SimLex-999 for Dutch

Lizzy Brans♣ and Jelke Bloem♢♠

♣ Human Computer Interaction, Utrecht University
♢ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
l.brans@students.uu.nl, j.bloem@uva.nl

Abstract

Word embeddings revolutionised natural language processing by effectively representing words as dense vectors. Although many datasets exist to evaluate English embeddings, few cater to Dutch. We developed a Dutch variant of the SimLex-999 word similarity dataset by gathering similarity judgements from 235 native Dutch speakers. Subsequently, we evaluated two popular Dutch language models, Bertje and RobBERT, finding that Bertje showed superior alignment with human semantic similarity judgments compared to RobBERT. This study provides the first intrinsic Dutch word embedding evaluation dataset, which enables accurate assessment of these embeddings and fosters the development of effective Dutch language models.

1. Introduction

The demand for effective communication across languages grows as the world becomes more interconnected. Word embeddings, mathematical representations of words in multi-dimensional space, have become crucial for natural language processing tasks such as machine translation and sentiment analysis (Mikolov et al., 2013). However, research on word embeddings has been mainly focused on English. All recent state-of-the-art language models were initially developed for English, and various English-language benchmarks and datasets for evaluating such models are available.

In particular, the English SimLex-999 dataset (Hill et al., 2015) is a reliable gold standard of word similarity ratings that can be used to benchmark word embeddings and large language models incorporating vector representations of word meaning. This dataset consists of a balanced set of 999 English word pairs, where each pair was rated for similarity by 50 native speakers of English. In intrinsic evaluation of word embeddings, these human similarity ratings of word pairs are correlated with cosine similarity values of embedding vectors of the two words in the pair. A model that yields cosine similarities that better correlate with the gold standard is assumed to better represent semantic similarities between words.

However, this focus on English language technology leaves other languages such as Dutch with fewer options for evaluating the quality of its language technology. Dutch is typically considered a fairly high-resource language, and it has importance as a widely spoken language in the Netherlands, Belgium and Suriname and as an official language of the European Union. Nevertheless, Dutch remains under-explored in the context of word embeddings. In particular, we are unaware of any Dutch word similarity datasets that can be used to intrinsically evaluate word embeddings. Dutch word embedding models, such as the Word2Vec-based ones of Tulkens et al. (2016), have exclusively been evaluated on extrinsic tasks, such as relation identification. Available datasets for such extrinsic tasks have been summarized in the DUMB benchmark (de Vries et al., 2023). Intrinsic evaluation could provide a more general type of evaluation that can also inform the use of language models for extrinisic tasks for which there is no specific Dutch gold standard available.

In this work, we address this gap by constructing a new word similarity dataset for Dutch, inspired by the established English SimLex-999 dataset (Hill et al., 2015), as well as MEN (Bruni et al., 2014), following Resnik's (1995) methods for collecting human semantic similarity judgments. We then use the created dataset to evaluate two state-of-the-art Dutch language models, Bertje and RobBERT, against the human semantic similarity judgments in our dataset. Besides providing a new perspective on the potential performance of these models at various NLP tasks, this also provides insights into the disparities and alignments between machine-learned and human-perceived semantic associations. These observations might inspire further research into human language comprehension and developing more human-like NLP models (Landauer and Dumais, 1997; Lake et al., 2017).

2. Background

The distributional hypothesis posits that words with similar meanings often appear in similar contexts (Harris, 1954). According to this hypothesis, words can be represented as dense vectors in multi-dimensional space, with the distance between vectors indicating semantic similarity (Turney and Pantel, 2010). Word embeddings, rooted in early vector

LREC-COLING 2024, pages 14832–14845

space models (Salton et al., 1975), are vectorised representations that allow machines to process word meanings and relationships, with a notion of meaning that is grounded in the distributional hypothesis.

Techniques like Latent Semantic Analysis (LSA) and count-based models, aided by methods like Singular Value Decomposition (SVD), were forerunners to modern embeddings (Deerwester et al., 1990; Klema and Laub, 1980). Earlier word embedding models like Word2Vec and GloVe offer static representations for each word without accounting for context-based dynamic meanings (Mikolov et al., 2013; Pennington et al., 2014). In contrast, transformer models, such as BERT (Bidirectional Encoder Representations from Transformers) and Generative Pre-trained Transformer (GPT), shift towards token-based or contextual embeddings (Devlin et al., 2019; Radford et al., 2018). Based on the transformer architecture, these models offer context-sensitive representations, capturing the nuances of word usage (Vaswani et al., 2017; Camacho-Collados and Pilehvar, 2018).

2.1. Word embedding evaluation

As word embedding models are grounded in a theory of distributional learning, word embeddings have often been evaluated against datasets of word pairs with gold standard word similarity scores elicited from native speakers of the language. Early distributional semantic models for English were evaluated in this way against the WordSim353 (Finkelstein et al., 2001) dataset, where 13 to 16 participants rated 353 English word pairs. Subsequently, the MEN dataset (Bruni et al., 2014) was developed for English with similar aims but on a larger scale. This dataset includes 3000 word pairs rated by crowd workers, though with only a few raters per pair.

These datasets were later criticized for not distinguishing the concepts of word relatedness and word similarity in the instructions to the participants. To address this, the SimLex-999 dataset was developed (Hill et al., 2015), consisting of 999 English word pairs covering different parts of speech and balanced for more abstract and more concrete words. The pairs were rated by 500 crowd workers with about 50 ratings per word pair. The workers were specifically instructed to rate for similarity, and given examples thereof, and the annotator instructions were made available along with the dataset.

In the subsequent era of contextual embeddings, word similarity scores were no longer the most obvious choice for embedding evaluation, as there is no context for contextual embedding models to take advantage of in this kind of benchmark. Nevertheless, they continued to be used for evaluating static embeddings 'distilled' from contextual embeddings, mainly in interpretability work (Rogers et al., 2021) or in situations that called for representations of word types containing lexical-semantic information. An example of this is Abdou et al. (2021), who study representations of color terms to find out whether they reflect the perceptual color space. Among others, Bommasani et al. (2020) documented that earlier layers of contextual embedding models perform better at word similarity tasks, and their distillation approach was used to perform such tasks. However, Ehrmanntraut et al. (2021) also argued that we might as well use static embedding models especially in contexts where interpretability is important, as they still perform well in comparison to contextual embedding models on similarity tasks.

