Estimating Word Concreteness from Contextualized Embeddings

Christian Wartena
Hochschule Hannover
Expo Plaza 12
30539 Hannover, Germany
christian.wartena@hs-hannover.de

Abstract

Concreteness is a property of words that has recently received attention in computational linguistics. Since concreteness is a property of word senses rather than of words, it makes most sense to determine concreteness in a given context. Recent approaches for predicting the concreteness of a word occurrence in context have relied on collecting many features from all words in the context. In this paper, we show that we can achieve state-of-the-art results by using only contextualized word embeddings of the target words. We circumvent the problem of missing training data for this task by training a regression model on context-independent concreteness judgments, which are widely available for English. The trained model needs only a few additional training data to give good results for predicting concreteness in context. We can even train the initial model on English data and do the final training on another language and obtain good results for that language as well.

1 Introduction

Word concreteness is one of the psycholinguistic norms of words that has been studied and collected for decades. These scores are obtained by presenting words to subjects and asking them to rate their concreteness on a 5- or 7-point Likert scale. Recently, there has also been more interest in studying the concreteness of specific word senses or words in a given context (see e.g. Gregori et al., 2020; Vandendaele and Grainger, 2022; Bruera et al., 2023; Collacciani et al., 2024).

In this paper, we propose a simple method to predict these contextualized concreteness scores. For the prediction of classical (non-contextualized) concreteness scores, several studies have obtained good results by training a regression model on static word embeddings. We do not have enough annotated data to train a regression model on contextualized embeddings and contextualized concreteness scores. However, we will show that we get good results by training a regression model on averaged embeddings and static concreteness scores, and then applying the trained model to contextualized embeddings to predict contextualized concreteness scores. The results can be further improved by fine-tuning the regression model on a small set of training data with context-dependent concreteness annotations. Using this model, we achieve state-of-the-art results with a system that is much simpler than those proposed in the literature. If we use multilingual embeddings, we can even do the final training in another language.

1.1 Organization of this paper

The remainder of the paper is organized as follows. In section 2 we describe the motivation for computing contextualized concreteness values and previous approaches to the problem. Section 3 describes our approach to the problem and gives details on all methods used. The data used for training and evaluation are given in section 4, the results are given in section 5.

2 Background and related work

Concreteness is a core semantic property of words that has received much attention in psycholinguistic research. Friendly et al. (1982) define concrete words as words that “refer to tangible objects, materials or persons which can be easily perceived with the senses”. Brysbaert et al. (2014) define concreteness as the degree to which the concept denoted by a word refers to a perceptible entity. Theijssen et al. (2011) point out that in general two concepts of concreteness are used that do not completely overlap, namely *sensory perceivability* and *specificity*. However, they also note that most subjects in tests interpret concreteness as *sensory perceivability*.

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 81–88
September 10-13, 2024 ©2024 Association for Computational Linguistics

2.1 Concreteness and ambiguity

Most studies that collected or predicted concreteness values for words either ignored the fact that many words have several senses or excluded ambiguous words, as was already noticed by Gilhooly and Logie (1980). Here, it has also has to be noticed that ambiguity in fact covers a large range of semantic phenomena from homonymy over irregular polysemy to regular polysemy (like e.g. the ambiguity between material and artifact, as in *glass* or object and information as in *book*), but also the distinction between *de re* and *de dicto* interpretation of a word, that might be strongly related to specificity and concreteness. Gilhooly and Logie (1980) found that the most concrete sense usually is the most dominant one. In addition, Đurđević et al. (2017) found that subjects rate mainly the dominant sense in these cases. In contrast Reijnierse et al. (2019) suggest, comparing their values to those of Brysbaert et al. (2014), that the presence of a metaphorical sense lowers the concreteness judgments for the words without any disambiguating information.

A few studies collected concreteness judgments for different word senses, among which (Gilhooly and Logie, 1980) for English, (Hager, 1994) for German, and more recently (Đurđević et al., 2017) for Serbian and both (Reijnierse et al., 2019) and (Scott et al., 2019) for English words. In order to obtain different senses for a word Gilhooly and Logie (1980) used all senses that came first to the mind of at least one of 40 subjects; Đurđević et al. (2017) compare different methods, including the use of a dictionary; Scott et al. (2019) use a list containing ambiguous words with sense indications but give no sources for these lists. Đurđević et al. (2017) included only polysemous words and thus excluded homonyms and words with different part of speech. Reijnierse et al. (2019) concentrate on one interesting aspect and only compare literal and metaphorical meanings of concrete words.

