AnlamVer: Semantic Model Evaluation Dataset for Turkish - Word Similarity and Relatedness

Gökhan Ercan
Department of Computer Engineering
Işık University, İstanbul, Turkey
gokhan.ercan@isik.edu.tr

Olcay Taner Yıldız
Department of Computer Engineering
Işık University, İstanbul, Turkey
olcaytaner@isikun.edu.tr

Abstract

In this paper, we present AnlamVer, which is a semantic model evaluation dataset for Turkish designed to evaluate word similarity and word relatedness tasks while discriminating those two relations from each other. Our dataset consists of 500 word-pairs annotated by 12 human subjects, and each pair has two distinct scores for similarity and relatedness. Word-pairs are selected to enable the evaluation of distributional semantic models by multiple attributes of words and word-pair relations such as frequency, morphology, concreteness and relation types (e.g., synonymy, antonymy). Our aim is to provide insights to semantic model researchers by evaluating models in multiple attributes. We balance dataset word-pairs by their frequencies to evaluate the robustness of semantic models concerning out-of-vocabulary and rare words problems, which are caused by the rich derivational and inflectional morphology of the Turkish language.

Title and Abstract in Turkish

AnlamVer: Anlambilimsel Model Ölçümleme Veri Kümesi - Kelime Benzerliği ve İlişkiselliği

Bu makalede, Türkçe için kelime benzerlik ve ilişkisellik görevlerini ayrı ayrı ölçümleyebilmek için tasarlanmış anlambilimsel bir model veri kümesi olan AnlamVer'i sunuyoruz. Veri kümemiz, 12 kişi tarafından her kelime çifti için ilişkisellik ve benzerlik puanları ayrı ayrı puanlanmış toplam 500 kelime çiftinden oluşmaktadır. Kelime çiftleri, dağılımsal anlambilimsel modelleri kelimelerin ve kelime çiftlerinin sıklık, morfoloji, somutluk ve ilişki tipleri (eş anlamlılık, karşıt anlamlılık vb.) gibi birden fazla niteliğine göre ölçümleyebilmek için seçilmiştir. Amacımız, anlambilimsel model araştırmacılarının modellerini birden fazla niteliğe göre ölçümleyerek içgörüler kazanmasını sağlamaktır. Veri kümesindeki kelime çiftleri, Türkçe'nin zengin türetimsel ve çekimsel yapısı kaynaklı sözlük-dışı ve seyrek-kelime problemlerine karşı anlambilimsel modellerin sağlamlığını ölçümleyebilmek amacıyla sıklık değerlerine göre dengelenmiştir.

1 Introduction

Unsupervised semantic modeling has recently gained a lot of attention in the NLP community due to the notion of high reusability of pre-trained models across a variety of higher level NLP tasks such as machine translation, word sense disambiguation and named entity recognition. Increasing computability of unsupervised distributional semantic modeling (DSM) methods enable researchers to increase the performance of NLP tasks by leveraging extracted semantic information from a high volume of unstructured texts at low costs. However, there are very few available methods and resources to evaluate semantic models intrinsically regardless of the higher level tasks' dynamics. Presenting word similarity and word relatedness (i.e., association) dataset AnlamVer, we aim at providing the semantic modeling field for Turkish with an intrinsic evaluation resource targeting morphology driven issues caused by the rich agglutinative nature of the language. We are not aware of the existence of such word similarity or

Proceedings of the 27th International Conference on Computational Linguistics, pages 3819–3836
Santa Fe, New Mexico, USA, August 20-26, 2018.

relatedness evaluation resources constructed for Turkish. In this paper, we describe design considerations and data collection guidelines we followed in the construction of such dataset as well as dataset statistics.

The main contributions of this paper include: (i) an introduction of a first word similarity and word relatedness evaluation dataset for a low-resource language Turkish[1], (ii) design considerations on the construction of a dataset where the main objective is balancing the words and the words-pairs by multiple morphological and semantic attributes, (iii) a novel analysis and visualization of a word similarity and relatedness dataset containing bi-dimensional values for each word-pair and, (iv) a publicly available web-based word similarity questionnaire software.[2]

2 Background and Design Motivations

Word similarity evaluation (i.e., wordsim) is one of the oldest intrinsic methods of semantic model assessment. For example, RG dataset Rubenstein and Goodenough (1965) is still being used today as one of the gold standards in the DSM research. Wordsim datasets are constructed by asking human subjects to assign numeric scores from 0 to 10 for every pre-selected word-pairs. In this section, we describe some issues of word similarity evaluation and design decisions we made to overcome such issues in our study.

2.1 Similarity and Relatedness Confusion

Linguistic Background

Linguists have been studying on statistical distributions of linguistic items (words) for a century. Although the distributional hypothesis *"words that occur in similar contexts, tend to have similar meanings"* is commonly traced back to Harris (1954), according to Sahlgren (2006), theoretical foundations of his distributional methodology go back to structuralist linguist Bloomfield (1887 - 1949) and Ferdinand de Saussure (1857 - 1913). De Saussure et al. (2011) pointed out that there can be distinctive functional roles of *signs* within the language system. He defined functional differences of linguistic elements in two (orthogonal) types which are widely studied with distributional relations in distributional semantics (DS) research today: *syntagmatic* and *paradigmatic*. Briefly, "words have a syntagmatic relation if they co-occur, and a paradigmatic relation if they share same neighbors" (Sahlgren, 2006). Paradigmatic words represent similar concepts or entities of the real world which are most likely substitutional in the context. One example is the synonyms like "clever" and "smart" in the sentence "She is very [clever | smart]." where two words are less likely to occur at the same sentence.

Lack of Distinction in Word Evaluation

Following the theoretical distinction of paradigmatic and syntagmatic relation types, one can easily apply such distinction to word evaluation by making the assumption that "similarity represents paradigmatic and relatedness represents syntagmatic type of relations.". However, semantic research has not paid as much attention to this distinction as necessary. Two most comprehensive DSM benchmark studies, Baroni et al. (2014) and Levy et al. (2015), reported model performances based on *wordsim* datasets such as RG (Rubenstein and Goodenough, 1965), WordSim-353 (Finkelstein et al., 2001) and MEN (Bruni et al., 2012). In their study, Hill et al. (2016) thoroughly describe the distinction between similarity and relatedness (i.e., sim-rel) caused limitation problems of such datasets. Hill et al. (2016) also define the criteria for evaluation datasets in three: representative, *clearly-defined*, consistent and reliable. Most wordsim datasets such as RG, MC (Miller and Charles, 1991), WordSim-353, and MEN do not satisfy clearly-defined criteria since their screen guidelines use both "similarity" and "relatedness", and "association" words in place of each others. One good example for guideline ambiguity is the following instructions from the WordSim-353 study: *"...please assign a numerical similarity score between 0 and 10 (0 = words are totally unrelated), 10 = words are VERY closely related..."*. Since our study aims to collect both similarity and relatedness scores from participants, we provided *clearly-defined* detailed instructions in the questionnaire screens (Figure 2).

