual word recognition task should be carefully designed by the experimenters because they contain phonological and orthographic features which might influence mental lexical processing. Phonologically, a word's graphemes have variations like consonants and vowels which have an influence on the mental processing (Frost, 1998). Orthographically a word is a conjunction of several graphemes based on some rules. They can be assessed on different aspects, e.g. length and frequency, as described below.

The length of a word as an approximation is the total number of graphemes in a word. Word frequency scores are obtained by counting how many times a word occurs in a text. Bigram and trigram frequencies of a word are obtained by counting the words in a lexical list that have some length of adjacent strings. For example, the word *duck* consist of three bigrams (du, uc, ck) and two trigrams (duc, uck).

If the location of the bigram in a word is to be considered in counting, it is a location-dependent frequency measure, otherwise the frequency measure is location independent. Orthographic Neighborhood size (ON) is obtained by counting the number of words in a lexical database into which a word can be transformed when a single grapheme is substituted. For example the word *song* has six orthographic neighbors including *sing* and *gong*, in the CELEX database (Duyck et al., 2004).

Orthographic Levenshtein distance 20 (OLD20) is obtained by taking the average of the 20 closest words in the unit of Levenshtein distance (LD) (Yarkoni et al., 2008). LD is a continuous metric that indicates the minimum number of insertions, deletions or substitutions required for turning one string to another (Levenshtein, 1966). Orthographic relatedness is a special form of similarity of two words, such as adjacently transposed letter (or grapheme) similarity (ATL), and subset/superset similarity. The former is the case when two words can be transformed to each other by transposing, such as *dart* and *drat* (Davis, 2005). Subset/superset similarity is manifested in the words *butter* and *utter*. (A word is best assessed on phonological grounds as the phonological word, but, due to lack of speech recordings that can give us reliable statistics, we use the orthographic representation as the proxy input to our statistics, with natural consequences on real precision. The word error rate in state of the art speech recognition models, even in those which are coupled with a language model, is about 23% for Turkish, which is too high; see (Dikici et al., 2013).

One good news for Turkish in this endeavour is its fairly transparent orthography and almost exclusively agglutinating morphology, by which we get allomorphy to be read off transparently from orthographic representations.)
Behavioral studies show that frequency of a word effects the behavioral data very strongly (Forster and Chambers, 1973). Generally, the reaction time for high-frequency words are significantly lower than low-frequency words because high-frequency words are processed faster than low-frequency words (Carroll and White, 1973). The behavioral effect of ON is not so consistent as that of the frequency variable; it is mostly a facilitation for words having higher ON scores. Therefore, the higher ON a word has, the shorter its response time is (Andrews, 1997). The effect of OLD20 on the behavioral data has a consistent pattern. Words having lower OLD20 values are recognized (and responded) faster than words having high OLD20 values (Yarkoni et al., 2008). For example, the table below shows OLD20 scores for some English words.

CONDITION		PISTACHIO	
Neighbor Words	Levenshtein Distance	Neighbor Words	Levenshtein Distance
conditions	1	distraction	4
coalition	2	hibachi	4
cognition	2	mustache	4
conditional	2	mustached	4
conditioned	2	mustaches	4
conditioner	2	pigtail	4
conduction	2	pistil	4
contrition	2	pitch	4
conviction	2	pitched	4
recondition	2	pitcher	4
rendition	2	pitches	4
addition	3	pitching	4
audition	3	psychic	4
collation	3	psycho	4
collision	3	abstain	5
commotion	3	abstraction	5
conception	3	antacid	5
concoction	3	attach	5
concretion	3	attache	5
conditioners	3	attached	5
OLD20: 2.4		OLD20: 4.3	

Table 1: OLD20 values for the words "condition" and "pistacho" and their 20 LD closest words (Adapted from (Yarkoni et al., 2008))