All of these approaches have the problem that a number of senses must first be determined for each word. This problem is avoided by the approach of Gregori et al. (2020), who presented words in a context to the subjects. Consequently, the result is not an inventory of concreteness values for word senses, but rather a resource for training and evaluating algorithms that predict the concreteness of a word in a given context.

2.2 Predicting concreteness

Recently, there has been growing interest in the concreteness of words in the field of computational linguistics. On the one hand side it turns out that concreteness values can be used for several tasks like e.g. detection of metaphors and non-literal language (Turney et al., 2011; Hill and Korhonen, 2014; Frassinelli and Schulte im Walde, 2019; Charbonnier and Wartena, 2021), lexical simplification (Jauhar and Specia, 2012) and multimodal retrieval (Hessel et al., 2018). On the other hand side, some effort was put in models predicting the concreteness of words. Most successful models use static word embeddings as an input to a regression model that predicts the concreteness score of a (not contextualized and/or disambiguated) word Tanaka et al. (2013); Paetzold and Specia (2016); Ehara (2017); Charbonnier and Wartena (2019, 2020).

Most studies that have tried to beat the baseline for the task of predicting concreteness in context organized by Gregori et al. (2020) have used concreteness values, either computed or looked up, from all other words in the sentence, taking advantage of the fact that concrete words tend to occur in the context of other concrete words and abstract words in the context of other abstract words (Tanaka et al., 2013; Frassinelli et al., 2017; Naumann et al., 2018). Only two submissions in the shared task produced results for the English dataset above the simple baseline that we will present below: The systems submitted by Bondielli et al. (2020) and Rotaru (2020). The system with the best results for the English test data from Bondielli et al. (2020), called Non-Capisco, simply takes some kind of weighted average of the general non-contextualized concreteness score of the target word, as given by Brysbaert et al. (2014), and the concreteness scores of all the other words in the sentence. Non-Capisco did not perform very well on the Italian data. Here, the Capisco-Transformers system from the same team performed much better. Capisco-Transformers uses a regression model on the sentence embedding computed by BERT. Note that this is different from our method sketched below: we also use BERT with a regression model, but we use the word embedding of the target word. To get enough data to train this model, they extend the provided training data by automatically generating variants of the provided training sentences and by collecting sentences for non-ambiguous words along with their static concreteness values. The system submitted by Rotaru

(2020), called ANDI, collects concreteness values from the target word and all other words in the sentence, further behavioral norms for all words, static embeddings from three pre-trained static models, and embeddings from four transformer-based models. All of these scores and embeddings are then used to train a regression model that predicts the contextual concreteness score.

In contrast to the ideas behind most of the approaches sketched above, our hypothesis is that contextualized word embeddings contain enough information about the context of a word, and that it should thus be possible to predict the concreteness of a word in a given context using only its contextualized embedding.

3 Methods

Since context independent concreteness values in the huge MT40k inventory of Brysbaert et al. (2014) are probably the values for the most dominant and most frequent sense, we might often make a very good guess for the context dependent values by simply taking the static value. So we will use these static values as a baseline. If a word form is not found in the data from Brysbaert et al., we use its lemma as provided in the test data set (see below).

The basic idea is that we train a regression model on word embeddings, assuming that some of the dimensions in the embeddings represent word concreteness. Since we do not have enough training data, we will train the regression model on static embeddings. To collect static embeddings, we use a large corpus and compute BERT (or RoBERTa) representations of all words, and for each word present in MT40k, we compute the average of all contextualized embeddings in the corpus. We take the average of the last 4 embedding levels. If a word has been split by the BERT tokenizer, we take the average of the embeddings of the parts. Alternatively, we could take the first embedding layer. This would eliminate the need to use a corpus to collect and average contextual embeddings. We will include this variant in the experiment and refer to it as L0 (layer 0). However, we do not expect good results from the models trained on the first layer, since there are many changes throughout the layers. To check if the regression model actually works, we randomly split the set of embeddings with concreteness values into a training set (95%) and a test set (5%) to evaluate the regression model. This is just a check to see whether the model works at all, and not an attempt to get state-of-the-art results for this task.