[1]Dataset is publicly available at `http://www.gokhanercan.com/anlamver`

[2]Source code and a demo application are publicly available at `http://www.gokhanercan.com/wsquest`

One Model Does Not Fit All

Agirre et al. (2009) detected the aforementioned sim-rel confusion and split the original WordSim-353 dataset into two datasets (WS-Rel and WS-Sim) by classifying word-pairs based on their relationship types. Thus, they solved the dataset's sim-rel distinction problem without re-scoring word-pairs.[3] They proposed two separate models for similarity and relatedness evaluation tasks. For example, they reported that the context-windows-based approach is better at capturing similarity (evaluated on WS-Sim) while the bag-of-words approach is at relatedness (evaluated on WS-Rel). Capturing the similarity seems arguably harder for *the distributional hypothesis* based unsupervised models compared to relatedness models. Examining the DSM benchmark study of Levy et al. (2015), average performances of all model configurations consistently perform the worst on the SimLex-999 similarity dataset (≈ 0.39) compared to relatedness (traditional wordsims) (≈ 0.70) datasets.[3] [4]

Similarly, Hill et al. (2016) focus only on similarity evaluation while clearly informing annotators about the sim-rel distinction in their SimLex-999 dataset work. In another dataset study, SimVerb-3500 (Gerz et al., 2016), only distributional verb semantics with a large scale (3,500 word-pairs) of verb similarity evaluation is considered. We observe that DSMs have been starting to be divided into more specific models (e.g., relatedness, similarity, antonymy), motivated by the better performance requirements of the higher-level tasks. As Faruqui et al. (2016) point out, intrinsic wordsim evaluation does not correlate well with extrinsic NLP tasks' evaluation results. Wordsim sim-rel confusion might be one of the reasons for this inconsistency. It is an open question whether a single pre-trained DSM can represent the semantics of a domain consistently across multiple higher-level tasks. We think that a *perfect* DSM would be a multi-model structure which could handle every specific relation types (e.g., relatedness, similarity, antonymy, hypernymy, meronymy) of the words with maximum performances. In this study, our dataset targets Turkish language specific DSMs with two main types of semantic relations (similarity and relatedness) by evaluating word-pairs on both at the same time. We are not aware of the existence of such dataset study for any language.

Sim-Rel Vector Space

Instead of splitting word-pairs into two distinct groups, we decided to get two scores for every word-pair: similarity and relatedness. By doing so, we can keep the dataset as a single unit while evaluating semantic models in two relation types. Two-dimensional structure of our evaluation data structure allows us to visualize the semantic space of the dataset through a scatter plot diagram we named "Sim-Rel vector space" (Figure 1).

Given x and y axes represent relatedness r and similarity s scores of each word-pair in the dataset respectively, and variables r and s (orthogonality of paradigmatic and syntagmatic relations) range from 0 and 10. Let SU similar-unrelated, SR similar-related, DU dissimilar-unrelated, DR dissimilar-related are categorical labels of possible semantic sub-spaces ss, $ss = f_1(r, s)$ function would be,

$$ss = f_1(r, s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

And let $t = 2$ denotes a threshold variable that represents boundary point of relation-type-spaces where $synonym, antonym, irrelevant$ are categorical labels of possible semantic relation-types rt, $rt = f_2(r, s)$ function would be,

$$rt = f_2(r, s) = \begin{cases} \text{synonym}, & \text{if } 10 - t \leq s \text{ and } 10 - t \leq r \\ \text{antonym}, & \text{if } 10 - t \leq r \text{ and } s \leq t \\ \text{irrelevant}, & \text{if } t \geq r \text{ and } t \geq s \end{cases}$$

[3] Since participants scored under ambiguous guidelines, WS-Sim is not inherently a similarity dataset. See (Hill et al., 2016).

[4] Low scores on RW dataset are plausible because it targets OOV and rare words problems.

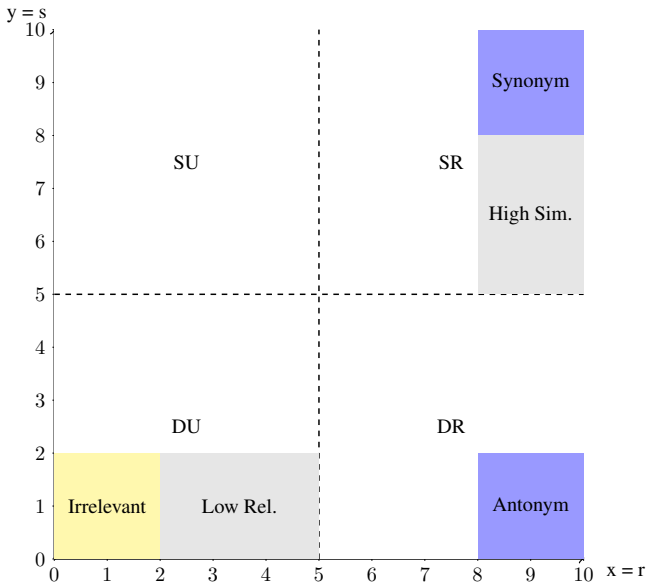


Figure 1: Sim-Rel vector space of word-pairs.

If we assume that participants are asked to score lower for similarity s (closer to 0) in the case of their antonym judgements for word-pairs, and asked to score higher (closer to 10) in the case of their synonym judgements[5], following the definition of the Sim-Rel vector space functions f_1 and f_2 above, followings can be inferred:

- A perfect DSM could assign word-pairs to every semantic sub-space ss with 100% accuracy.
- No word-pair instance is expected to be assigned to a similar-unrelated SU sub-space. Semantically, all highly similar word-pairs should also be highly related. For example, "car - automobile" word-pair is highly similar. Since they are very likely to share many common neighbors in their contexts, their relatedness score should be high, too.
- Word-pairs could be accepted as synonyms if their rt value is assigned to $synonym$ varying by the t parameter. The same rule applies to the $irrelevant$ value, too. We intuitively chose the threshold $t = 2$ value for Sim-Rel semantic space. We kept $t = 2$ same for all axes and relations for the sake of model and visualization simplicity. We leave the theoretical or empirical investigation of selection strategies of such threshold values for future work.
- Antonym-DR-overlapping problem: No DSM could perfectly assign rt as an $antonym$. Boundaries between the antonyms and dissimilar-related DR word-pairs are semantically ambiguous. Our bi-dimensional evaluation model cannot differentiate between the two. For example, word-pairs "tense - loose" and "red - rose" could get closer r and s scores while the first one can semantically be an antonym but the latter is obviously not. Asking to score lower for antonym judgements is a common method in *similarity* datasets (Hill et al., 2016) (Gerz et al., 2016). Storing two inherently different relation types' scores (synonym, antonym) in a single s variable is the root cause of the problem. It is the downside of our Sim-Rel vector space model. We leave the resolution of this problem for future work.