3. Data and its use

We make use of a stem list compiled by us from BOUN corpus. The corpus has 490 million words, which is the largest Turkish database, collected from the Web by (Sak et al., 2011). They make use of 55,000 lexical items for the purpose of wide-coverage morphological processing with the same corpus, out of which we extracted approximately 32,000 stems excluding proper names and words that have initial onset CC clusters (e.g. *tren* 'train', *kral* 'king'), which are mostly borrowed. The rationale for this choice is to reduce the possibility of overfitting in pseudoword generation and in selecting orthographic neighbors. This stem database covers 346 million words out of 490 million in BOUN. The final count that is input to our tools is 24,414 stems, with full coverage of their syllabifications.
We provide the syllabification database and stem frequency database as public resources. The most frequent 3 stems have approximately 8.6 million occurrences each, the next frequent 3 about 4 million each, out of approximately 346 million token occurrences. Around 2,000 stems have unique occurrence, and 1,300 have only two occurrences. This is a Zipfian distribution, as expected.
Using a very large corpus for descriptive statistics provides us a reliable base on which we can conduct psycholinguistic experiments. A lexical decision task is conducted on a group of subjects, and several hypothesis-testing models were derived from the results. The stimuli set included 250 words and 250 pseudowords. The words in the stimuli set varied in the orthographic scores of word frequency, ON, and OLD20. KelimetriK is used for the selection of words with specified orthographic scores. Wuggy with the Turkish plug-in is used for generating the pseudowords. It was hypothesized that word frequency, ON and OLD20 scores have a specific influence on the behavioral results. Moreover, based on previous studies, the OLD20 variable is expected to have a stronger behavioral effect than the ON variable (Yarkoni et al., 2008). We report the results of these experiments.

4. KelimetriK Tool

KelimetriK is a query based program designed by the first author to provide output to the orthographic scores of word frequency, bigram and trigram frequency, ON, OLD20, ATL and subset/superset similarity. KelimetriK is a useful guide for psycholinguistic experimenters with which we can query the orthographic scores while selecting words for the stimuli sets. These kinds of query software are available in other languages such as N-watch for English (Davis, 2005), and BuscaPalabras for Spanish (Davis and Perea, 2005).
KelimetriK bases its orthographic calculations on a stem list. This list also contains the words' frequency scores which were counted from the BOUN corpus. Because the orthographic calculations are based on the stem list, querying the words with the same orthographic representation but with a different meaning is limited in KelimetriK. It is also not possible to query more than one word at the same time.

5. Wuggy Pseudoword Generator with Turkish Plug-in

The original Wuggy software was developed by (Keuleers and Brysbaert, 2010). The software has a flexible algorithm, which covers various languages. The Wuggy algorithm generates pseudowords from one or more template words using an efficient method. The syllabified version of the words are used as the input, which are decomposed into their sub-syllabic components. The bigram frequency

chains are constructed by the entire words' onset, nucleus and coda patterns.

The Turkish plug-in has been incorporated into the software by us. Three pieces of information are provided: a lexical list as a corpus with word frequency values, a syllabified version of the words in the list, and a regular expression that describes the sub-syllabic pattern of the words (the onset-nucleus-coda pattern). The lexical list for Turkish words is taken from a Turkish stem list which is a 24,414-words database that provides information about the words' frequency values as mentioned above. The words in the Turkish stem list is syllabified using the hyphenation algorithm which was designed for the TEX type setting system (MacKay, 1988). The sub-syllabic pattern for Turkish words is defined in the system, for example the nucleus of the Turkish words can only be the vowels *a, e, ı, i, o, ö, u, ü*.

6. Methods of the Behavioral Study

In this section we show the usefulness of our resources for psycholinguistic experimentation. The stimuli set composed of 250 five-letter Turkish words selected randomly from the Turkish stem list. They varied in the dimensions of word frequency, ON, OLD20 and imageability, the last of which we did not control yet, which is upcoming work. The pseudowords were generated using Wuggy with Turkish plug-in. Each word in the stimuli set is used as a template for generating the pseudowords. The chosen pseudowords have the smallest sub-syllabic element transition frequency among the other nine candidate pseudowords.

The behavioral study was conducted on 37 subjects (21 male and 16 females with mean age of 27.18). Each item in the stimuli set was presented randomly only one at a time as a trial. The subjects were expected to report whether a presented stimulus was a word or a pseudoword as quickly and accurately as possible. The maximum response time of a trial was 2000 milliseconds. The inter-stimulus interval between the trials was 500 milliseconds. Software of the task was specially designed for the experiment on Java programming language using the PsychWithJava library (`http://hboyaci.bilkent.edu.tr/PsychWithJava/`).

7. Results of the Behavioral Study

Response times (RT) were derived from the behavioral data. Most of the words in the stimuli set have ON scores in the range 0-5 (85% of the words), and OLD20 scores in the range 2.0-2.5. The RT difference among words and the pseudowords is (significantly) 58.83 milliseconds [t(36)=-7.756, p<0.05].

