Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas

Maximilian Köper **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,schulte}@ims.uni-stuttgart.de

Abstract

This paper presents a collection of 350 000 German lemmatised words, rated on four psycholinguistic affective attributes. All ratings were obtained via a supervised learning algorithm that can automatically calculate a numerical rating of a word. We applied this algorithm to abstractness, arousal, imageability and valence. Comparison with human ratings reveals high correlation across all rating types. The data discussed in the present paper is accessible at:
http://www.ims.uni-stuttgart.de/data/affective_norms/

Keywords: affective norms, abstractness, arousal

1. Introduction

Abstract words refer to things that we cannot perceive directly with our senses (*idea*, *politics*, ...), concrete words on the other hand refer to things that we can perceive (*image*, *scent*, ...,). A large subset of concrete words have a high imageability, these are words that refers to things that we can actually see. Valence determines the pleasantness of a word (*gift* vs. *punishment*). Arousal describes the intensity of emotion provoked by a stimulus (*alert* vs. *calm*). These psycholinguistic attributes are often called affective norms. Information about the affective meaning of words is used by researchers working in multiple fields, such as sentiment analysis, metaphor detection, word processing and lexical decision.

Currently, a number of databases offers affective norms for words in different languages, including English (Altarriba et al., 1999; Bradley and Lang, 1994; Stevenson et al., 2007; Warriner et al., 2013), German (Võ et al., 2006; Võ et al., 2009; Lahl et al., 2009; Kanske and Kotz, 2010; Schmidtke et al., 2014), Spanish (Redondo et al., 2007), and Finnish (Eilola and Havelka, 2010). There is a number of resources that focuses on a single rating type only: Brysbaert et al. (2014) collected 40.000 abstractness ratings for English. The MRC Psycholinguistic Database[1] contains roughly eight thousand abstractness ratings. However all of these resources cover only a small proportion of a language due to the fact that human annotators are required to obtain ratings. Especially the German resources contain usually less than 3000 words. In addition most resources focus on nouns only and therefore lack other word classes such as verbs or adjectives. A notable exception is the work by Turney et al. (2011), who used a method that overcome these limitations by applying an algorithm from Turney and Littman (2003). This algorithm uses distributional semantics (vector representations of words) and learns to assign a rating score to unseen words based on other known labelled training instances. Using this method, they were able to learn abstractness ratings for 114 501 English words. In this work we apply the same method as in Turney et al. (2011) and: (i) we learn ratings for German, (ii) we learn four different rating types, (iii) recent advances in the field of word representation learning (namely the methods of Mikolov et al. (2013)) allow us to learn ratings for a large vocabulary ($\approx 350k$ words).

[1] http://www.psych.rl.ac.uk/

2. Training Data & Preprocessing

Preparing the data is done in three steps: first, we collect available German ratings. Then we extend the data by translating some of the English ratings into German. Finally we merge all resources together.

In more detail, the algorithm that assigns a rating score requires labelled training data. Therefore we first collected several affective ratings for German. Table 1 lists all available resources together with the ratings, that we used. As

Source	Words	Abs.	Ar.	Val.	Img.
Võ et al. (2009)	2902	✗	✓	✓	✓
Lahl et al. (2009)	2654	✓	✓	✓	✗
Kanske and Kotz (2010)	1000	✓	✓	✓	✗
Schmidtke et al. (2014)	1000	✗	✗	✗	✓
MRC∩Brys *(EN → GER)*	3266	✓	✗	✗	✗
# Unique words		5237	4848	4848	2901

Table 1: German Resources

it can be seen, not every resource contains all four types of rating. In addition, we decided to translate some of the English abstractness ratings from MRC and (Brysbaert et al., 2014) to increase the number of available training instances. Here we computed the intersection of both resources and used a translation tool[2] to translate the words from English to German. Missing and double translations were removed from the list, resulting in a final list of 3 266 additional words together with their abstractness ratings. We finally computed a total set of words (=unique words) for each rating type by mapping all ratings to the same scale, namely $[0, 10]$. We did this mapping by using a continuous function (see equation 1). This function maps numbers from an interval $[min, max]$ to a new interval $[a, b]$.