2.2 Out-Of-Vocabulary and Rare Words Problems

Turkish is an agglutinative language which has a highly inflectional and derivational morphology. The agglutinative nature allows forming new words by stringing stem, morphemes and suffixes together. In

[5]Scoring closer numbers to zero for antonyms is a common convention for similarity datasets (Hill et al., 2016), (Gerz et al., 2016).

Turkish, words are bound-morphemes, which means that there can be only one lexical stem (root) of a word. Since Turkish has many productional affixes (e.g., CHk, CA, CI, lHK, SHz, HmsH), theoretically unlimited surface words can be generated. Table 1 shows inflections and derivations of various words in morphologically decomposed forms where every word share the same lexeme "maymun" (monkey). In this study, all morphological decompositions are performed by the toolkit from Görgün and Yildiz (2011).

Word	Decomposition	Sense	Form	Frequency
maymun	maymun	monkey	root form	very
maymunları	maymun + lAr + sH	their monkeys	inflectional	medium
maymunsu	maymun + sI	ape, like monkeys	derivational (usual deviance)	rare
maymungilleri	maymun + gil + lAr + yH	family of monkeys, primades	derivational (acceptable deviance)	oov
maymuncuk	maymun + CHk	skeleton key, picklock (tool)	derivational (deviant)	rare

Table 1: Morphological decomposition of various words sharing the same lexeme.

Simple word-based models ignore the internal structure of words which reduces the model's capabilities and qualities. The main problem is zero (i.e., unseen, out-of-vocabulary) or low occurrence (i.e., rare words) of a testing word in the training corpus. Distributional semantics (DS) community has been developing more complex subword-level (i.e., compositional) models to overcome *out-of-vocabulary* (OOV) and *rare words* problems. DSMs for morphologically-rich languages must alleviate OOV and rare words problem to generalize better. RW dataset (Luong et al., 2013) provides word frequency (i.e., rareness) based evaluation strategy to compositional model developers. Similarly, we aim to balance our dataset's word-pool by words' frequencies to assess generalization powers of such models.

In addition to the traditional OOV and rare words evaluation strategy, we applied another concept that we named *made-up words* by injecting *novel* (i.e., made-up, fake) words into the word-pool of our dataset. Vecchi et al. (2017) applied this concept to their phrase-level models to test the model's creative capabilities (i.e., generalization power). The main idea is as follows: even if people hear a word for the first time and it might sound odd to them, people have the intuition to make sense of the intended meaning. Could DSMs do the same? In our subword-level case, we assume that Turkish affixes can change the meanings of the words in a consistent manner, which is called *acceptable semantic deviance*. For example, the word "maymungilleri" (family of monkeys) is a made-up word and it sounds odd to any native Turkish speaker (Table 1) in the first place. But almost every native speaker can understand what it's meant to some extent. This productivity feature of a language can be seen as a substantial model generalization potential for a researcher. However, the downside of the concept is that the assumption does not always hold as in the word "maymuncuk" (skeleton key, a tool) (Table 1). In this example, the word is derived from the root word (i.e., lexeme) "maymun" (monkey) by getting the affix "cHk" with a valid state transition but its *one sense's* meaning shifts to an entirely different space. It is a very challenging problem for compositional DSMs. Vecchi et al. (2017) name this type of semantically-lossy derivations as *deviants*.

2.3 Dataset Translation Issues

Before starting the dataset construction phase, we have considered translating the existing well-known wordsim datasets to Turkish. After completing MC dataset's translation, we conclude that constructing a new dataset would be more meaningful and reliable than translating the existing ones. The main reasons behind our decision can be summarized as follows:

1. Both words in the source word-pair map to the same single word in the target language: *"football - soccer"* $\rightarrow$ *"futbol - futbol"*.

2. A word in the source word-pair maps to a multi-word phrase: *"asylum - madhouse"* $\rightarrow$ *"tımarhane - akıl hastanesi"*. Traditional wordsim datasets and DSMs ignore phrases for the sake of model and evaluation simplicity. We left phrases out of the scope of this study.

3. Meaning shifts in translations require re-annotation from human resources for every word-pair. The human annotation stage is one of the costly parts of the study. [6]

4. We aim to balance words and word-pairs in as many attributes as possible such as word frequency, derivations, inflections, concreteness and relation types. Word frequencies and morphological features are pretty much language-dependent.

3 Dataset Design and Methodology

Design motivations we borrowed from the previous section can be summarized as follows: (i) collecting two-dimensional relatedness and similarity scores from participants while clearly-defining distinctions of such concepts, (ii) making it language-specific morphological dataset which can evaluate DSMs' generalization power concerning OOV, rare words and semantic deviance scenarios, and (iii) balancing the dataset by multiple morphological and semantic attributes as much as possible. Due to the time and budget limitations, we set the target dataset size of the project to 1.000 scores (500 word-pairs) as most of the wordsim datasets include fewer scores (SimLex-999=999, RG=65, M30, WordSim-353=353, RW=2,034, MEN=3,000). We planned dataset construction process in three stages: word candidates, word pool and word-pairs selections (Table 2).

	Stage 1	Stage 2	Stage 3
	1) Word Candidates (starts)	2) Word-Pool Selection	3) Word-Pairs Selection
Goals	1.1) Reusing existing resources	2.1) Balancing word attributes by estimations	3.3) Balancing word-pairs by estimations
Input	1.2) TKN (600) + MC (39)	2.2) Stage1 output (639) + new derivational words (99)	3.2) 320 Stage2 words
Process	1.3) Attaching frequencies, morphological tags	2.3) Filtering for balancing	3.3) Mapping pairs (every word used 2-5 times building word-pairs)
Output	1.4) 639 words	2.4) 320 words	3.4) 500 word-pairs (ends)

Table 2: Three stages of dataset construction.

3.1 Word Candidates Selections

TKN Dataset

Since Turkish is a low-resource language in NLP research, we aimed to re-use existing resources as much as possible. We investigated word candidates that already have some informative attributes. We used word norms "Türkçe Kelime Normları" (TKN) dataset (Tekcan and Göz, 2005) which is constructed for a psycholinguistics study. TKN consists of 600 Turkish words which are balanced in terms of *concreteness* (half concrete, half abstract) values. TKN's concreteness values are annotated by 100 voluntary university students. As in the English USF word norms dataset (Nelson et al., 2004), TKN words range between 1 and 7. Lower values denote more abstract and higher values denote more concrete concepts. For instance, the word "mutluluk" (happiness) takes 1.85 while the word "gül" (rose) 6.79. By choosing candidates from TKN, we enabled model developers to evaluate their models on various concreteness levels. Unfortunately, TKN contains very frequent and root-formed word dataset with 480 words in lexical root form where none of the 120 non-root form words are inflected.[7] Those limitations led us to add 99 words (with no concreteness values) manually on the next stage in order to achieve the balancing goal of the dataset.