The aforementioned datasets are predominantly English-centric and may not encapsulate linguistic and cultural nuances of other languages, which could lead to biased or incomplete assessments of word embeddings (Faruqui et al., 2016; Agirre et al., 2009). To address this, and with SimLex-999 gaining widespread use, the dataset has been adapted to various other languages. This adaptation typically involved translating the word pairs from the English SimLex-999 to the target language and then asking native speakers of the target language to rate the pairs using a translated version of the instructions. We are aware of German, Italian, Russian (Leviant and Reichart, 2015), Estonian (Kittask and Barbu, 2019), Portugese (Querido et al., 2017), Vietnamese (Van Tan et al., 2017), Polish (Mykowiecka et al., 2018), Czech (Kliegr and Zamazal, 2018), Thai (Netisopakul et al., 2019) and Swedish (Hengchen and Tahmasebi, 2021) versions of SimLex-999. The subsequent MultiSimLex dataset (Vulić et al., 2020) covers 13 languages in a similar manner, based on crowdsourced ratings: Mandarin, Yue, Welsh, English, Estonian, Finnish, French, Hebrew, Polish, Russian, Spanish, Kiswahili and Arabic, with subsequent extensions to Icelandic (Daníelsson et al., 2021) and Buddhist Sanskrit (Lugli et al., 2022).

For some languages, translated versions of SimLex-999 are available but with no re-rating specific to that language, including Spanish (Etcheverry and Wonsever, 2016), Croatian (Mrkšić et al., 2017), Slovenian (Pollak et al., 2020) and various other languages (Barzegar et al., 2018). This approach may be less reliable when word pair similarity differs between languages (Leviant and Reichart, 2015).

For yet other languages, there are resources available that are similar in scope, but specific to that language, such as AnlamVer for Turkish, based on the SimLex-999 annotation guidelines (Ercan and Yıldız, 2018), Finnish FS300 (Venekoski and Vankka, 2017) based on a sub-

set of SimLex-999, the Japanese Word Similarity and Association Norm (Inohara and Utsumi, 2022), SimRelUz for Uzbek (Salaev et al., 2022) and USWS-21 for Urdu (Muneer et al., 2023).

We are not aware of any resources for Dutch in any of these three categories. One type of available resource for Dutch from which some semantic similarity gold standard may be derived is those used by psychologists to study associative priming, such as the dataset of Drieghe and Brysbaert (2002). However, this is about relatedness, not similarity, and this dataset is quite small in scope with 20 usable word pairs.

Therefore, we developed a Dutch version of SimLex-999, with re-rating by Dutch native speakers. Our dataset includes Dutch nouns, verbs, and adjectives, aligning with the English SimLex-999 to ensure linguistic coherence and enable cross-language comparisons.

3. Methodology

To build a Dutch version of SimLex-999, we first translated the word pairs to Dutch. To ensure consistency and comparability with the original SimLex-999, corresponding nouns, verbs, and adjectives were used (Hill et al., 2015). The selection process involved identifying appropriate Dutch translations or equivalents for the English words in these datasets and carefully considering differences in meaning, usage, and cultural context. When multiple Dutch translations were available, the German version of the SimLex-999 dataset was used as a reference point, given the closer linguistic relationship between Dutch and German (Leviant and Reichart, 2015).

Table 1 showcases this translation process, displaying English word pairs from SimLex-999 alongside their German and Dutch counterparts.

English	German	Dutch
meat - sandwich	fleisch - sandwich	vlees - broodje
meter - yard	meter - yard	meter - decimeter
dumb - dense	blöd - unterbelichtet	dom - onderontwikkeld

Table 1: Examples SimLex-999 translation

These examples underline the careful considerations made during translation, focusing on achieving semantic accuracy and cultural relevance. For example, the word *yard* in its distance sense is not used much in Dutch as it is part of the British Imperial system, while in Dutch-speaking countries, the metric system is used. Therefore, we used the word *decimeter* for the Dutch dataset. Although a yard is much longer than a decimeter, we cannot translate it to *meter* as this would yield the pair *meter-meter*, so we chose another related unit of measure that is reasonably frequent in Dutch, the decimeter. In other cases, we chose a word of similar frequency if an exact translation was not available. In this manner, the selection process aimed for the closest possible alignment between the original English words and their Dutch counterparts, thus enhancing the validity and applicability of the Dutch version of the SimLex-999 dataset.

3.1. Re-rating

Next, we conducted a questionnaire in which the Dutch word pairs were re-rated. Native Dutch speakers aged 16 or older participated in the study. Similar to the original English SimLex-999 study, the aim was to ensure a certain level of language proficiency. The chosen age criterion ensured that participants had completed a significant portion of their education in Dutch. Furthermore, this broad age range was required by our ethics committee, based on EU GDPR regulations, and more detailed demographic data collection would have been considered personal data, complicating the data collection procedure and its shareability. In addition, to minimise judgment errors, individuals with any language disorders were excluded from the study. Several recruitment channels were utilised, including online forums, social media, and educational institutions, which targeted a wide range of potential participants from different walks of life.

The data was collected through an online questionnaire hosted on the Qualtrics platform, chosen for its user-friendly interface and efficient distribution to a large audience (Qualtrics, 2023). Before participating, all participants read an information brochure and provided informed consent. A unique personal ID number was provided to each participant, allowing them to withdraw their answers at any time, ensuring their anonymity. The study was approved by the corresponding author's faculty ethics committee.

After providing informed consent, participants answered personal questions to ensure they met the eligibility criteria. These questions assessed whether participants met the age requirement, their fluency in Dutch, and the absence of language disorders. After determining eligibility, the study directed participants to an instructions page with a Dutch translation of the original English instructions from the SimLex-999 dataset (Hill et al., 2015). These instructions are shown in Appendix B. The instructions emphasise that participants should rate word pairs based on similarity rather than relatedness and provide clear examples to guide them.