[2] Translation was done by applying the following java-google-translate-text-to-speech API: https://code.google.com/p/java-google-translate-text-to-speech/

We set $a = 0$ and $b = 10$. We computed mean values in case of overlapping words. For every rating type we divided the number of unique words randomly into two sets: training data (90%) and test data (remaining 10%). For the abstractness set, we made sure that the test data contains only human labelled data and not the translated ratings. The test data is later used to validate the algorithm by comparing the original ratings with the ratings created by the algorithm.

$$f(x) = \frac{(b-a)(x-min)}{max-min} + a \quad (1)$$

3. Algorithm

The core idea of the algorithm from Turney and Littman (2003) is that the degree of abstractness (or arousal, valence, imageability) can be expressed by comparing a given word w_i with a list of positive and a list of negative paradigm words. Each word is represented by a high-dimensional vector (based on context counts). A rating score $R(w_i)$ is then computed by simply calculating the similarity with all positive paradigm words minus the similarity with all negative paradigm words:

$$R(w_i) = \sum_{p_j \in positive} sim(w_i, p_j) - \sum_{n_j \in negative} sim(w_i, n_j) \quad (2)$$

Similarity is measured using cosine distance. The algorithm begins with an empty set of paradigm words and adds one word at a time to the paradigm list, alternating between adding a word to the positive paradigm words and then adding a word to the negative paradigm words. At each step, we add the paradigm word that results in the highest Pearson correlation with the ratings of the training data. This is a form of greedy forward search without backtracking. The algorithm terminates when 20 paradigm words have been added to both lists (after 40 iterations). Finally we assign each word in our vocabulary a rating score by using equation 2. This score is then rescaled to a numerical number within $[0, 10]$ by using the function from equation 1.

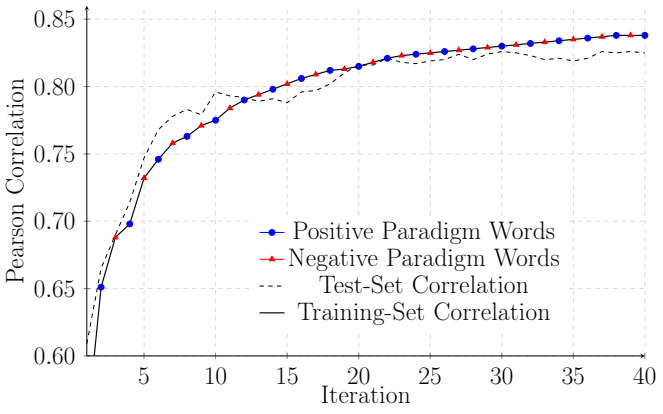

Figure 1: Example progress: train and test correlation, learning abstractness

To validate how well the algorithm works on unseen words we simultaneously measure the Pearson correlation with the (unknown) test data ratings. Note that these ratings did not influence the paradigm word selection (training process). Figure 1 shows the increasing Pearson correlation for training and test data when paradigm words are added with regard to the abstractness ratings. The figure shows that both correlations obtain already a high correlation (> 0.80) after only 15 iterations. While the training correlation increases with each iteration, it can be observed that in some cases the next paradigm word decreases the correlation for the test data.

4. Distributional Information

In order to apply the algorithm explained in Section 3., word representations were required. Since the training process considers every possible word of the vocabulary when selecting the next paradigm word, it is especially useful to work with low-dimensional word representations. Recent work by Mikolov et al. (2013) introduced an efficient way to learn low dimensional but reliable word representations. In addition Baroni et al. (2014) and Köper et al. (2015) compared these representations across a variety of tasks, showing superior performance to traditional count vector space approaches. We used the *word2vec toolkit*[3] and applied it to a lemmatised version of the DECOW14AX German web corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015). This corpus contains about 20 billion tokens. We ignored words that occurred less than 100 times in the corpus. In addition we tuned the hyper-parameter settings on two German word correlation tasks: *Gur350* (Zesch and Gurevych, 2006) and *Gur65* (Gurevych, 2005). These tasks compare distributional similarity (cosine) with human-annotated similarity values. Finally we took the vectors that obtained the best performance, in terms of Pearson correlation. The final model used the skip gram architecture with a symmetrical window of size 3, 400 dimensions, and SubSampling with $t = 1e^{-5}$.

5. Ratings

Using the algorithm described in Section 3. together with the word vectors explained in Section 4., we were able to obtain a total of 351 617 ratings (86% nouns, 10% verb, 4% adj+adv). After 40 iterations, the final correlation scores between training and test data are presented in Table 2.