[6] Considering human resources, questionnaire software and data pre/post-processing costs.

[7] 108 words having one derivations, 12 words having two derivations.

3.2 Word-pool Selections

Having 600 candidate words that are transferred from the first stage, our goal was to narrow down them to 320 words (word-pair candidates) while preserving our dataset balancing requirements. Table 3 shows word-pool grouping attributes along with the number of words and percentages.

Frequency-based Balancing

Considering the importance of OOV and rare words problems in modeling for morphologically rich and productive languages, our priority was to balance our word-pool based on word frequencies. RareWords dataset (Luong et al., 2013) addresses this issue by grouping words by their frequencies into four groups $(5-10, 10-100, 100-1000, 1000-10000)$. Since the RareWords dataset is designed for the English language (which is relatively less inflectional and productive than Turkish), researchers may assume that words with frequencies lower than five are most likely to be junk or non-English words. In our case, a single Turkish lexicon can take thousands of surface forms. Our own corpus coverage analysis shows that 47% of word types (277K) occur only once in the corpus, which is compatible with Boun Corpus word coverage statistics (Sak et al., 2011). Therefore, we couldn't ignore the words that have zero or less than five frequencies. We applied a different grouping strategy, where the first group is OOV (zero frequency) and rare words in five groups $RW1, RW2, RW3, RW4, RW5$. Table 3 displays how OOV and rare words (RW) groups are represented in the word-pool. We made the frequency analysis on Boun Corpus (Sak et al., 2011) which consists of roundly 3.2 million token types (i.e., vocabulary size). We defined frequency groups $(0-32, 32-320, 320-3200, 3200-32000, 32000-\infty)$ by using the $gr(n, voc, g)$ function below, where g is the number of groupings, n is the index of each group varying between 1 to g, and voc is the vocabulary size of the given corpus. The only exception is the minimum and maximum values of the first and last groups are fixed to 0 and ∞ respectively. Ampersand symbols (&) denotes string concatenations:

$$gr(n, voc, g) = (voc \times 10^{-(g-n+3)}) \text{ \& "-" \& } (voc \times 10^{-(g-n+2)})$$

3.3 Word Pair Selections

In the word-pair selection stage, we matched words from word-pool with each other to form new pairs. We defined a constraint that every word in the word-pool should occur in the matching word-pairs up to five times. The primary goal of this mapping stage was to find, 500 word-pair relations, which are balanced explicitly by new type attributes: estimated semantics relations are synonym, antonym, hypernym, meronym etc. We manually matched and estimated the semantic type of the relationships. For example, we manually picked two potentially similar words "otomobil" (automobile) and "araba" (car) from the pool and flagged them as a strong synonym potential. We defined 50 synonym, 50 antonym, 50 meronym, 50 hypernym relations. Similarly, we grouped word-pairs by estimated magnitudes (low, medium, high) of relatedness relations too. Table 4 shows the number of actual instances and percentages of such estimation-based groupings and morphological groupings of word-pairs. Finally, we ended up with 500 manually-selected, grouped word-pairs. Table 5 shows some sample annotated word-pair instances from the final dataset.

4 Questionnaire Design

4.1 Platform

We built a web-based application to collect data from human annotators. Participants were asked to score similarity and relatedness relationship for every 500 word-pairs. In total, every participant scored 1,000 answers for 500 word-pairs. We split the questionnaire into two sections. The first section starts with describing what similarity relation is, along with five examples with detailed descriptions. Similar to the Simlex-999 guidelines, we asked users to score low for antonyms and score high for synonyms. We described similarity as follows (showing first two sentences only):
"Two words are similar if they refer to the same thing, person, concept or action. Similar things share common concrete or abstract attributes. For example, 'tea' and 'coffee' are quite similar because both

	G0	G1	G2	G3	G4	G5	Total
Frequency	OOV	RW1	RW2	RW3	RW4	RW5	
	31 9.6%	33 10.3%	30 9.3%	62 19.3%	111 34.6%	53 16.5%	320 100%
Concreteness	no value	abstract	medium	concrete			
	149 46.5%	35 10.9%	30 9.3%	106 33.1%			320 100%
Root Form	root	non-root					
	182 56.8%	138 43.1%					320 100%
Derivations	no der.	der1	der2+				
	198 61%	81 25.3%	41 12.8%				320 100%
Inflections	no inf.	inf1	inf2+				
	277 86.5%	17 5.3%	26 8.1%				320 100%

Table 3: Groupings of the word-pool.

	G0	G1	G2	G3	G4	G5	Total
Est. Synonyms	synonym	antonym	other				
	50 10%	50 10%	400 80%				500 100%
Est. Relatedness	high	medium	low				
	200 40%	150 30%	150 30%				500 100%
Est. Rel. Type	hyponym	meronym	other				
	50 10%	50 10%	400 80%				500 100%
OOV	no oov	any oov	two oov				
	434 86.8%	66 13.2%	42 8.4%				500 100%
Min. Derivations	no der.	der1	der2+				
	231 46.2%	166 33.2%	103 20.6%				500 100%
Min. Inflections	no inf	inf1	inf2+				
	424 84.8%	32 6.4%	44 8.8%				500 100%
Min. RareWord	rw0 (oov)	rw1	rw2	rw3	rw4	rw5	
	66 13.2%	65 13%	62 12.4%	130 26%	142 28.4%	35 7%	500 100%

Table 4: Groupings of the word-pairs.

are relaxing hot drinks gathered from nature and irreplaceable beverages for friendly conversations." Figure 2 shows a snapshot from the initial guideline screen for similarity annotation. When a participant presses the "ileri" (next) button, first word-pair page appears, asking to score 20 word-pairs per screen (Figure 3).

4.2 Participants

All 12 participants are native Turkish speakers who voluntarily participated in the questionnaire. Eight participants are female, and four participants are male. Both mean and median values of ages are 33.5

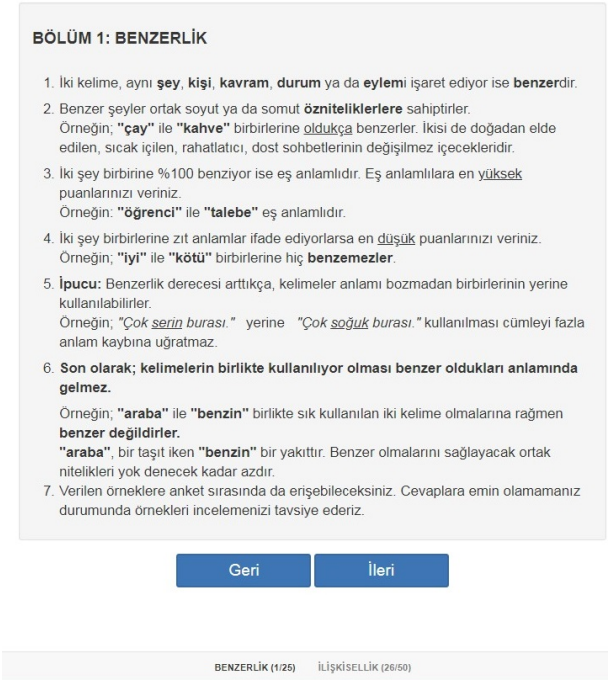


Figure 2: Similarity instructions page.