Participants rated the semantic similarity of word pairs in the new dataset using a continuous sliding scale from 0 to 10 (Resnik, 1995; Majewska et al., 2021). This random assignment aimed to minimise potential biases related to the order in which word pairs were presented. Aggregating these ratings,

the mean of the scores was used for each word pair, and standard deviations are also included in the dataset as a measure of variance. To ensure the quality of our dataset, we conducted manual checks of all annotations. Annotators who completed only a small portion of the questionnaire were specifically reviewed for accuracy. This process helped us maintain the integrity and reliability of our dataset. These steps follow Hill et al. (2015). The aggregated scores serve as the word similarity values in our dataset, which we subsequently use for evaluating Dutch word embeddings.

The new dataset's reliability and validity were assessed using several statistical methods. Inter-rater reliability was examined using the intraclass correlation coefficient (ICC, Koo and Li, 2016), and at least ten responses per word pair were gathered to reinforce reliability. We examine the validity of our dataset in terms of content, construct and criterion validity (Cronbach and Meehl, 1955). These concepts are not often cited in NLP but are commonly used in quantitative research, particularly in the social sciences, to validate metrics and procedures. Content validity is defined as "the extent to which a test reflects the various components of the construct it is intended to measure", construct validity is "the extent to which a measure accurately assesses the construct or latent attribute that it is intended to measure", and criterion validity is "evaluating the validity of a measure based on its relationship to a specific criterion" (APA, 2020). We believe this is also a useful way of describing and comparing the validity of NLP benchmarks beyond specifics of the particular evaluation procedure.

For Dutch SimLex-999, content validity is supported by the fact that the selected words covered a broad range of semantic domains, and include nouns, verbs, and adjectives. To assess construct validity, we examine the relationship between the dataset and the original English SimLex-999 dataset which is more established as a benchmark of semantic association. To address criterion validity, we compare the dataset's performance in evaluating the embeddings contained in Dutch language models with previous evaluation attempts that use other evaluation metrics.

3.2. Intrinsic evaluation of Dutch language models

In addition to creating this dataset, we use our new Dutch SimLex-999 dataset to intrinsically evaluate two state-of-the-art Dutch language models, Bertje and RobBERT. Bertje is a Dutch-language version of Google AI's BERT (Devlin et al., 2019), having been trained on a vast array of Dutch internet text (De Vries et al., 2019). RobBERT, mirroring Facebook AI's RoBERTa model, benefits from advanced training approaches and a rich Dutch dataset from the OSCAR corpus (Delobelle et al., 2020; Liu et al., 2019; Scheible et al., 2020). While SimLex-999 is typically used to evaluate static word embeddings, we chose to evaluate embeddings distilled from contextual embedding models as this is a recently popular use case (cf. section 3.2).

The evaluation procedure involves correlating model-calculated similarity scores with the averaged human judgments from the datset using Spearman's rank correlation (Spearman, 1904). Similarity scores were obtained by separately embedding each word from a word pair without any context, and computing the cosine similarity between the resulting embeddings. We perform this procedure for every transformer layer of each model to test which layer best encodes human word similarity judgements. We expect that layer 0 yields the best correlation with human word similarity ratings, as it corresponds to a contextless static embedding and the human similarity ratings were also elicited without context. Lastly, we identified instances where model outcomes significantly deviated from human ratings by qualitative analysis.

As BERT's WordPiece tokenizer (Schuster and Nakajima, 2012) and RoBERTa's Byte-Pair Encoding (BPE) tokenizer (Liu et al., 2019) approaches may segment longer or less frequent words into tokens, we also consider the issue of subtokenized words. When intrinsically evaluating a model, it is possible to either construct embeddings for them by combining the embeddings of their subtokens, or to skip word pairs with subtokenized words as we would do with OOV words for static embedding models. We explore the effects of these options.

In the former condition, we apply an averaging method to the subtoken embeddings, in line with practices for models using the WordPiece and BPE tokenisers (Devlin et al., 2019; Liu et al., 2019). This is called subword pooling (Bommasani et al., 2020). This approach typically leads to better performance on low-frequency words.

4. Results

From the initial 250 participants, 235 native Dutch speakers were selected as raters. Fifteen were excluded due to self-reported language disorders, such as dyslexia, potentially impacting their Dutch similarity ratings. Data was collected using the Qualtrics platform (Qualtrics, 2023). Participants were given a unique set of 100 word pairs to evaluate. Some participants did not complete the questionnaire, resulting in 10 to 19 responses per word pair, with an average of 15 responses. For each pair, the mean rating was taken to determine the final gold standard SimLex-999 score.

We measured reliability using the intraclass cor-

relation coefficient (ICC, Koo and Li, 2016), for which we divided annotators into random groups. The average rater ICCs displayed excellent consistency across all groups, with values between 0.84 and 0.96. Single-rater ICCs, indicative of individual rating consistency, ranged from fair (0.26) to good (0.59) across the groups. These values confirm the dataset's reliability, signifying stability in similarity ratings and minimising the potential impact of individual judgment discrepancies or biases.

The dataset's validity was evaluated in three dimensions: content, construct, and criterion (Cronbach and Meehl, 1955). Following Hill et al. (2015) the word pairs cover a range of semantic domains, ensuring content validity. Criterion validity was assessed by comparing the dataset's performance in evaluating Dutch word embeddings to past evaluations, which is detailed in section 5.1. Our construct validity metric is the correlation between ratings in our dataset and ratings in SimLex-999 for the English and German versions of SimLex-999.

This type of correlation analysis has also been performed to assess cross-language similarity in MultiSimLex (Vulić et al., 2020). Spearman's rank correlation coefficient (Spearman, 1904) served as the metric to measure the degree of semantic similarity between different language versions of the SimLex-999 dataset. High correlation coefficients signify a strong agreement in the semantic similarity between two languages.

