Predicting the Concreteness of German Words

Jean Charbonnier and Christian Wartena
Hochschule Hannover
Expo Plaza 12, 30539 Hannover, Germany
{jean.charbonnier, christian.wartena}@hs-hannover.de

Abstract

Concreteness of words has been measured and used in psycholinguistics already for decades. Recently, it is also used in retrieval and NLP tasks. For English a number of well known datasets has been established with average values for perceived concreteness. We give an overview of available datasets for German, their correlation and evaluate prediction algorithms for concreteness of German words. We show that these algorithms achieve similar results as for English datasets. Moreover, we show for all datasets there are no significant differences between a prediction model based on a regression model using word embeddings as features and a prediction algorithm based on word similarity according to the same embeddings.

1 Motivation

A number of properties of words, mainly of semantic nature, have been studied and used in psycholinguistic research for decades. These properties often referred to as (affective) word norms, include concreteness, imagery[1], age of acquisition, valence, and arousal. In the present work we will focus on concreteness. 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". Similarly, Brysbaert et al. (2014) define concreteness as the degree to which the concept denoted by a word refers to a perceptible entity, but found that subjects largely rated the haptic and visual experiences, even if they were explicitly asked to take into account experiences involving any senses.

Concreteness seems to play an important role in human language processing (Borghi et al., 2017). Concreteness also has been used for various computational linguistic tasks like detection of metaphors and non-literal language (Turney et al., 2011; Hill and Korhonen, 2014; Frassinelli and Schulte im Walde, 2019), lexical simplification (Jauhar and Specia, 2012), multimodal retrieval (Hessel et al., 2018) or estimating the stability of word embeddings (Pierrejean and Tanguy, 2019).

Traditionally, word norms are obtained by asking subjects to estimate the value for each property on a Likert scale. Recently, also various approaches have been proposed to predict the concreteness of words. On three different datasets we will test two algorithms that have given very good results for English data and compare the results in section 4 after we have discussed the most common approaches to predict word concreteness (section 2) and presented the concreteness data available for German (section 3).

2 Related work

We find basically three approaches to predict the concreteness of a word: (1) adopting the concreteness value from similar, related or neighboring words; (2) identifying a dimension in word embeddings that corresponds to concreteness; (3) training a regression model on features of words

2.1 Adopting concreteness of related words

Liu et al. (2014) predict values for imagery, a word norm that strongly correlates with concreteness, by using the values from synonyms and hypernyms found in WordNet.

[1] Most authors seem to use the term *imagery*, while others also use *imageability* and *visualness*. In German the term *Bildhaftigkeit* is the most common one, while also *Vorstellbarkeit* is found. We will use *imagery* throughout this paper.

Rabinovich et al. (2018) predict the concreteness of words indirectly by assigning a concreteness value to sentences in which a word occurs. The concreteness value of a sentence is based on the presence of seed words. The set of seed words is constructed by selecting words with derivational suffixes that are typical for highly abstract nouns. The correlation between the predicted values and the manual assigned values from various subsets of the dataset from Brysbaert et al. (2014) and the 4,295 concreteness values[2] from the MRC (Medical Research Council) Psycholinguistic Database (Coltheart, 1981) ranges from 0.66 to 0.74.

Turney et al. (2011) compute the degree of concreteness of a word as the sum of the similarities between the word and n abstract paradigm words minus the sum of the similarities between the word and n concrete paradigm words. The paradigm words are found as follows: first one concrete and one abstract paradigm word are selected such that the concreteness values for all words in the training data, predicted by using the similarity with these two words is maximized. Then a second concrete and abstract word are added that again maximize the correlation. This process is repeated until n abstract and concrete words are found. Turney et al. (2011) limit the selection to 20 abstract and concrete paradigm words. Using half of the MRC data for training and half for testing, they found Spearman correlation coefficient of 0.81 between predicted and observed concreteness values. To compute the similarity between words they use count based word embeddings of 1000 dimensions trained on a $5 \cdot 10^{10}$ word web corpus. The same approach is followed by Köper and Schulte im Walde (2016) to predict concreteness values for German words using word vectors trained with word2vec (Mikolov et al., 2013) on the DE-COW14AX German Web corpus. For training and testing they merged concreteness values from Kanske and Kotz (2010) (called Leipzig Word Norms below) and Lahl et al. (2009) (called WWN below) and added in addition translations from sets of English word norms for training. 90% of the data were used for training, 10% for testing. The Pearson correlation between the test data and the predicted values for concreteness/abstractness was 0.825.