For the regression models we use a Support Vector Regression (SVR) model with a polynomial kernel from the SciKitLearn library. For all parameters we use standard settings. As a second model we use a multilayer perceptron (MLP) implemented with PyTorch. The MLP has three hidden layers (512, 256 and 128 dimensions, resp.) with ReLU activation and a dropout probability of 20% for each layer. The MLP is trained for 25 epochs using Mean Square Error as loss function and the Adam optimizer with a learning rate of $1 \cdot 10^{-5}$ and a small weight decay to ensure that the model will not focus too much on a few embedding dimensions and neglect others that might be important in the context dependent task. In all cases we use a batch size of 15.

The regression models can be applied immediately to predict the contextualized concreteness scores. Since we have a small set of training data, we can use it to further improve the predictions. In the case of SVR, we add the extracted contextualized embeddings along with the contextualized concreteness scores to the training data. In the case of MLP, we continue to train the model on the additional data. Here we train for 50 epochs and use a smaller weight decay. We will refer to these models as models with extended training.

We also predict concreteness values for Italian. For Italian, we do not have a repository of static concreteness values for a large number of words. To overcome this deficiency, we use a pre-trained multilingual language model, collect word embeddings for the English (!) words again, and train regression models on these data. We then apply the multilingual model and the regression model to the Italian data. For the extended training, we use both the English and the Italian training data.

4 Materials

We use three different pretrained language models, BERT base uncased (Devlin et al., 2019), RoBERTa base (Liu et al., 2019), and BERT multilingual, all obtained from the Hugging Face repository (`https://huggingface.co/`). We found that using BERT large does not improve the results, probably because the regression part gets more parameters to train.

As a corpus to collect BERT embeddings that

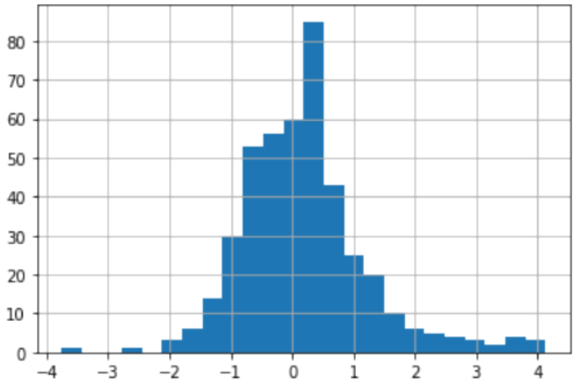

Figure 1: Histogramm of the deviation of the contextualized deviations of the English test data (Gregori et al., 2020) from the MT40k values.

are averaged to obtain static embeddings we use the list of all words from the concreteness data from Brysbaert et al. (2014) and three corpora. The list of single words is used to ensure that every word with a concreteness value has an embedding. This makes the results from the first experiment better comparable to other studies on those data. Each word is given as a sentence to the language model to compute an embedding. Next we used the Brown corpus (Kucera and Francis, 1967). This corpus has a balanced distribution over different genres and might help to include words and word senses not present in the other corpora. The two larger corpora are a collection of 300,000 sentences from a 2016 Wikipedia dump and a corpus of 300,000 sentences from newspapers from 2020, both obtained from `https://wortschatz.uni-leipzig.de/en/download/English` (Goldhahn et al., 2012). These corpora have to be included to collect enough data to compute averaged embeddings for all words.

Static concreteness values were obtained from the collection from Brysbaert et al. (2014), called MT40k, that has ratings for 37,058 words and 2,896 short phrases. When using the BERT tokenizer we could find 29,007 of these words in our corpora. When using the Roberta tokenizer we find 28,122 words (BERT and RoBERTa use the same subword tokenizer, but apparently slightly different pre-tokenizers to split the sentence into words). Embeddings are computed for all single words in Brysbaert's dataset except for a small number of stop words to speed up the data collection process.

Finally, we use the annotated data from Gregori et al. (2020) to finalize the training and to evaluate the models. The provided trial data, that we will use for training as well, consist of 100 sentences, the test set of 434 sentences. In each sentences one word is marked and annotated with a concreteness score. Furthermore, the part of speech and the lemma for the target word are given. In order to investigate how much the values in the test set deviate from the values from the MT40k values we rescaled the later to the range from 1 to 7 and for all 434 examples we subtracted the contextualized value from the static one. The distribution of these differences is shown in the histogram in Fig. 1. Here we see that in most cases there is only a very small deviation from the static value, suggesting that the baseline using MT40k values might give quite good results.

Beside the English trial and test data Gregori et al. (2020) also provide Italian data. For Italian the test data consists of 450 annotated sentences and the trial data (used for training) of 100 sentences.