Dataset Comparison	Correlation
English - German	0.7478
English - Dutch	0.7487
German - Dutch	0.6785

Table 2: Cross-Language Similarity Agreement

Table 2 summarises the Cross-Language Similarity Agreement findings. The results validate the Dutch SimLex-999 dataset's similarity to the English and German versions, serving as evidence of construct validity. Of course, we do not expect maximum correlation, as word meanings differ between languages even though we translated the word pairs as closely as possible. However, since these languages are all related West Germanic languages, a high correlation is to be expected of a valid metric. The English-German comparison yielded a Spearman correlation of 0.7478, aligning with prior findings and suggesting a substantial semantic similarity between these two languages. A comparison between English and Dutch produced a Spearman correlation of 0.7487, similar to the English-German correlation, signifying a high degree of semantic similarity. Furthermore, a comparison between German and Dutch resulted in a Spearman correlation of 0.6785. This is surprising as we might expect German and Dutch to be closer together. The correlation values might have different statistical power due to differences in the number of participants (approximately 50 per pair for English, 13 per pair for German, 10-19 per pair for Dutch).

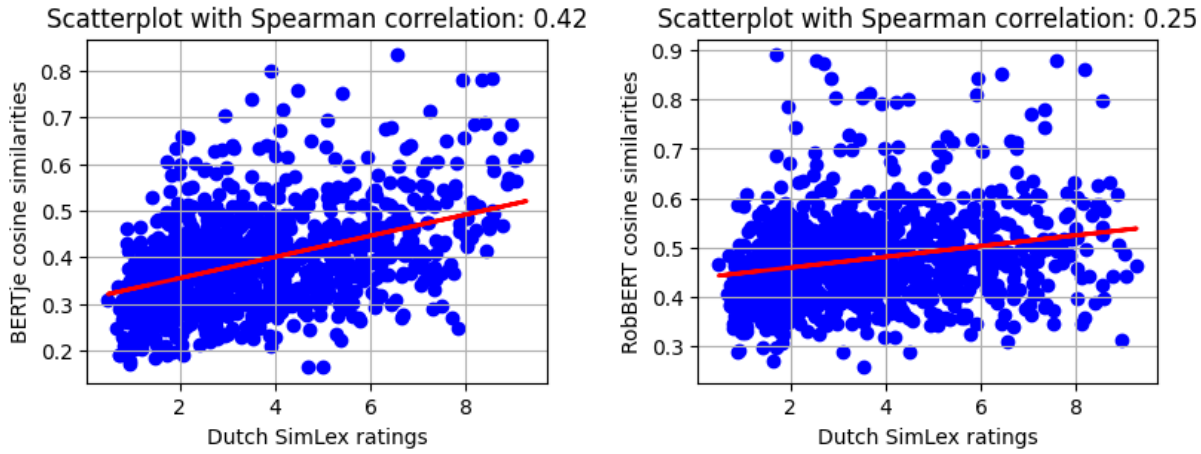
4.1. Comparison of Dutch language models

Next, we evaluate two widely used transformer-based Dutch language models, Bertje and RobBERT-v2, against the novel SimLex-999 Dutch dataset. Using Spearman correlation coefficients, we assess the statistical relation between the model's predictions and human-rated similarity.

Firstly, to be able to use Spearman correlation, the relationship between human-rated and model similarities must be monotonic. We observe that this is the case for both models from the scatterplots in Figure 1. Figures 2 and 3 show the correlation values between the human ratings of word pair similarity and cosine similarities of word pairs in all layers of the models. For Bertje, the highest Spearman correlation is observed in Layer 0, with a correlation coefficient of 0.409. When combining layers 0 and 3, a correlation of 0.421 is reached. Correlations with human judgements mostly decrease across the layers down to 0.213 at layer 11. Adjectives display the highest correlation at 0.474, with nouns at 0.449 and verbs significantly lower at 0.213 at layer 0. Using the SUBTLEX-NL word frequency database (Keuleers et al., 2010), high-frequency words achieve a 0.444 correlation at layer 0, while low-frequency words hit 0.420. In higher layers, low-frequency words have higher correlations than high-frequency words.

RobBERT's highest overall correlation is 0.207 in Layer 0 and 0.247 when combining Layers 0 and 5. For this model, nouns are the part-of-speech with the highest correlation at 0.263. At layer 0, low-frequency words correlate better with the human judgements than high-frequency words (0.284 and 0.194 respectively).

The trends observed in Bertje's performance align with the expected patterns of Transformer-based models. Early layers usually adequately represent semantic relationships, while later layers focus more on syntactic patterns (Zhang et al., 2018). Layer 0, which corresponds to static word embeddings, yields the highest correlation scores in both models. Combining high-scoring early layers slightly increases correlation with human judgements. Overall, RobBERT-v2 shows lower correlation scores than Bertje.



(a) Bertje-DutchSimLex scatterplot (layers 0+3)

(b) RobBERT-DutchSimLex scatterplot (layers 0+5)

Figure 1: Scatterplots with predicted model similarities and similarity ratings from Dutch SimLex-999.

R	Word 1	Word 2	SimL	Cos
1	rondzwerven *roam*	dwalen *wander*	7.83	0.247
2	competentie *competence*	vermogen *ability*	7.73	0.271
3	film *movie*	filmrol *film*	3.92	0.798
4	zelfverzekerd *confident*	zeker *sure*	7.49	0.295
5	elastisch *elastic*	flexibel *flexible*	7.32	0.289
...	...	...	...	
995	botten *bone*	enkel *ankle*	1.93	0.292
996	oud *old*	vers *fresh*	1.69	0.278
997	rijkdom *wealth*	bekendheid *fame*	2.46	0.327
998	hart *heart*	operatie *surgery*	1.80	0.286
999	onderzoeken *investigate*	herzien *examine*	3.24	0.379

(a) Largest and smallest differences for Bertje.

Word 1	Word 2	SimL	Cos
aanmoedigen *encourage*	ontmoedigen *discourage*	1.71	0.891
vlug *quick*	snel *rapid*	8.95	0.313
afwezigheid *absence*	aanwezigheid *presence*	2.54	0.878
desorganiseren *disorganise*	organiseren *organise*	2.68	0.873
aandacht *attention*	interesse *interest*	8.13	0.347
...	...	...	...
heldin *heroine*	held *hero*	8.55	0.797
klein *tiny*	enorm *huge*	1.82	0.372
arm *arm*	ader *vein*	2.05	0.389
hout *wood*	papier *paper*	3.55	0.481
bad *bath*	vrouw *wife*	1.61	0.358

(b) Largest and smallest differences for RobBERT.