2.2 Concreteness in word embeddings

Rothe et al. (2016) try to find low-dimensional feature representations of words in which at least some dimensions correspond to interpretable properties of words. One of these dimensions is concreteness. For training and testing they use Google News embeddings and two subsets of frequent words from the norms of Brysbaert et al. (2014). For their test set of 8,694 frequent words they found a moderate correlation with the human judgments (Kendall's $\tau = 0,623$). Similarly, Hollis and Westbury (2016) looked which dimension of word embeddings correlate to one of the classical word norms. They found no direct correlations, but after reducing the number of dimensions for a set of words by applying Singular Value Decomposition, they found a strong correlation between one of the dimensions and concreteness.

2.3 Regression models for concreteness

Tanaka et al. (2013) train a regression model to predict concreteness values. As features they use a small number of manually constructed co-occurrence features, like co-occurrence with sense verbs. For training and evaluation they use a subset of 3,455 nouns from the MRC Database. Pearson's correlation and Kendall's τ between the values from the database and their predictions are 0.688 and 0.508, respectively.

Paetzold and Specia (2016) train a regression model to predict four word norms, among which concreteness. Like many other studies, they use the data from the MRC database. As features they use word embeddings trained on a set of various large corpora and a number of word features extracted from WordNet. For each word norm they use half of the words to train the model and half of the words for evaluation. For concreteness they find a Pearson correlation coefficient of 0.869.

Ehara (2017) trains regression models to predict four word norms for Japanese and English words. As features they use word embeddings trained with word2vec and a probability distribution of words over topics found using Latent Dirichlet Allocation. They use a subset of 1,842 words from the MRC data, from which 1,342 words are used for training and 500 for testing. When both feature sets are trained on the British National Corpus (BNC) and used in combination, the best regression model gives a Pearson correlation of 0.87 and a Spearman correlation coefficient of 0.876 on the test data.

[2] The current version of MRC has concreteness values aggregated from different sources for 8,288 words. We assume that a previous version provided concreteness values for 4,295 words.

Ljubešić et al. (2018) used a regression model as well with pre-trained fastText word embeddings (Mikolov et al., 2018). They found a Spearman correlation coefficient of 0.887 between the predicted concreteness values and the values from Brysbaert et al. (2014) and a Spearman correlation of 0.872 on the MRC data, in both cases using 3-fold cross validation. A similar result was found by Charbonnier and Wartena (2019), who reach a Pearson correlation coefficient of 0.91 on the data from Brysbaert et al. (2014), using the same vectors and 10-fold cross validation. Here a minor improvement could be realized using part of speech and frequent suffixes as additional features.

Though all studies use different data and different versions of the MRC Psycholinguistic database, use different splits and different number of folds for cross validation and finally use different correlation coefficients, all studies report very similar results. The correlations that are found are all in the range of correlations found between various sets of concreteness values (see Charbonnier and Wartena, 2019, Table 2).

3 Data

Both, for English and German, various word norms with concreteness values have been created, though some are quite small and only available as printed supplements to older publications.

The dataset created by Baschek et al. (1977) and Wippich and Bredenkamp (1979) consists of 1698 words (800 nouns, 400 adjectives, 498 verbs) is one of the oldest and still one of the largest word norms for German. We will refer to this dataset as the Göttingen Word Norms. We removed 40 verbs containing an underscore, especially all reflexive verbs (e.g. *sich_wünschen*; to wish), from the data set. For number of words the experiment was repeated and two values are given. We only use the first value in these cases.