Recently, both the English and Italian have been extended and are described in much more detail (Montefinese et al., 2023). In this paper we do not yet use these extended data sets.

5 Results

First, we have a look at the results of the regression models on the random split of the static embeddings and MT40k values. These results are given in Table 1. We observe that all results are very good and close to or even slightly better than the results obtained by Charbonnier and Wartena (2019) who used precomputed static embeddings along with additional morpho-syntactic features. However, here we just used a random split, whereas Charbonnier and Wartena (2019) used cross-validation. We do not see large differences between the classifiers or the language models used. The results using the first level embeddings is slightly below the other results. The scatter plot in Figure 2 gives a visual impression of the correlation between the averaged human scores and the values predicted by the MLP using RoBERTa embeddings. At this point we can conclude that in all cases the models learned to predict static concreteness values, the task they were trained for. Next we will see, whether these models are able to predict contextualized concreteness values.

The results for the prediction of the contextualized embeddings are given in Table 2 and visual-

Table 1: Results of predicting concreteness values form a random split of the MT40k data and averaged word embeddings. Both Pearson and Spearman correlation between the predicted and real values are given. The test set has 1847 word-concreteness pairs.

Method	Pearson	Spearman
SVR - BERT	0.913	0.901
SVR - BERT ML	0.892	0.887
SVR - RoBERTa	0.898	0.890
SVR - RoBERTa L0	0.850	0.852
MLP - BERT	0.910	0.897
MLP - BERT ML	0.891	0.887
MLP - RoBERTa	0.902	0.893
MLP - RoBERTa L0	0.858	0.856

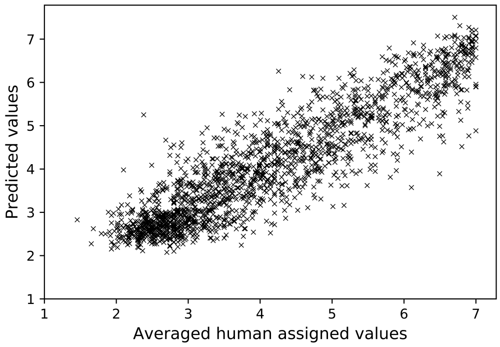

Figure 2: Scatter plot of the concreteness values from MT40k and values predicted by the MLP using RoBERTa embeddings for our test set (randomly selected 1847 words from MT40k)

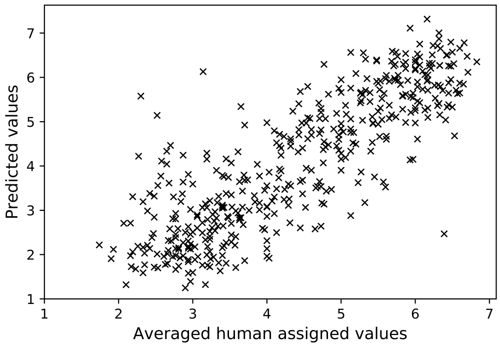

Figure 3: Scatter plot of the contextualized concreteness values and values predicted by the MLP using RoBERTa embeddings for the test data from Gregori et al. (2020) (100 words)

Table 2: Correlation for various regression models, simple base line and state of the art system between predicted and gold standard concreteness values of words in context (N=100). English dataset.

Method	Pearson	Spearman
MT40k Baseline	0.759	0.752
ANDI (SoTA)	0.834	**0.833**
Non-Capisco	0.785	0.787
Capisco-Trans	0.504	0.501
SVR - BERT	0.791	0.793
SVR - BERT ML	0.771	0.767
SVR - RoBERTa	0.820	0.810
SVR - RoBERTa L0	0.446	0.451
MLP - BERT	0.776	0.775
MLP - BERT ML	0.760	0.754
MLP - RoBERTa	0.800	0.790
MLP - RoBERTa L0	0.494	0.483
SVR - BERT - ext.	0.813	0.814
SVR - BERT ML - ext.	0.803	0.804
SVR - RoBERTa - ext.	0.828	0.818
SVR - RoBERTa L0 - ext.	0.341	0.328
MLP - BERT - ext.	0.818	0.816
MLP - BERT ML - ext.	0.790	0.786
MLP - RoBERTa - ext.	**0.838**	0.830
MLP - RoBERTa L0 - ext.	0.420	0.420

ized for one model again in a scatter plot in Figure 3. We see that the simple baseline gives very good results, as expected when looking at the small deviations in Figure 1. Applying the pre-trained regressor to the test data already gives correlations that are clearly above this baseline. The final training in all cases improve the model. The best model, using a MLP and RoBERTa embeddings give results that are very close to the state of the art results from (Rotaru, 2020) and clearly better than the Capisco systems. Furthermore, we see that using first embedding level for the training phase does not give good results, as already expected. The results from the multilingual BERT model are slightly behind those from BERT base, but not very much.