Table 3: The words with the highest and lowest absolute difference in human similarity score and model cosine distance (scaled to the SimLex-999 1-10 rating scale, but only original values shown). Highest differences, i.e. errors, are at the top. The words printed in italics are the equivalent words from the English SimLex-999, they are not necessarily the most obvious translation of the Dutch word.

4.2. Error analysis

To gain some qualitative insights into these results, we compare for both models the top 5 worst predicted similarity scores and the top 5 best predicted similarity scores. These results are shown in tables 3a for Bertje and 3b for RobBERT. Word pairs from our Dutch SimLex-999 are shown with their equivalent words from the English SimLex-999. As SimLex similarity ratings are on a scale of 1 to 10 and cosine similarities are on a scale from 0 to 1, we scaled the cosine similarities to a scale of 1 to 10 and then computed the absolute difference between the scaled cosine similarities from the models and the human ratings. The table is sorted by absolute difference, though this value is not shown for reasons of space.

We observe a number of characteristic errors at the top of the table. For both

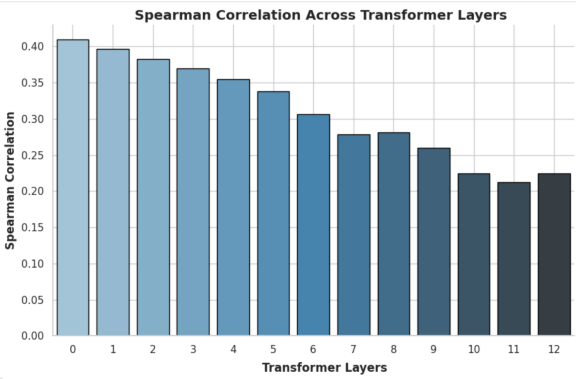


Figure 2: Spearman Correlation Across Transformer Layers in Bertje.

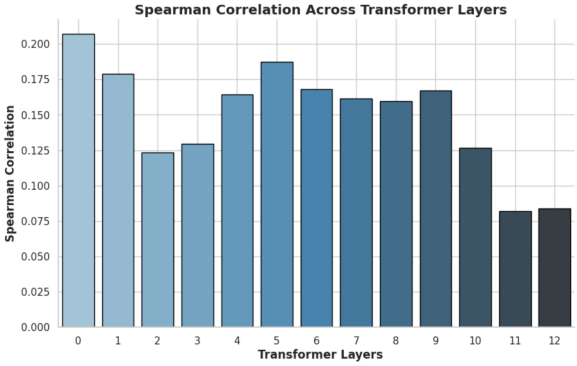


Figure 3: Spearman Correlation Across Transformer Layers in RobBERT.

models, low-frequency words appear at the top of the error list. In particular, RobBERT struggles with predicting antonym relations (*aanmoedigen-ontmoedigen*, 'encourage-discourage', *aanwezigheid-afwezigheid*, 'absence-presence'). This suggests that RobBERT encodes relatedness rather than semantic similarity. This is a common problem in distributional semantic modeling and an important motivation for the development of the original SimLex-999 benchmark.

Conversely, Bertje appears to struggle more with ambiguous words, especially ones that are rated similar by humans. In the pair that is ranked 4, the word *zeker* 'sure' is also often used as a confirmatory statement in conversations: 'for sure'. In pair 5, the word *flexibel* 'flexible' has many metaphoric extensions that may be common in internet data. In pair 3, the word *filmrol* can either mean a role in a movie or a roll of film. In pair 2, the word *vermogen* 'ability' has extended meanings related to finance and energy. An interesting succesful example for Bertje is pair 998, *hart-operatie* 'heart-surgery', where we might expect interference from the association between these two words, but Bertje correctly represents them as semantically unrelated.

Both models yield better correlations with unrelated terms (scored low by the raters), though this could be an artifact of our scaling procedure.

4.3. Effect of the tokenizer

To obtain these similarity scores from Bertje and RobBERT, we made the common assumption that when a word is not in the model's vocabulary, its embedding can be reconstructed by averaging the embeddings of its subtokens, as discussed in section 3.2. However, this assumption may not always hold and this process may affect model performance. We observed a large difference in the number of subtokenized words for both models. Subtokenized words are in some sense out-of-vocabulary (OOV), although this is not quite comparable to OOV words in static embedding models. Contextual embedding models are designed to be used with tokenizers that split infrequent words into subtokens, thus learning a vocabulary that consists of whole-word tokens and subtokens. In this analysis, we make a distinction between such subtokenized words, and words that are in the model's vocabulary as a whole.

We observe that RobBERT's BPE-based vocabulary contains far fewer of the Dutch SimLex-999 words than Bertje's WordPiece-based vocabulary. BPE builds the vocabulary by considering all symbols used to write words, which results in a higher number of subtokenized words when compared to other tokenisation methods (Sennrich et al., 2016).

Of the 999 word pairs, 139 contained words that are subtokenized for Bertje. Some examples included *molecuul* 'molecule', *volbrengen* 'to accomplish', *rechtvaardigheid* 'justice' and *afwijken* 'to deviate'. For the RobBERT model, 550 words are subtokenized, with examples like *erkennen* 'to acknowledge', *secretaresse* 'secretary', *somber* 'dreary' and *aanzien* 'prestige'. This far larger reliance on subtokens for Dutch is a potential explanation for RobBERT's lower correlation with human similarity judgements. To investigate this possibility, we also compute the correlations over the in-vocabulary words only, excluding all word pairs containing at least one subtokenized word. These results are shown in figures 4 and 5. Bertje achieved lower correlation scores in this condition. Bertje's highest Spearman correlation reached 0.280 in Layer 0 and 0.276 when combining Layers 0 and 3 in this condition. Conversely, RobBERT's results for subtokenized words are better, reaching 0.461 at layer 0. This is a higher correlation than Bertje's overall result, though it drops off faster in subsequent layers.