Lahl et al. (2009) collected values for 2,654 words using crowdsourcing to build a dataset called the Web Word Norms (WWN). For the WWN 3,907 subjects committed 190,212 ratings, each for at most 50 words. On average each word has 24 ratings. They used a 11-point scale were 0 stands for the most concrete and 10 for the least concrete judgment.

Kanske and Kotz (2010) collected ratings for valence, arousal and concreteness for 1000 nouns. This dataset is known as the Leipzig Affective Word Norms. Only nouns were used to reduce the variance other word classes would introduce. The experiment was done in 2006 with 32 native speaker. On two separate days the participants rated the words 3 times on a 9-point scale, each time for one of the three ratings. This was repeated 2 years later with two groups, one with 22 repeating participants from 2006 and a second with 32 fresh participants. The words were collected from the Duden dictionary and a previous word list by the same authors. Only 1 and 2-syllable words and no compound nouns were allowed.

The Berlin Word Norms (Vo et al., 2009) and the word norms determined by Schmidtke et al. (2014) contain values for valence, arousal and imagery but no values for concreteness. Some more word norms for German, including concreteness, are published by Hager and Hasselhorn (1994).

3.1 Merged Dataset

In order to have a larger dataset for German, providing more training data for supervised prediction algorithms, we created a merged dataset.

The overlap of the datasets is quite small (see Table 1), the correlation between the values for the overlapping parts, however, is high (around 0.9). Since the Leipzig Word Norm uses low values for concrete and high values for abstract words, the correlation between this and the other datasets is negative.

For the merged data set we use the 7 point scale where 1 means abstract and 7 means concrete. We do not simply rescale the values but use linear regression on the overlapping parts such that the values for the words in overlapping parts are as close as possible. We take the values from the Göttingen Word Norms as an anchor and transform the other values using the slope and the intercept. The transformed concreteness thus is defined as

$$C' = \alpha + \beta C \qquad (1)$$

where C is the original value. For WWN $\alpha = 0.776$ and $\beta = 0.608$ and for the Leipzig Word Norms $\alpha = 7.39$ and $\beta = -0.540$. Finally, we take the average from all datasets if a word is present in more than one source. The dataset thus offers empirical concreteness values for 4,182 German words. In Table 1 we see the high correlation of the values in the merged data with those in the original datasets. The merged dataset can be downloaded from http://textmining.wp.hs-hannover.de/datasets.html

Table 1: Size of the intersections and the Pearson correlation between the concreteness values in the datasets. As the Merged set is a composition of the other dataset, the intersection is always equal to the size of the other dataset.

	Merged WN		Göttingen WN		WWN	
	Inters.	Correl.	Inters.	Correl.	Inters.	Correl.
Göttingen WN	1698	0.997				
WWN	2654	0.969	680	0.900		
Leipzig WN	1000	-0.985	127	-0.928	488	-0.875

Table 2: Results of 5-fold cross validation using different methods for all datasets. All results are averaged Pearson correlation coefficients. For Turney we used 20 words per class.

	Merged	Göttingen WN	WWN	Leipzig WN
SVR	0.861 (± 0.026)	0.862 (± 0.040)	0.851 (± 0.023)	0.890 (± 0.027)
Turney et al.	0.849 (± 0.012)	0.842 (± 0.033)	0.851 (± 0.020)	0.901 (± 0.017)

4 Methods

For each dataset we use two methods to predict the concreteness values in a five-fold cross validation scheme. We compare the method of Turney et al. (2011) described above in section 2. Following Turney et al. (2011) and Köper and Schulte im Walde (2016) we use 20 abstract and 20 concrete prototype words. As a second method we use Support Vector Regression (Drucker et al., 1997) and grid search to find optimal hyper parameters ($\gamma = 1$, $C = 10$ with an *rbf* kernel). As features we use the pre-trained Word-embeddings from fastText for German (Grave et al., 2018).