Table 3 gives the results for the Italian data. The results show that the language transfer is successful but the results are behind those from Rotaru (2020) (but better than those of the Capisco systems).

If we inspect the largest errors that were made, we do not find a very clear pattern. Eventually we can get the impression that the model gives too low scores for words referring to specific things that are not clearly perceivable with the senses like, *fear*,

Table 3: Correlation for various regression models and state of the art system between predicted and gold standard concreteness values of the Italian dataset (N=100).

Method	Pearson	Spearman
ANDI (SoTA)	**0.749**	**0.749**
Non-Capisco	0.557	0.557
Capisco-Trans	0.625	0.617
SVR - BERT ML	0.666	0.671
MLP - BERT ML	0.648	0.652
SVR - BERT ML - ext.	0.715	0.715
MLP - BERT ML - ext.	0.732	0.732

answer, *idea*, *advantage*, *success*, etc. and gives too high scores especially to verbs like *hit*, *kick*, *eat* in cases where they do not refer to a physical action. We also find cases, where RoBERTa obviously misinterpreted the sentence, like in *Sign your name in ink in the space provided by the four blank lines.* where *space* gets the score 3.66 instead of 5.6. While *space* usually is some quite abstract word indicating a large range of options to do something, here the word refers to a very concrete area on a piece of paper.

Furthermore, let us have a look at the types of distinctions the model can make. The following sentences are all taken from the Brown Corpus. The concreteness value predicted by the extended Roberta/MLP model is added as a subscript to the word. In the first pair of sentences we see that the model clearly distinguishes homonyms with different concreteness values:

(1) a. Not even an empty cartridge **case**$_{5.9}$ could be found.

b. In this **case**$_{2.4}$ the district manager was led to see the errors of his ways.

Regular polysemy, here between a building and an institution is also captured:

(2) a. John entered the vast **church**$_{6.7}$ and climbed the tower steps to the bells.

b. Surveys show that one out of three Americans has vital contact with the **church**$_{5.1}$.

Finally, we compare two sentences with literal and figurative use of a word. Here we see that the figurative use get still a high concreteness value but clearly a lower one that the literal use.

(3) a. He ran a **finger**$_{7.0}$ down his cheek, tracing the scratch there.

b. Lawrence could not put his **finger**$_{5.6}$ on it precisely, and this worried him.

6 Discussion and Conclusion

We have shown that concreteness of words is a semantic word property that can be derived form a BERT-based word embedding and that can be effectively predicted for word senses in a specific context using only these embeddings, without the need to use information from other words in the sentence. The presented approach is much simpler than previous approaches that used up to 7 different embeddings and had to be trained on many different semantic properties. Our results are close to the state of the art, but do not clearly outperform it. Since the inter-annotator agreement in this type of annotation is usually not very high and the dataset is quite small, it may also be the case that the highest possible agreement with human concreteness scores is already achieved.

The downside of the proposed approach is, that we need to compute averaged embeddings on a large amount of data, as we see that simply using the first (context independent) layer does not give the desired results. This is not only time consuming but also makes the results dependent on the corpus used for this task.

Using multilingual embeddings we also can apply the model to a different language than the language from the training data.

7 Limitations

The main limitation in this study is the availability of annotated data. We have only two very small datasets and only for two languages. However, the topic of the paper is exactly about the approach how to deal with the absence of a large training dataset. A further limitation is that we did not do hyper parameter optimization or model selection for the regression models. We did not do so since we had limited computing resources but also to avoid the risk of overfitting on the small amount of data available. However, it is very likely that slightly better results can be obtained when selecting optimal number of training epochs, layer dimensions, etc.

8 Ethical Considerations

The research presented here did not involve any experiments with humans or animals. All experiments where done with a very limited amount of computational resources and thsu without a high energy consumption and enviromental impact. The research results are rather theoretical and will not have a direct impact on the working or living circumstances of anyone. We hope that this research will contribute to the understanding of large language models and natural languyage processing in general. Here we rather believe that a better understanding of these methods and a more widespread dissemination of this knowledge helps to identify and deal with possible threats from this technology.