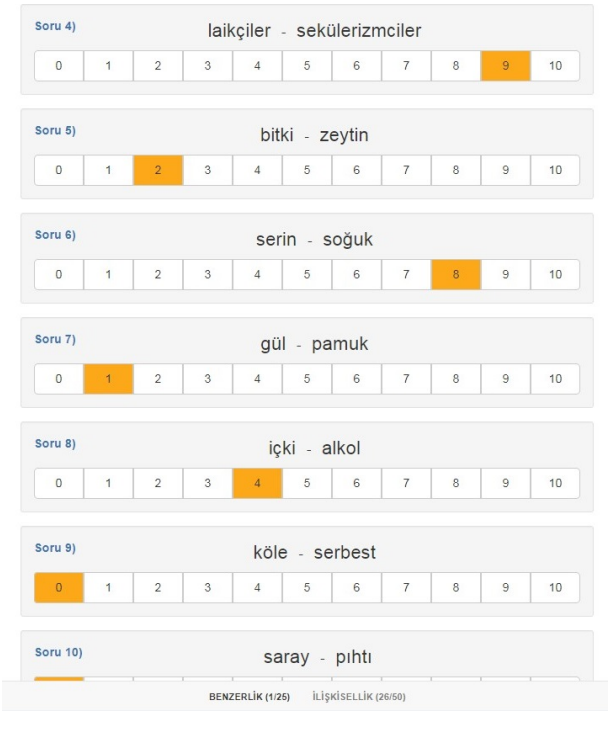


Figure 3: Word-pair annotation page.

with the standard deviation of 9.3. Nine participants are university graduates (seven participants with the master's degree), two participants are undergraduate, and a participant holds a high school degree. Participants were asked to join the questionnaire remotely using their web browsers by following the invitation link sent to their mailboxes. Questionnaire software WSQuest's[8] responsive layout support, allowed users to score quickly from mobile and tablet devices. They are asked to read user guidelines carefully since no participant had the prior knowledge about the flow of the annotation process, and word similarity and relatedness concepts. Initial guideline screen informed users that they could score the questionnaire freely at any time of the day in three days since questionnaire software allows users to continue their sessions easily as long as they keep the last completed URL of the software. Without giving any breaks, it took participants' 75 minutes on average to complete the entire questionnaire.

5 Dataset Analysis

Final (actual) similarity s and relatedness r values of the dataset seems consistent with our estimations. Under the aforementioned Sim-Rel vector space model assumptions and configuration, scatter plot of the average values of s and r yielded a visual similar to our expectations (Figure 4). Our observations about the actual data distribution as follows:

- SU subspace remain empty as expected. Participants proved that word-pairs cannot be similar and unrelated at the same time.
- Antonym-DR-overlapping problem holds. We observed very close word-pair values for antonym and DR word-pairs. For example, average s and r scores of "kırmızı - gül" (red - rose) word-pair are 1.16 and 7.16 respectively. On the other hand, an antonym-estimated word-pair "şeffaf - opak" (transparent - opaque) has exactly the same scores as the former one (see Table 5).
- Participants scored word-pairs that include made-up words normally as regular word-pairs. For example, annotators scored the word-pair "atatürkist - kemalci" (atatürk+İST - kemal+CH) as 8.75

[8]See appendices or `http://www.gokhanercan.com/wsquest` for complete user screen guidelines.

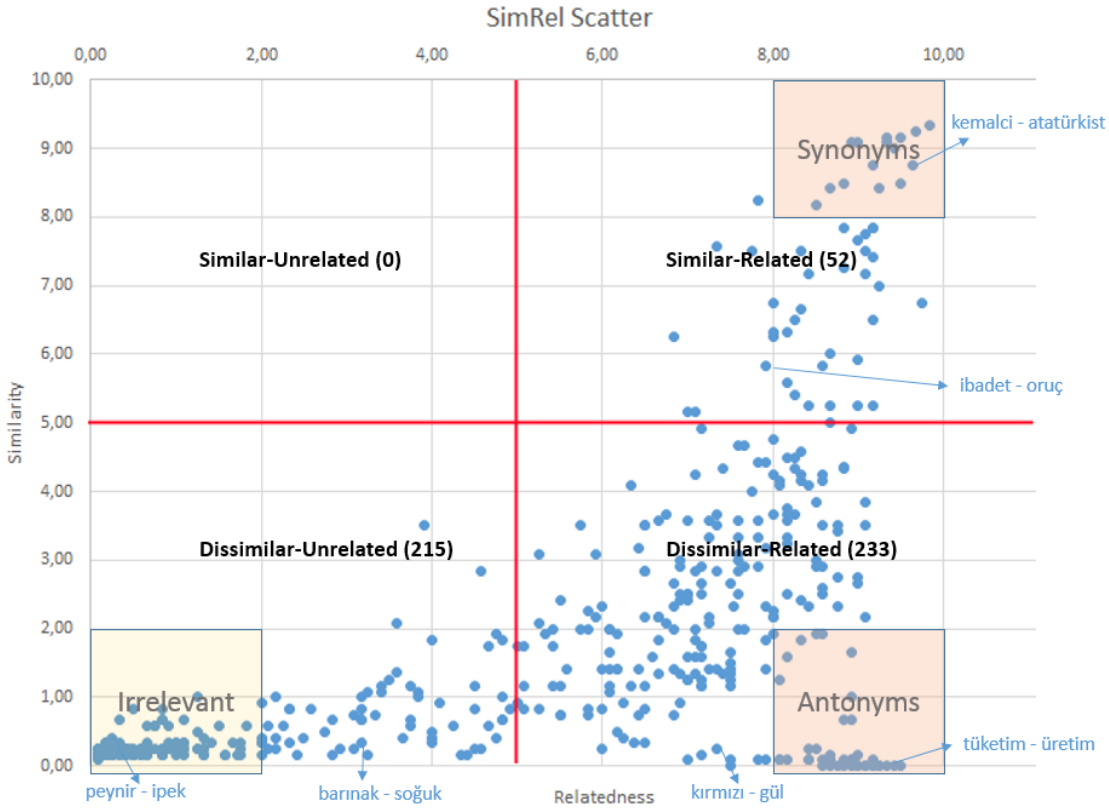


Figure 4: Scatter plot of the final (actual) dataset. Data-points denote participants' avg. Sim-Rel score of each word-pair where y axis is s and x axis is r. Member counts of $ss\{SU, SR, SU, DR\}$ semantic sub-spaces are in the parenthesis.

similar and 9.63 related (see Table 5) where neither of surface forms has the common usage in Turkish (OOV and RW1 respectively). Both İST and CH suffixes have usages to change words' meaning to "ideological adherence to a person/thing" where both names "Atatürk" and "Kemal" are parts of the name "Mustafa Kemal Atatürk" who is the founder of the Republic of Turkey.