All test were done using 5-fold cross validation. We use stratified sampling for the Göttingen WN and the Merged dataset to ensure that each fold has the same number of nouns, verbs and adjectives. For the other dataset we use random splits.

5 Results and Discussion

The results for all datasets and both methods are given in Table 2. We see in general very high correlation values for all datasets and both methods. All correlation values are in a similar range as the correlations between the datasets.

We can make some interesting observations. The first remarkable fact is, that for all datasets there is no significant difference between the results from the method of Turney et al. (2011) and the regression model. As far as we know, these methods have not been compared directly before. This result is quite surprising, since there are many aspects of the meaning of a word that determine the word similarity. All of these aspects are used to find the similar words on which the concreteness prediction is based in the method of Turney *et al.* It has to be noted that the search of the prototype words in Turney's method is extremely slow and not feasible for large datasets.

Furthermore, we see that our implementation of the method of Turney *et al.* gives slightly better results for WWN and the Leipzig Word Norms than the result found by Köper and Schulte im Walde (2016), who used a random split of the unification of those two datasets (0.844 and 0.891 vs. 0.825). Besides the possibility that they have chosen a disadvantageous split, we see two differences: In the first place we used different word embeddings to compute the word similarity. Secondly, they added concreteness values from English datasets with German translations to the training data. This is only helpful if concreteness is invariant under translation. This might be not the case.

6 Conclusions and Future Work

Datasets with concreteness values for German are smaller and less easily accessible than those for English. One contribution of the present work is that we aggregated a consistent dataset with over 4000 concreteness ratings from three different sources.

A possibility to obtain more concreteness ratings is to train a model on available ratings and predict ratings for other words. We show that prediction methods that have been tested for English only before yield similar results for German. Moreover, we show that two of the best available methods that have not been compared on the same data before, yield similar results with no significant differences on 4 different data sets.

In near future we will extend the merged dataset with values from some smaller and older studies.

References

Ilse-Lore Baschek, Jürgen Bredenkamp, Brigitte Oehrle, and Werner Wippich. 1977. Determination of imagery, concreteness and meaningfulness of 800 nouns. *Zeitschrift für experimentelle und angewandte Psychologie*, 24(3):353–396.

Anna M Borghi, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. 2017. The challenge of abstract concepts. *Psychological Bulletin*, 143(3):263.

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.

Max Coltheart. 1981. The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

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 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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Willi Hager and Marcus Hasselhorn, editors. 1994. *Handbuch deutschsprachiger Wortnormen*. 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.

Geoff Hollis and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, 23(6):1744–1756.

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.

Philipp Kanske and Sonja A. Kotz. 2010. Leipzig affective norms for german: A reliability study. *Behavior Research Methods*, 42(4):987–991.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598.

Olaf Lahl, Anja S. Göritz, Reinhard Pietrowsky, and Jessica Rosenberg. 2009. Using the world-wide web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 german nouns. *Behavior Research Methods*, 41(1):13–19.

Ting Liu, Kit Cho, G. Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz, Ignacio Cases, and Ching-sheng Lin. 2014. Automatic expansion of the MRC psycholinguistic database imageability ratings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2800–2805, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Bénédicte Pierrejean and Ludovic Tanguy. 2019. Investigating the stability of concrete nouns in word embeddings. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 65–70, Gothenburg, Sweden. Association for Computational Linguistics.

E. Rabinovich, B. Sznajder, A. Spector, I. Shnayderman, R. Aharonov, D. Konopnicki, and N. Slonim. 2018. Learning Concept Abstractness Using Weak Supervision. *ArXiv e-prints*.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777. Association for Computational Linguistics.

David S Schmidtke, Tobias Schröder, Arthur M Jacobs, and Markus Conrad. 2014. Angst: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior research methods*, 46(4):1108–1118.

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.

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.

Melissa LH Vo, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The Berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538.

Werner Wippich and Jürgen Bredenkamp. 1979. *Bildhaftigkeit und Lernen*, volume 78 of *Wissenschaftliche Forschungsberichte*. Steinkopff-Verlag, Darmstadt.