This indicates that Bertje benefits from the use of subtokens to handle complex words, while the RobBERT BPE tokenizer might not be optimal for Dutch, even though we used RobBERT v2, which uses a Dutch-specific tokenizer. This appears to be the main reason for RobBERT's performance

deficit compared to Bertje on this benchmark. This result also shows that the approach of combining subtoken embeddings is essential for good performance also for Dutch, as it is for a higher-resource language like English (Hu et al., 2019). However, the ideal strategy for composing token embeddings from subtoken embeddings (e.g. averaging, summing or otherwise) may vary depending on the specifics of the task and dataset at hand.

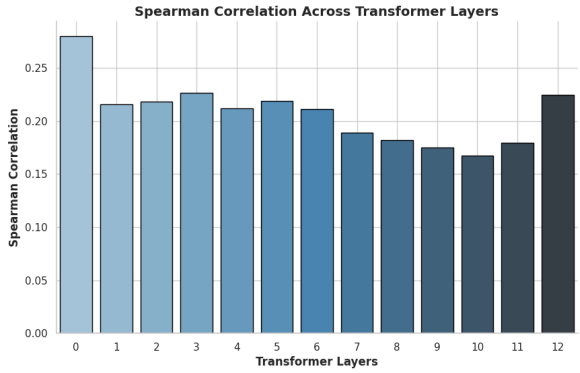


Figure 4: Spearman Correlation across transformer layers in Bertje without subtokens.

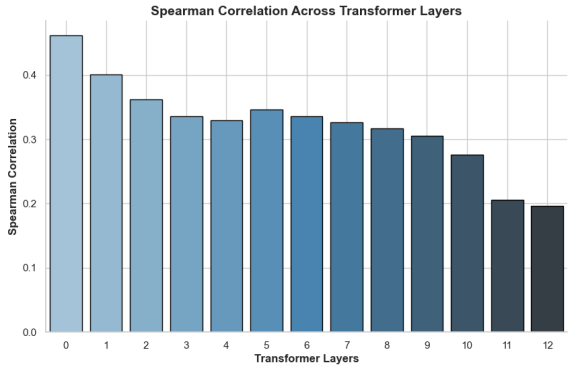


Figure 5: Spearman Correlation across transformer layers in RobBERT without subtokens.

The data indicates that Bertje generally outperforms RobBERT on the SimLex-999 dataset. Our error analysis shows that RobBERT appears to have more difficulty with distinguishing word similarity from word relatedness, and Bertje appears to have more difficulty with semantically highly ambiguous words. Differences in performance can be attributed to model architectures and training datasets, highlighting how these factors influence the resulting semantic representations.

5. Discussion

In this work, we have introduced the Dutch SimLex-999 resource, which should put Dutch language technology on a more equal playing field with that of higher-resource languages. Additionally, we provided insights into the performance of two commonly used Dutch language models. Nevertheless, our findings leave several points open to discussion which may have implications for the generalisability and interpretation of the findings.

5.1. Comparison to other studies

Study	Model	Corr.
Leviant & Reichart (2015)	W2V (EN)	0.266
Leviant & Reichart (2015)	W2V (DE)	0.354
Leviant & Reichart (2015)	W2V (IT)	0.308
Leviant & Reichart (2015)	W2V (RU)	0.260
Chronis & Erk (2020)	BERT (L8)	0.608
Ehrmanntraut et al. (2021)	BERT	0.476
Shahmohammadi et al. (2021)	GloVe	0.408
de Vos et al. (2022)	BERT	0.384
This Study	Bertje	0.421
This Study	RobBERT	0.247

Table 4: Comparison of SimLex-999 Spearman correlations across different studies

To better contextualize our results, it is informative to examine the results of similar experiments for other languages. The performance of the Dutch Bertje and RobBERT on the Dutch SimLex-999 dataset is compared to the performance of other models on their respective language's version of SimLex-999 in Table 4.

In this landscape, Bertje's peak Spearman correlation of 0.421 (with layer combination) surpasses many of these results, except for the correlation reported by the two English BERT models (Chronis and Erk, 2020; Ehrmanntraut et al., 2021). On the other hand, RobBERT's highest correlation of 0.247 is notably lower, especially when compared to other BERT-based models.

We can observe that the highest correlation with human similarity judgements was obtained by taking embeddings from the 8th layer of English BERT, though this study of Chronis and Erk (2020) was specialized for lexical-semantic tasks by using centroids of clusters of token embeddings. The additional contextual information this provides likely accounts for the higher correlation, and this would be an interesting direction to explore for Dutch for improving results on the task of semantic relationship estimation. In our setup, without any context, layer 0 tended to yield the highest correlations. However, combining layer 0 with another relatively highly correlating layer (layers 3 or 5 for our models) resulted in somewhat higher correlations than using a single layer.

More broadly, our Dutch SimLex-999 dataset enables other cross-linguistic model comparisons involving Dutch to be made in future work. Such

comparisons can discern language-specific characteristics from model-specific influences, thereby advancing multilingual NLP.

5.2. Implications for extrinsic tasks

Previous evaluations involving the English SimLex-999 dataset have raised concerns about focusing solely on the "interpretability" of word embeddings (Gladkova and Drozd, 2016). This intrinsic evaluation approach might not leverage the potential of distributional semantics fully. In response, combining intrinsic evaluations with extrinsic evaluations is recommended to understand the models' real-world applicability. These concerns equally apply to the Dutch version of the dataset. For Bertje and RobBERT, it is also evident that model performance can vary based on the extrinsic tasks they are used for (De Vries et al., 2019; Delobelle et al., 2020). Intrinsic evaluation results do not always correlate with extrinsic evaluation results.

Furthermore, differences in tokenisation methods between Bertje and RobBERT influenced the subtokenized word count, and the model that performed worse had far more subtokenized items and poorer performance on those items. This invites the question of what the ideal tokenisation approach for the Dutch language is.