5.1 Post-processing and Inter-annotator Argument

Since questionnaire includes many uncommon OOV and rare words pairs, it allows users to skip that word-pair empty if they don't have any idea about the meanings. However, null answering rate (0.1%) is quite lower than we expected. In order to calculate ranking correlations properly, we replaced null answers with the average score of all users' answers for that question. Among 16 participants, we do some post analysis on collected data. We detected that one participant achieved marginally low (0.32 min, 0.57 max) Spearman ranking correlation score compared to the other participants. After a little further investigation, we noticed that this participant had completed the test only in 25 minutes. It is three times faster than what we estimated for a high-quality annotation. Similarly, we increased the overall data quality by removing three more participant's answers too. After post-processing, we computed 0.748 average pairwise inter-annotator ($apia$) score where the highest pairwise correlation of users is 0.847, and the lowest one is 0.474. Even though dataset's $apia$ score is lower than we expected, 0.748 is still higher than the most word similarity datasets (WS-Sim=0.667, MEN=0.68, 0.67=SimLex-999). Based on the study of (Snow et al., 2008), more than ten annotators are statistically acceptable for word similarity evaluation task's reliability.

word1	word2	sim.	rel.	oov	conc.	ss	der#	inf#
otomobil (automobile)	araba (car)	9.16	9.33	no	6.87	SR (syn.)	0	0
üşengen* (lazy)	üşengeç (lazy)	8.25	7.83	one	3.06	SR	2	0
atatürkist* (adh .to Atatürk)	kemalci (adh. to Kemal)	8.75	9.63	one	-	SR	2	0
kitaplıklar (bookshelves)	kitaphane* (place with books)	7.16	8.41	no	-	SR	2	1
kemalci (adh. to Kemal)	kemalizmcilerden (from ...)	5.25	8.66	one	-	SR	3	3
kırmızı (red)	gül (rose)	1.16	7.16	no	6.79	DR	0	0
şeffaf (transparent)	opak (opaque)	1.16	7.16	no	4.37	DR	0	0
zarar (loss)	kazanç (profit)	0.18	8.8	no	3.25	DR (ant.)	0	0
gevşek (loose)	heykel (statue)	0.16	0.16	no	-	DU (irr.)	0	0
üşengen* (lazy)	yedigen (heptagon)	0.16	0.25	two	-	DU	2	0

Table 5: Sample word-pairs from the final dataset. Words with asterisks (*) are made-up words. (adh = adherent, conc = concreteness, ss = semantic sub-space, syn = synonyms, ant = antonyms, irr = irrelevant, der# = total derivations, inf# = total inflections)

6 Conclusion

We presented a semantic model evaluation dataset for the Turkish language. Turkish morphology requires complex semantic models to alleviate OOV and rare words problems. Since the dataset includes 13% OOV and 26% rare-word-pairs (RW1 and RW2), we think that it will be a challenging intrinsic evaluation task for DSM researchers. Hopefully, AnlamVer-evaluated distinct similarity and relatedness models correlate better with the higher level NLP tasks. For future work, we are planning to construct a bigger dataset, leveraging existing lexical resources such as Turkish WordNet (Ehsani et al., 2018) which already includes manually-annotated synonymy (i.e., synsets), antonymy and hypernymy relations.

Acknowledgements

This work was supported by Tübitak project 116E104.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Ferdinand De Saussure, Wade Baskin, and Perry Meisel. 2011. *Course in general linguistics*. Columbia University Press.

Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):24.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Onur Görgün and Olcay Taner Yildiz. 2011. A novel approach to morphological disambiguation for turkish. In *Computer and Information Sciences II*, pages 77–83. Springer.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Institutionen för lingvistik.

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

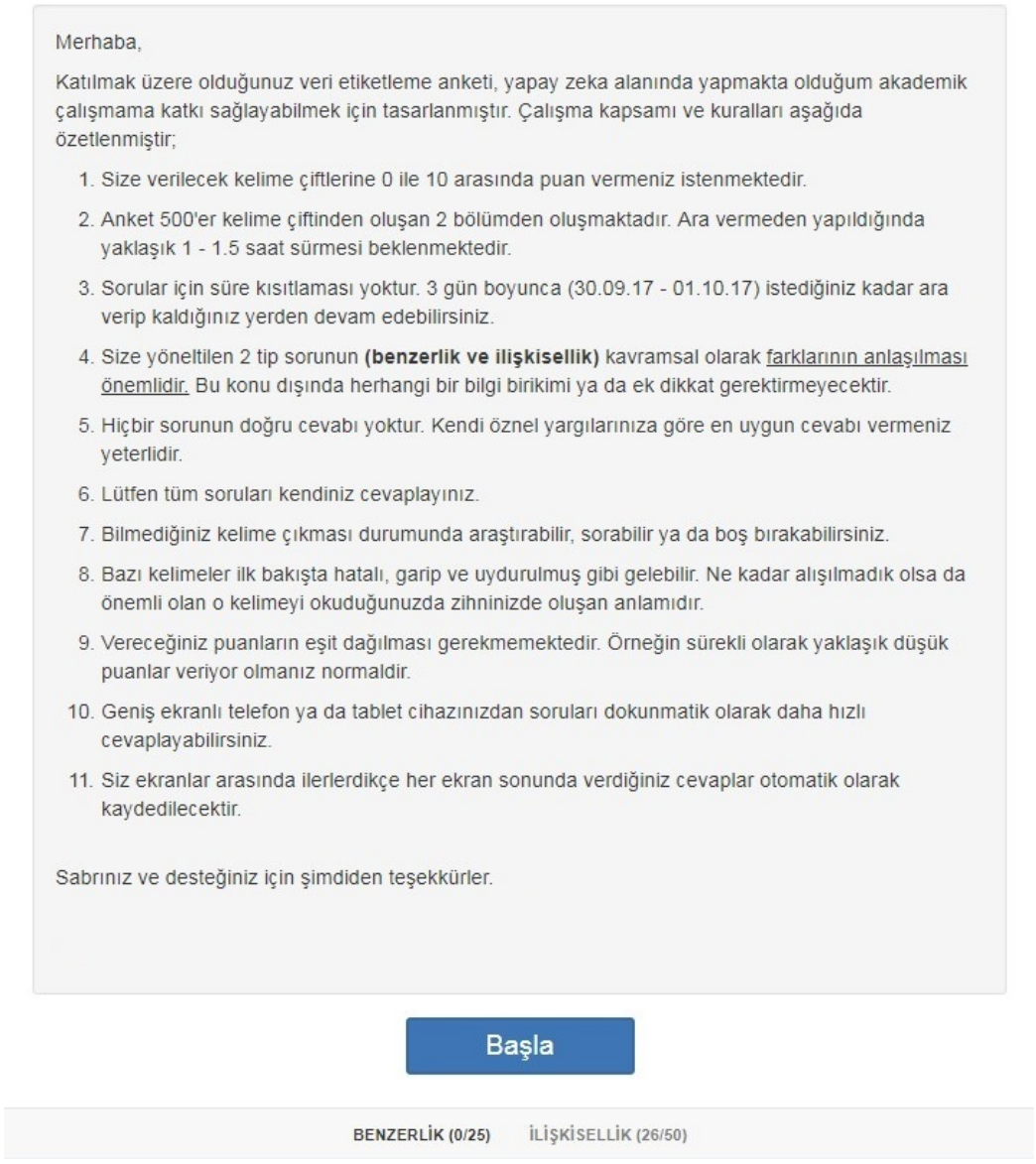
Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Ali İ Tekcan and İlyas Göz. 2005. Türkçe kelime normları. *İstanbul Boğaziçi Üniversitesi*.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, 41(1):102–136.