In order to investigate the effect of word frequency, ON, and OLD20, the words were grouped according to the score levels per variable. Three different conditions were obtained for statistical comparisons: high versus low frequency, ON, and OLD20. We report them in this order.

Mean frequency scores in the *Low-frequency* condition is 3.654 per-million (sdv= 2.253); it is 460.542 per-million (sdv=460.542) in the *High-frequency* condition. Mean ON scores in the *Low-ON* condition is 0.96 (sdv= 0.79). It is 5.13 (sdv=2.82) in the *High-ON* condition. Mean OLD20 scores in the *Low-OLD20* condition is 1.63 (sdv= 0.17). It is 1.92 (sdv=0.13) in the *High-OLD20* condition. Three separate paired samples t-test analyses are conducted to test the significance of the difference among low versus high-frequency words.

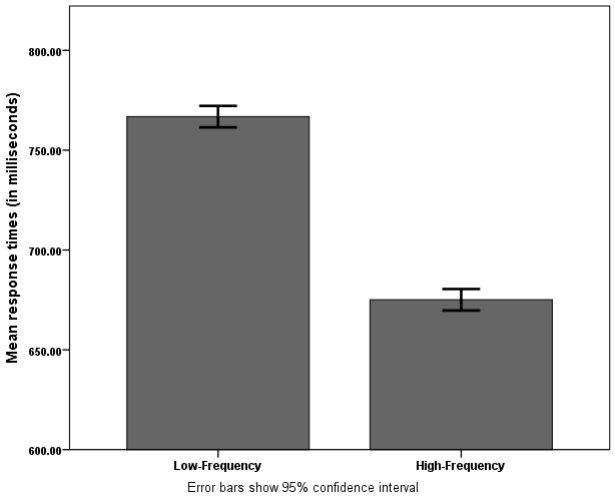

Figure 1: Mean response times of words having low and high word frequency scores.

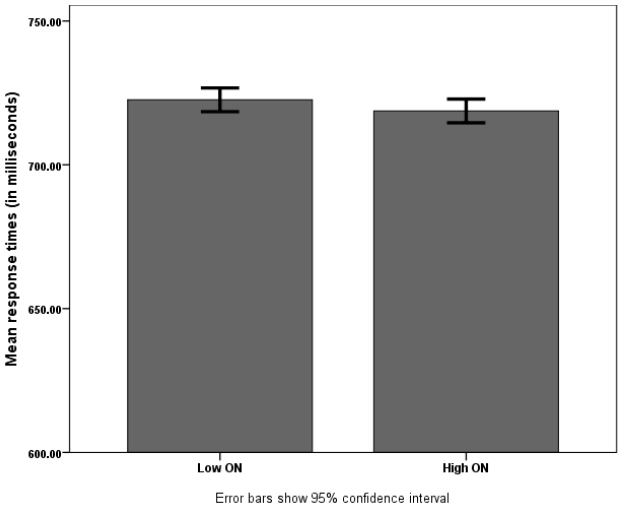

Figure 2: Mean response times of words having low and high ON scores.

Figure 1 shows the mean response times of words in low and high frequency condition. Words with high-frequency scores are significantly lower in mean RT than the words with low-frequency scores [t(36)= 17.295, p<0.05]. Figure 2 shows the mean response times of words in low/high and ON condition. There is no significant difference among low and high ON condition [t(36)=0.948, p<0.05]. Figure 3 shows the mean response times of words in low and high OLD20 condition. Words with high OLD20 scores are significantly lower than the words with low-frequency scores [t(36)=2.127, p<0.05].

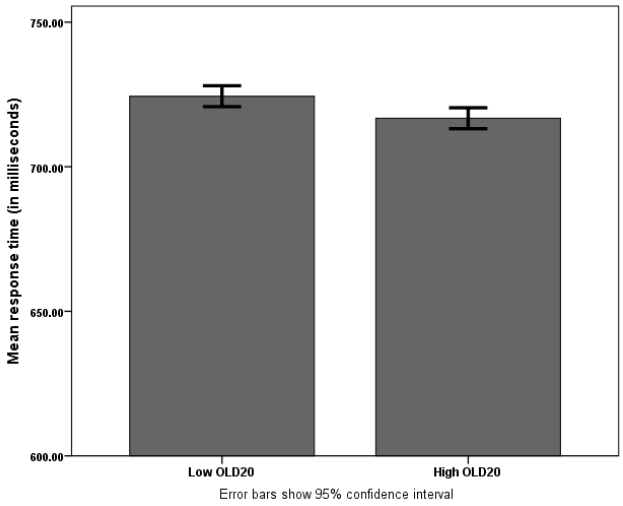

Figure 3: Mean response times of words having low and high OLD20 scores.