This research primarily centred on the intrinsic evaluation of Dutch language models, a methodology focused on assessing the models' ability to accurately capture semantic relationships between words (Bellegarda, 2000). Integrating extrinsic evaluations usually provides deeper insights when evaluating for a specific task (Foster et al., 2014; Khurana et al., 2023), and may yield different results. For instance, in an extrinsic evaluation of Dutch emotion detection tasks, RobBERT outperformed Bertje (De Bruyne et al., 2021). Moreover, in another task-focused evaluation, Bouma (2021) probed Dutch language models' ability to predict the appropriate use of relative pronouns in complex sentences. While Bertje performed best in the masked language modelling probing task, RobBERT significantly improved fine-tuning, particularly highlighting the model's capacity to adapt and learn from task-specific data. These differential performances indicate the need for the availability and use of multiple evaluation methods.

5.3. Further work

Fine-tuning Bertje and RobBERT on the Dutch SimLex-999 dataset might enhance their performance on semantic similarity tasks, as seen in English models (Shi et al., 2023; Ding et al., 2023). This could lead to improved performance on downstream tasks. Additionally, employing diverse evaluation sets can offer a well-rounded view of a model's capabilities (Alivanistos et al., 2022; Xu et al., 2023).

The evaluation of a broader range of models on Dutch SimLex-999 would provide a better picture of the state of Dutch language technology, including static word embedding models or embeddings from multilingual models such as mBERT. The potential of generative large language models such as the LLAMA and GPT families of models for Dutch remain largely unexplored. Assuming that lexical-semantic representations can be obtained, evaluating such models using the Dutch SimLex-999 dataset could provide insights into their Dutch semantic capabilities, potentially benefiting numerous downstream applications. Dutch SimLex-999 could also be used as a benchmark for intrinsically evaluating multimodal embeddings, as done by Pezzelle et al. (2021) for English, comparing the similarities between vector representations of images to human similarity judgements.

The tokenizers employed by Bertje and RobBERT were mainly designed with English in mind. Using different tokenisation methods, like morphological or character-level tokenisers, might provide better results for the more complex morphology of Dutch (Kettunen, 2014). Such tokenisers could offer meaningful tokens for complex Dutch words or capture nuances of the language more effectively.

6. Conclusion

By developing a Dutch version of the SimLex-999 dataset, our work opens up the possibility to carry out intrinsic evaluations of Dutch word embeddings for the first time. Our rating procedure showed high agreement between the raters providing semantic similarity judgements, and the dataset shows significant overlap in semantic similarities with the English and German versions (Hill et al., 2015; Leviant and Reichart, 2015).

We used this dataset to evaluate two prominent Dutch language models, Bertje and RobBERT (De Vries et al., 2019; Delobelle et al., 2020), examining performance per layer, part-of-speech and by lexical frequency. These findings assist in model selection, emphasizing each model's specific strengths and limitations. A qualitative analysis of model performance on specific word pairs from the dataset with specific linguistic properties can help to better understand the strengths and weaknesses of each model.

This dataset advances Dutch natural language processing by offering a broadly applicable benchmark for word embedding quality, and we encourage the community to intrinsically evaluate other Dutch language models using this benchmark.

7. Ethical considerations and limitations

Despite rating the Dutch SimLex-999 dataset with 235 native Dutch speakers, the dataset's broad applicability might be limited due to the demographic characteristics of the participants. Since collection of participants' personal information was restricted by privacy concerns, providing detailed demographic data was impossible. In particular, it is possible that the intuitions of speakers from smaller populations where standard Dutch is used, such as Belgian Dutch and Surinamese Dutch speakers, are less well represented in the dataset.

The translation process from English to Dutch might have brought in biases, and having been done by only 2 experts, lacked diverse representation. While effort was made to maintain meaning, linguistic subtleties and cultural differences between the translators and raters could have influenced the dataset.

Although MultiSimLex a multilingual followup to SimLex-999 exists (Vulić et al., 2020), we focused on adapting the original dataset to Dutch. As the scope of MultiSimLex is larger, it would have required more resources to adapt even for a single language. However, this restricts the comparability of our results to evaluations done using MultiSimLex.

Our evaluation was limited to contextual word embedding models, while most previous benchmarking using the English SimLex-999 involved static word embeddings. Of course, the Dutch SimLex-999 can be used to evaluate static embeddings as well in future work.

As is the case for English, there is no guarantee that higher scores on an intrinsic evaluation benchmark yield higher scores on an extrinsic task (Gladkova and Drozd, 2016). Furthermore, it is important to note that no single metric can summarize all aspects of model quality (Valmeekam et al., 2024). Semantic similarity-based intrinsic evaluation does not tell us anything about the presence of harmful biases in a model, for example. Models are best evaluated on various different tasks.

8. Bibliographical References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132.

Eneko Agirre, Enrique Alfonseca, Keith B 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: Conference of the North American Chapter of the Association of Computational Linguistics, Companion Volume: Short Papers*, pages 19–21.

Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as probing: Using language models for knowledge base construction. In *2022 Semantic Web Challenge on Knowledge Base Construction from Pre-Trained Language Models, LM-KBC 2022*, pages 11–34. CEUR-WS. org.

APA. 2020. APA PsycTests methodology field values.

Siamak Barzegar, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and André Freitas. 2018. Semr-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

J.R. Bellegarda. 2000. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Gosse Bouma. 2021. Probing for Dutch relative pronoun choice. *Computational Linguistics in the Netherlands Journal*, 11:59–70.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Online. Association for Computational Linguistics.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Hjalti Daníelsson, Steinunn Rut Friðriksdóttir, and Steinþór Steingrímsson. 2021. Icelandic multi-SimLex (21.06). CLARIN-IS.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Emotional RobBERT and insensitive BERTje: combining transformers and affect lexica for dutch emotion detection. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 257–263. Association for Computational Linguistics.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2023. DUMB: A benchmark for smart evaluation of Dutch models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7221–7241.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard A Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Denis Drieghe and Marc Brysbaert. 2002. Strategic effects in associative priming with words, homophones, and pseudohomophones. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(5):951.

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. Type-and token-based word embeddings in the digital humanities. In *CHR*, pages 16–38.