.

Appendices


Merhaba,
Katılmak üzere olduğunuz veri etiketleme anketi, yapay zeka alanında yapmakta olduğum akademik çalışmama katkı sağlayabilmek için tasarlanmıştır. Çalışma kapsamı ve kuralları aşağıda özetlenmiştir;
1. Size verilecek kelime çiftlerine 0 ile 10 arasında puan vermeniz istenmektedir.
2. Anket 500'er kelime çiftinden oluşan 2 bölümden oluşmaktadır. Ara vermeden yapıldığında yaklaşık 1 - 1.5 saat sürmesi beklenmektedir.
3. Sorular için süre kısıtlaması yoktur. 3 gün boyunca (30.09.17 - 01.10.17) istediğiniz kadar ara verip kaldığınız yerden devam edebilirsiniz.
4. Size yöneltilen 2 tip sorunun (benzerlik ve ilişkisellik) kavramsal olarak farklarının anlaşılması önemlidir. Bu konu dışında herhangi bir bilgi birikimi ya da ek dikkat gerektirmeyecektir.
5. Hiçbir sorunun doğru cevabı yoktur. Kendi öznel yargılarınıza göre en uygun cevabı vermeniz yeterlidir.
6. Lütfen tüm soruları kendiniz cevaplayınız.
7. Bilmediğiniz kelime çıkması durumunda araştırabilir, sorabilir ya da boş bırakabilirsiniz.
8. Bazı kelimeler ilk bakışta hatalı, garip ve uydurulmuş gibi gelebilir. Ne kadar alışılmadık olsa da önemli olan o kelimeyi okuduğunuzda zihninizde oluşan anlamıdır.
9. Vereceğiniz puanların eşit dağılması gerekmemektedir. Örneğin sürekli olarak yaklaşık düşük puanlar veriyor olmanız normaldir.
10. Geniş ekranlı telefon ya da tablet cihazınızdan soruları dokunmatik olarak daha hızlı cevaplayabilirsiniz.
11. Siz ekranlar arasında ilerlerdikçe her ekran sonunda verdiğiniz cevaplar otomatik olarak kaydedilecektir.
Sabrınız ve desteğiniz için şimdiden teşekkürler.
Başla
BENZERLİK (0/25)
İLİŞKİSELLİK (26/50)