8. Conclusion

We have extracted a Turkish stem list with wide-coverage, from a widely used dataset (BOUN), with particular emphasis on syllabification to avoid overfitting in pseudoword generation and to avoid reporting lower precision in orthographic neighbors. The highest ranking stem has approximately 9 million occurrences in a corpus of 346 million words, and lowest rank (one) about 2,000 occurrences. 24,414 stems out of 55,000 were selected for inclusion in visual word recognition studies, which require a stem list, frequency and syllabification. We showed the usefulness of large resources for such tasks.

The most effective variable on the Turkish lexical decision data is word frequency because it had the highest reaction time difference among low and high conditions. There were 92 milliseconds of lexical decision latency difference among low and high frequency conditions. This result is consistent with previous studies conducted on other languages. For example, in an English lexical decision study, low-frequency scores had 196 milliseconds higher reaction time scores for low-frequency words than the high-frequency ones (Forster and Chambers, 1973).

OLD20 is also an effective variable on the Turkish lexical decision data. The effect obtained with our data had the reverse pattern compared with the other lexical decision studies. Reaction times for the high OLD20 words were shorter than the low OLD20 words, and this is an opposite effect when compared with a previous English lexical decision study (Yarkoni et al., 2008). Thus, OLD20 had an inhibition effect on the Turkish lexical decision data rather than a facilitation effect. This may have to do with the language's agglutinating morphology, where unique non-stem combinations (unique suffix groups) can be quite large, as much as 50,000 according to (Sak et al., 2011). We hope to provide a better understanding of this result soon. A future study should replicate the task with the words that have equal frequency scores. The effect of ON on the Turkish lexical decision data is also assessed in this study. We have found it to be ineffective.

The present study confirmed that the effect of the linguistic variables of word frequency and OLD20 is also present in the Turkish lexical decision data when compared with the other languages' lexical decision studies. We need large resources to carry out such experiments, and two such resources and tools are provided in this work. Our tools and datasets are publicly available at github.com/beguyumaz/turkish-visual-word-rec-tools.

9. Acknowledgements

We are grateful to Nihan Ketrez who first suggested that we design the Turkish part of Wuggy, which started this project.

10. References

Andrews, Sally. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, 4(4):439–461.

Carroll, John B and White, Margaret N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1):85–95.

Cibelli, Emily. (2012). Shared early pathways of word and pseudoword processing: Evidence from high-density electrocorticography. UC Berkeley Phonology Lab Annual Report.

Davis, Colin J and Perea, Manuel. (2005). Buscapalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in spanish. *Behavior Research Methods*, 37(4):665–671.

Davis, Colin J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37(1):65–70.

Dikici, Erinç, Semerci, Murat, Saraçlar, Murat, and Alpaydın, Ethem. (2013). Classification and ranking approaches to discriminative language modeling for ASR. *IEEE Trans. on Audio, Speech and Language Processing*, 21(2):292–300.

Duyck, Wouter, Desmet, Timothy, Verbeke, Lieven PC, and Brysbaert, Marc. (2004). Wordgen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods*, 36(3):488–499.

Forster, Kenneth I and Chambers, Susan M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635.

Frost, Ram. (1998). Toward a strong phonological theory of visual word recognition: true issues and false trails. *Psychological bulletin*, 123(1):71.

Keuleers, Emmanuel and Brysbaert, Marc. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3):627–633.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

MacKay, Pierre A. (1988). Turkish hyphenations for tex. *TUGboat*, 9(1):12–14.

Mainy, Nelly, Jung, Julien, Baciu, Monica, Kahane, Philippe, Schoendorff, Benjamin, Minotti, Lorella, Hoffmann, Dominique, Bertrand, Olivier, and Lachaux, Jean-Philippe. (2007). Cortical dynamics of word recognition. *Human brain mapping*, 29(11):1215–1230.

Sak, Haşim, Güngör, Tunga, and Saraçlar, Murat. (2011). Resources for Turkish morphological processing. *Language resources and evaluation*, 45(2):249–261.

Yarkoni, Tal, Balota, David, and Yap, Melvin. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.