References

Alessandro Bondielli, Gianluca E. Lebani, Lucia C. Passaro, and Alessandro Lenci. 2020. Capisco@ concretext:(un) supervised systems to contextualize concreteness with norming data. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR-WS.org.

Andrea Bruera, Yuan Tao, Andrew Anderson, Derya Çokal, Janosch Haber, and Massimo Poesio. 2023. Modeling brain representations of words' concreteness in context using gpt-2 and human ratings. *Cognitive Science*, 47(12):e13388.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 176–187.

Jean Charbonnier and Christian Wartena. 2020. Predicting the Concreteness of German Words. In *Proceedings of Konvens / SwissText*.

Jean Charbonnier and Christian Wartena. 2021. Verbal idioms: Concrete nouns in abstract contexts. In *KONVENS 2021, Düsseldorf, Germany, 06–09 September 2021*.

Claudia Collacciani, Andrea Amelio Ravelli, and Marianna Bolognesi. 2024. Specifying genericity through inclusiveness and abstractness continuous scales. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15126–15136, Torino, Italia. ELRA and ICCL.

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.

Yo Ehara. 2017. Language-independent prediction of psycholinguistic properties of words. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 330–336.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte im Walde. 2017. Contextual characteristics of concrete and abstract words. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Diego Frassinelli and Sabine Schulte im Walde. 2019. Distributional interaction of concreteness and abstractness in verb–noun subcategorisation. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 38–43, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Friendly, Patricia E. Franklin, David Hoffman, and David C. Rubin. 1982. The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14(4):375–399.

K. J. Gilhooly and R. H. Logie. 1980. Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation*, 12(4):428–450.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Lorenzo Gregori, Maria Montefinese, Daniele P. Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. Concretext@ evalita2020: The concreteness in context task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR-WS.org.

Willy Hager. 1994. Bildhaftigkeit, Konkretheit-Abstraktheit und Bedeutungshaltigkeit von 63 mehrdeutigen Substantiven. In Willi Hager and Marcus Hasselhorn, editors, *Handbuch deutschsprachiger Wortnormen*, chapter 3.6, pages 212–217. Hogrefe Verlag für Psychologie, Göttingen.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2194–2205, New Orleans, Louisiana. Association for Computational Linguistics.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–731.

Sujay Kumar Jauhar and Lucia Specia. 2012. Uowshef: Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 477–481.

Henri Kucera and Winthrop Francis. 1967. Computational Analysis of Present-Day American English. Technical report, Brown University Press, Providence, RI.

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.

Maria Montefinese, Lorenzo Gregori, Andrea Amelio Ravelli, Rossella Varvara, and Daniele Paolo Radicioni. 2023. Concretext norms: Concreteness ratings for italian and english words in context. *PLOS ONE*, 18(10):1–19.

Daniela Naumann, Diego Frassinelli, and Sabine Schulte im Walde. 2018. Quantitative Semantic Variation in the Contexts of Concrete and Abstract Words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 76–85, New Orleans, LA, USA.

Gustavo Paetzold and Lucia Specia. 2016. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 435–440, San Diego, California. Association for Computational Linguistics.

W. Gudrun Reijnierse, Christian Burgers, Marianna Bolognesi, and Tina Krennmayr. 2019. How polysemy affects concreteness ratings: The case of metaphor. *Cognitive Science*, 43(8):e12779.

Armand Stefan Rotaru. 2020. Andi@ concretext: Predicting concreteness in context for english and italian using distributional models and behavioural norms (short paper). In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR-WS.org.

Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270.

Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. 2013. Estimating content concreteness for finding comprehensible documents. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 475–484, New York, NY, USA. ACM.

Daphne Theijssen, Hans van Halteren, Lou Boves, and Nelleke Oostdijk. 2011. On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands Journal*, 1:61–77.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dušica Filipović Đurđević, Aleksandar Kostić, and Zorana Đinđića. 2017. Number, relative frequency, entropy, redundancy, familiarity, and concreteness of word senses: Ratings for 150 serbian polysemous nouns. In *Selected Papers From the 4th and 5th Workshop on Psycholinguistic, Neurolinguistic and Clinical Linguistic Research*, volume 2 of *Studies in Language and Mind*, pages 13–50. Filozofski fakultet u Novom Sadu.

Aaron Vandendaele and Jonathan Grainger. 2022. Now you see it, now you don't: Flanker presence induces the word concreteness effect. *Cognition*, 218:104945.