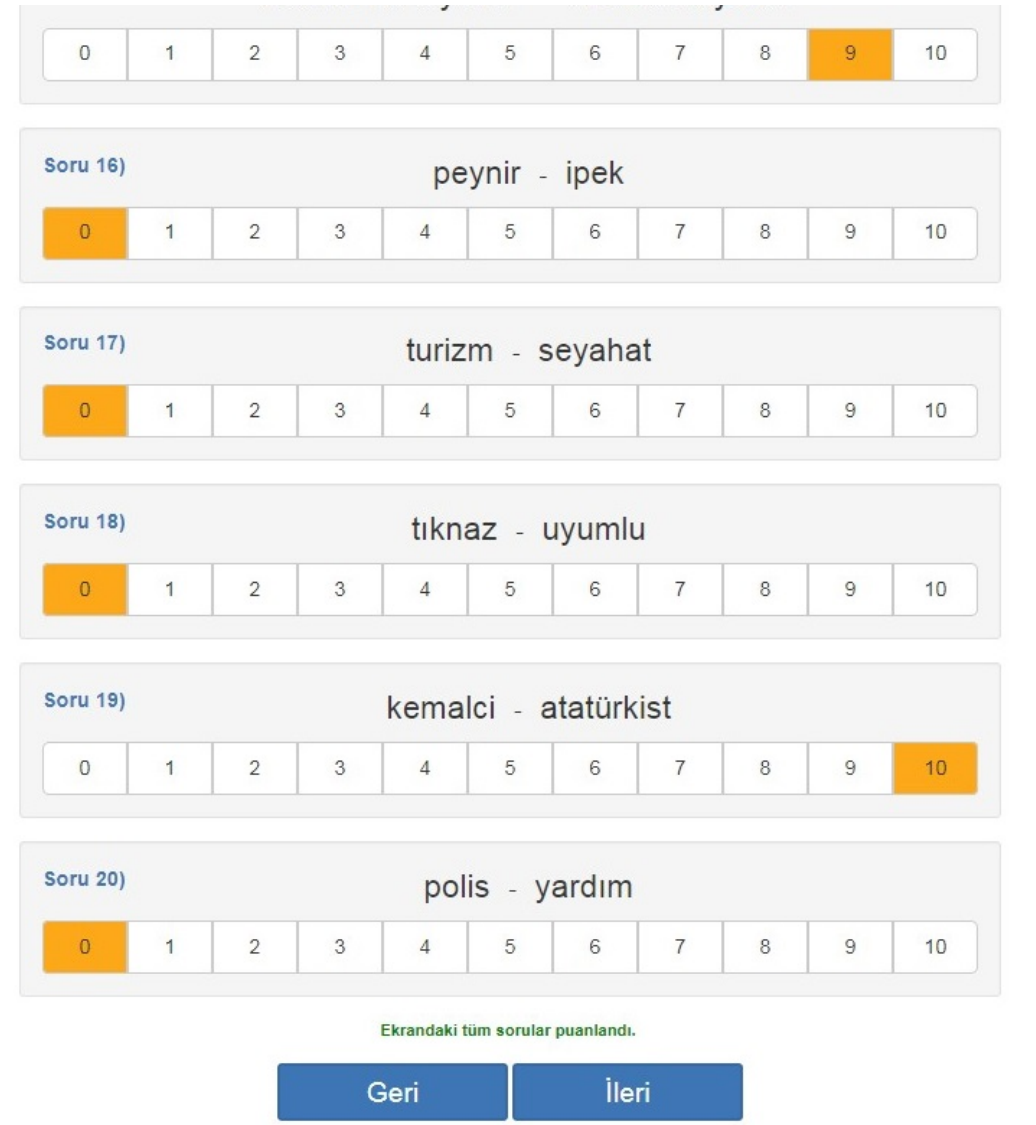
BÖLÜM 1: BENZERLİK

1. İki kelime, aynı **şey**, **kişi**, **kavram**, **durum** ya da **eylem**i işaret ediyor ise **benzer**dir.
2. Benzer şeyler ortak soyut ya da somut **özniteliklerlere** sahiptirler.
   Örneğin; **"çay"** ile **"kahve"** birbirlerine oldukça benzerler. İkisi de doğadan elde edilen, sıcak içilen, rahatlatıcı, dost sohbetlerinin değişilmez içecekleridir.
3. İki şey birbirine %100 benziyor ise eş anlamlıdır. Eş anlamlılara en yüksek puanlarınızı veriniz.
   Örneğin: **"öğrenci"** ile **"talebe"** eş anlamlıdır.
4. İki şey birbirlerine zıt anlamlar ifade ediyorlarsa en düşük puanlarınızı veriniz.
   Örneğin; **"iyi"** ile **"kötü"** birbirlerine hiç **benzemezler**.
5. **İpucu:** Benzerlik derecesi arttıkça, kelimeler anlamı bozmadan birbirlerinin yerine kullanılabilirler.
   Örneğin; *"Çok serin burası."* yerine *"Çok soğuk burası."* kullanılması cümleyi fazla anlam kaybına uğratmaz.
6. **Son olarak; kelimelerin birlikte kullanılıyor olması benzer oldukları anlamında gelmez.**

   Örneğin; **"araba"** ile **"benzin"** birlikte sık kullanılan iki kelime olmalarına rağmen **benzer değildirler.**
   "araba", bir taşıt iken **"benzin"** bir yakıttır. Benzer olmalarını sağlayacak ortak nitelikleri yok denecek kadar azdır.
7. Verilen örneklere anket sırasında da erişebileceksiniz. Cevaplara emin olamamanız durumunda örnekleri incelemenizi tavsiye ederiz.

Geri İleri

BENZERLİK (1/25) İLİŞKİSELLİK (26/50)


0 1 2 3 4 5 6 7 8 9 10
Soru 16)
peynir - ipek
0 1 2 3 4 5 6 7 8 9 10
Soru 17)
turizm - seyahat
0 1 2 3 4 5 6 7 8 9 10
Soru 18)
tıknaz - uyumlu
0 1 2 3 4 5 6 7 8 9 10
Soru 19)
kemalci - atatürkist
0 1 2 3 4 5 6 7 8 9 10
Soru 20)
polis - yardım
0 1 2 3 4 5 6 7 8 9 10
Ekrandaki tüm sorular puanlandı.
Geri
İleri
BENZERLİK (1/25)
İLİŞKİSELLİK (26/50)

BÖLÜM 2: İLİŞKİSELLİK

1. Bu bölümde aynı kelime çiftlerini **ilişkisellik** derecesi bakımından değerlendirmeniz beklenmektedir.
2. İlişkisellik bir önceki bölümdeki benzerliğe oranla çok daha kolay belirlenebilmektedir.
3. Yüksek ilişkili kelimeler birbirleri ile alakalıdırlar ve sıklıkla benzer bağlamlar içinde kullanılırlar. Örneğin; **"benzin"** ile **"araba"** kelimeleri benzer bağlamlar içinde kullanıldıklarından oldukça ilişkilidirler.
4. Kelimelerin yüksek ilişkili olabilmesi için ortak özniteliklerinin olmasına ihtiyaç yoktur. Örneğin; **"kahve"** ve **"fincan"** çiftinde, biri içecek diğeri eşya olmasına rağmen çift oldukça ilişkilidir ve hatta birbirlerini hatırlatırlar. Bununla beraber; **"faiz"** ve **"fincan"** kelimelerinin ilişkileri oldukça azdır.
5. Benzer ve zıt anlamlı kelimelerin ilişki seviyeleri de genellikle yüksektir. Örneğin; *"İyi kötü ve çirkin."* cümlesinden anlaşılacağı gibi "iyi" ve "kötü" benzer bağlamlarda kullanılırlar. **"iyi"** ve **"kötü"** oldukça ilişkilidirler. Aynı şekilde; benzer anlamlı **"öğrenci"** ve **"talebe"** kelimeleri de benzer bağlamlarda sık geçerler ve oldukça ilişkilidirler.
6. Verilen örneklere anket sırasında da erişebileceksiniz. Cevaplara emin olamamanız durumunda örnekleri incelemenizi tavsiye ederiz.

Geri İleri

BENZERLİK (25/25) **İLİŞKİSELLİK (26/50)**

İLİŞKİSELLİK: Yüksek ilişkili kelimeler birbirleri ile alakalıdırlar ve sıklıkla bir arada kullanılır, birbirlerini hatırlatırlar. ☑ Otomatik olarak alttaki soruya geç

- "benzin" ve "araba" kelimeleri sıklıkla benzer bağlamlarda geçerler ve oldukça ilişkilidirler.
- "kahve" ve "fincan" kelimeleri birbirlerini hatırlatırlar, oldukça ilişkilidirler.
- "faiz" ve "fincan" kelimeleri oldukça ilişkisizdirler.
- "iyi" ve "kötü" gibi zıt anlamlı kelimeler oldukça ilişkilidirler.
- "öğrenci" ve "talebe" gibi yüksek benzerlikte kelimeler genellikle aynı bağlamlarda geçerler ve oldukça ilişkilidirler.

Soru 501) mızrap - barınak

0	1	2	3	4	5	6	7	8	9	10

Soru 502) kırmızı - gül

0	1	2	3	4	5	6	7	8	9	10

Soru 503) suçlu - şüphe

0	1	2	3	4	5	6	7	8	9	10

Soru 504) laikçiler - sekülerizmciler

0	1	2	3	4	5	6	7	8	9	10

Soru 505) bitki - zeytin

0	1	2	3	4	5	6	7	8	9	10

BENZERLİK (25/25) **İLİŞKİSELLİK (26/50)**

Anket başarıyla tamamlandı. Pencereyi kapatabilirsiniz.

Değerli vaktinizi ayırdığınız için sonsuz teşekkürler.

Sevgiler,

Geri

BENZERLİK (25/25) **İLİŞKİSELLİK (51/50)**