	Abs	Ar	Val	Img
Training Correlation	0.838	0.796	0.827	0.832
Test Correlation	0.825	0.784	0.798	0.789

Table 2: Final Pearson correlation after 40 iterations

It can be seen that all correlations are sufficiently high. Furthermore the training data provides only slightly higher correlations than the test data. We can therefore assume that the algorithm delivered reliable ratings even for words that were not used for training. Table 3 lists some words together with their respective rating score. For each rating type and word class we present two words with high rating scores and two words with a low rating score.

[3] https://code.google.com/p/word2vec/

↑↓	Adj+Adv	*English*	Rating	Verb	*English*	Rating	NN	*English*	Rating
Abstractness-Concreteness									
↑	aufgeblättert	exfoliated	8.44	entlangrutschen	slip along	7.97	Uniformtasche	uniform bag	10.00
↑	beinlang	leg-length	8.32	bekleckseln	blot	7.92	Vampirgebiss	vampire ivories	9.61
↓	paradox	paradox	0.63	negieren	negate	0.77	Selbstläuterung	self purification	0.52
↓	rechtfertigbar	justifiable	0.36	innewohnen	inhere	0.64	Willenlosigkeit	abulia	0.66
Arousal									
↑	bestialisch	brutish	9.33	vergewaltigen	rape	9.85	Bandenrivalität	gang rivalry	10.0
↑	gewalttätig	violent	9.25	umbringen	kill	9.32	Blutbad	bloodbath	9.83
↓	satzweise	blockwise	0.81	flechten	weave	1.35	Wortfamilie	word family	0.48
↓	ausgerollt	rolled out	0.71	einfüllen	pour in	1.22	Holzdeckel	wood cover	0.00
Imageability									
↑	neonerleuchtet	neon-lighted	9.83	weiterschnüffeln	continue sniffing	8.99	Granatangriff	grenade attack	10.0
↑	schwarzverhüllt	black-cloaked	9.61	emporzüngeln	upwardslicking	8.71	Polizeijeep	police Jeep	0.87
↓	gewiss	certain	0.58	auffassen	understand	1.08	Zielstellung	objective	0.59
↓	unstrittig	indisputable	0.41	elaborieren	elaborate	1.08	Sinnfreiheit	mindless	0.46
Valence									
↑	wundervoll	wonderful	9.69	beschenken	endow	8.35	fitnessangebot	Fitness offer	10.0
↑	wunderbar	marvellous	9.69	genießen	enjoy	8.28	Frühlingsrezept	spring recipe	9.49
↓	katastrophenmässig	disastrous	0.39	zermürben	demoralize	0.60	Falschdiagnose	misdiagnosis	0.06
↓	ausblutend	bleeding to death	0.37	frikassieren	fricassee	0.57	Essensentzug	food deprivation	0.00

Table 3: Example words with high and low ratings per rating type and word class.

When comparing all ratings pairwise with each other (Figure 2) we observe a high Pearson correlation (0.81) between imageability and concreteness and a moderate negative correlation between arousal and abstractness and valence. wile all other pairwise comparisons exhibit no correlations.

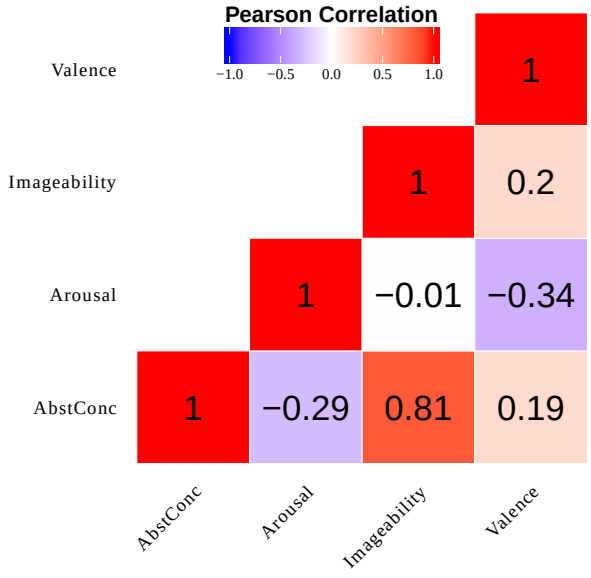

Figure 2: Comparing all ratings pairwise

When comparing the four rating types across different word classes (Figure 3) we can observe that the median is always close to ≈ 5, which is the middle of our scale. In addition we can see that nouns tend to have a higher rating for imageability and abstractness (high degree = rather concrete). This observation is not surprising, since nouns tend to be visual whereas verbs and adjectives are rather abstract.