Gökhan Ercan and Olcay Taner Yıldız. 2018. Anlamver: Semantic model evaluation dataset for Turkish word similarity and relatedness. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3819–3836.

Mathias Etcheverry and Dina Wonsever. 2016. Spanish word vectors from Wikipedia. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3681–3685.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A Smith. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–8.

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.

Mary Ellen Foster, Manuel Giuliani, and Amy Isard. 2014. Task-based evaluation of context-sensitive referring expressions in human–robot dialogue. *Language, Cognition and Neuroscience*, 29(8):1018–1034.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

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

Simon Hengchen and Nina Tahmasebi. 2021. SuperSim: a test set for word similarity and relatedness in Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 268–275.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning

for out-of-vocabulary words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4102–4112, Florence, Italy. Association for Computational Linguistics.

Keisuke Inohara and Akira Utsumi. 2022. JWSAN: Japanese word similarity and association norm. *Language Resources and Evaluation*, pages 1–29.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Emmanuel Keuleers, Marc Brysbaert, and Boris New. 2010. Subtlex-nl: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, 42(3):643–650.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Claudia Kittask and Eduard Barbu. 2019. Is similarity visually grounded? Computational model of similarity for the Estonian language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 541–549, Varna, Bulgaria. INCOMA Ltd.

Virginia Klema and Alan Laub. 1980. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176.

Tomáš Kliegr and Ondřej Zamazal. 2018. Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353. *Data & Knowledge Engineering*, 115:174–193.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach.

Ligeia Lugli, Matej Martinc, Andraž Pelicon, and Senja Pollak. 2022. Embeddings models for Buddhist Sanskrit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3861–3871.

Olga Majewska, Diana McCarthy, Jasper JF van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2021. Semantic data set construction from human clustering and spatial arrangement. *Computational Linguistics*, 47(1):69–116.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Iqra Muneer, Ghazeefa Fatima, Muhammad Salman Khan, Rao Muhammad Adeel Nawab, and Ali Saeed. 2023. Developing a large benchmark corpus for Urdu semantic word similarity. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–19.

Agnieszka Mykowiecka, Malgorzata Marciniak, and Piotr Rychlik. 2018. Simlex-999 for Polish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ponrudee Netisopakul, Gerhard Wohlgenannt, and Aleksei Pulich. 2019. Word similarity datasets for Thai: Construction and evaluation. *IEEE Access*, 7:142907–142915.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Senja Pollak, Ivan Vulić, Andraž Pelicon, Andraž Repar, Carlos Armendariz, Purver Matthew, and Nikola Ljubešić. 2020. Simlex-999 Slovenian translation SimLex-999-sl 1.0.

Qualtrics. 2023. Qualtrics survey software. Accessed: 2023-05-24.

Andreia Querido, Rita Carvalho, João Rodrigues, Marcos Garcia, João Silva, Catarina Correia, Nuno Rendeiro, Rita Valadas Pereira, Marisa Campos, and António Branco. 2017. LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of Portuguese. *Revista da Associação Portuguesa de Linguística*, (3):265–283.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, pages 448–453.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ulugbek Salaev, Elmurod Kuriyozov, and Carlos Gómez-Rodríguez. 2022. SimRelUz: Similarity and relatedness scores as a semantic evaluation dataset for Uzbek language. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 199–206.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. GottBERT: a pure German language model.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yiwen Shi, Jing Wang, Ping Ren, Taha ValizadehAslani, Yi Zhang, Meng Hu, and Hualou Liang. 2023. Fine-tuning BERT for automatic ADME semantic labeling in FDA drug labeling to enhance product-specific guidance assessment. *Journal of Biomedical Informatics*, page 104285.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4130–4136.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2024. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36.

Bui Van Tan, Nguyen Phuong Thai, and Pham Van Lam. 2017. Construction of a word similarity dataset and evaluation of word similarity techniques for Vietnamese. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 65–70. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Viljami Venekoski and Jouko Vankka. 2017. Finnish resources for evaluating language model semantics. In *Proceedings of the 21st Nordic conference on computational linguistics*, pages 231–236.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic

similarity. *Computational Linguistics*, 46(4):847–897.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

A. Code and dataset

The following link provides access to the repository containing the Dutch SimLex-999 dataset, as well as the code used for evaluating Bertje and RobBERT: https://github.com/lizzybrans/Simlex999-Dutch

B. Annotator instructions

Instructies

Twee woorden zijn synoniemen als ze zeer vergelijkbare betekenissen hebben. Synoniemen vertegenwoordigen hetzelfde type of dezelfde categorie van dingen.

Hier zijn enkele voorbeelden van synoniemenparen:

- *kop / mok*
- *buurt / wijk*
- *jaloezie / afgunst*

In de praktijk kunnen woordparen die niet exact synoniem zijn, toch zeer vergelijkbaar zijn.
Hier zijn enkele bijna synoniemen:

- *alligator / krokodil*
- *liefde / genegenheid*
- *kikker / pad*

In tegenstelling hiermee zijn de volgende woordparen wel gerelateerd, maar niet erg vergelijkbaar.
De woorden vertegenwoordigen totaal verschillende soorten dingen:

- *auto / band*
- *auto / snelweg*
- *auto / ongeluk*

In deze enquête word je gevraagd om woordparen te vergelijken en te beoordelen hoe vergelijkbaar ze zijn door middel van een schuifregelaar die varieert van 1 tot 10. Onthoud dat het hier gaat om gelijkenis en niet om gerelateerdheid.

Het is belangrijk om te proberen de enquête zo snel en efficiënt mogelijk in te vullen. Probeer daarom zo snel mogelijk te werken en vertrouw op je intuïtie of onderbuikgevoel als Nederlandstalige.

Er is geen juist antwoord op deze vragen. Het is volkomen redelijk om je intuïtie te gebruiken, vooral wanneer je wordt gevraagd om woordparen te beoordelen die je niet vergelijkbaar vindt.

Figure 6: Instructions Dutch SimLex-999

The annotation instructions were translated from (Hill et al., 2015). The instructions emphasised that participants should rate word pairs based on similarity rather than relatedness and provided clear examples to guide them.