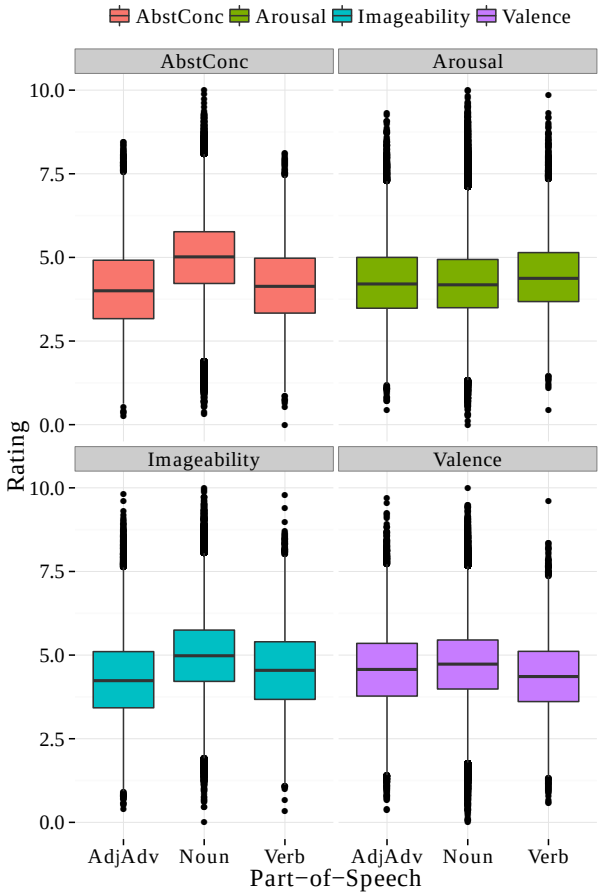

Figure 3: Distribution across Part-of-Speech.

6. Conclusion

This paper presented a large database of automatically generated affective norms for German. To the best of our knowledge this resource is currently the largest affective dictionary for the German language. We hope that the provided norms are useful for a variety of applications and support future research in different research areas.

7. Acknowledgements

The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

8. Bibliographical References

Altarriba, J., Bauer, L., and Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods*, 31(4):578–602.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *J Behav Ther Exp Psychiatry*, 25(1):49–59, March.

Brysbaert, M., Warriner, B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Eilola, T. and Havelka, J. (2010). Affective norms for 210 british English and Finnish nouns. *Behavior Research Methods*, 42(1):134–140.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, IJCNLP'05, pages 767–778, Berlin, Heidelberg. Springer-Verlag.

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

Köper, M., Scheible, C., and Schulte im Walde, S. (2015). Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 40–45, London, UK, April. Association for Computational Linguistics.

Lahl, O., Göritz, A., Pietrowsky, R., and Rosenberg, J. (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):13–9.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Redondo, J., Fraga, I., Padrón, I., and Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior Research Methods*, 39(3):600–605.

Schäfer, R. and Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Schäfer, R. (2015). Processing and querying large web corpora with the cow14 architecture. In Piotr Bański, et al., editors, *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), Lancaster, 20 July 2015*, pages 28 – 34.

Schmidtke, D., Schröder, T., Jacobs, A., and Conrad, M. (2014). Angst: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior Research Methods*, 46(4):1108–1118.

Stevenson, R., Mikels, J., and James, T. (2007). Characterization of the affective norms for english words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, October.

Turney, P., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Võ, M., Jacobs, A., and Conrad, M. (2006). Cross-validating the berlin affective word list. *Behavior Research Methods*, 38(4):606–609.

Võ, M., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M., and Jacobs, A. (2009). The berlin affective word list reloaded (bawl-r). *Behavior Research Methods*, 41(2):534–538.

Warriner, A., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Zesch, T. and Gurevych, I. (2006). Automatically creating datasets for measures of semantic relatedness. In *COLING/ACL 2006 Workshop on Linguistic Distances*, pages 16–24, Sydney, Australia.