A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis

Bettina M. J. Kern
University of Vienna, Austria

Andreas Baumann
University of Vienna, Austria

Thomas E. Kolb
TU Wien, Austria

Katharina Sekanina
University of Vienna, Austria

Klaus Hofmann
University of Vienna, Austria

Tanja Wissik
Austrian Academy of Sciences, Vienna, Austria

Julia Neidhardt
TU Wien, Austria

Abstract

The domain of German polarity dictionaries is heterogeneous with many small dictionaries created for different purposes and using different methods. This paper aims to map out the landscape of freely available German polarity dictionaries by clustering them to uncover similarities and shared features. We find that, although most dictionaries seem to agree in their assessment of a word's sentiment, subsets of them form groups of interrelated dictionaries. These dependencies are in most cases an immediate reflex of how these dictionaries were designed and compiled. As a consequence, we argue that sentiment evaluation should be based on multiple and diverse sentiment resources in order to avoid error propagation and amplification of potential biases.

2012 ACM Subject Classification Computing methodologies → Cluster analysis

Keywords and phrases cluster analysis, sentiment polarity, sentiment analysis, German, review

Digital Object Identifier 10.4230/OASIcs.LDK.2021.37

Supplementary Material *Software (Source Code)*: `https://github.com/bettina-mj-kern/LDK_2021`

Funding This research was funded by the City of Vienna (Digitaler Humanismus grant, MA7-737909/19).

1 Introduction

Sentiment analysis is a popular tool to draw emotional information from language data. One approach for sentiment detection is the use of lexical resources, i.e., sentiment dictionaries containing a list of words and their corresponding sentiment information. An ample selection of sentiment dictionaries exists for the English language. German, however, with its abundance of compound words, inflections and derivational suffixes, poses more of a challenge for automated sentiment analysis [6] and for the development of adequate tools and methods.

Many sentiment dictionaries contain sentiment ratings on more than one aspect than just polarity. The dimensional view is a common conception of emotion, in which emotions are characterised as quantitatively different from each other on a number of dimensions [8, 24, 34]. Accordingly, different dimensions can be used to describe the sentiment information of a word, polarity being one of them. Different sentiment dictionaries make use of different conceptualisations of emotions and include different dimensions, such as arousal, valence, or dominance, to capture the emotional content of words. This makes them difficult to compare. For this paper, we thus focus on the common denominator for most German sentiment resources: sentiment polarity, also often referred to as valence and sometimes as evaluation.

The assessment of polarity is one of the most basic tasks in sentiment classification and reflects whether a word or a text snippet is positive or negative. While the number and the definition of emotional dimensions may vary between dictionaries, most of them contain ratings on sentiment polarity. Dictionaries also differ in how they encode polarity: some of them provide categorical polarity labels (eg. NEG, POS), while others give numerical values which allow measuring sentiment intensity in addition to sentiment orientation. When it comes to the German language, sentiment resources are relatively small in size with regard to the number of words they encompass, mostly containing up to a few thousand words [18].

Although there are websites[1] listing different databases, data sets and sentiment dictionaries for German sentiment analysis, none of them are exhaustive. Currently, there is no central, comprehensive go-to online resource for German sentiment analysis.

Our paper aims to collect and compare the available resources and map out the landscape of German polarity dictionaries for sentiment analysis. To this end, we analyse the similarities between the 15 sentiment dictionaries identified by our search by means of a divisive clustering method.

The results show that while most dictionaries seem to make relatively comparable predictions about a word's sentiment, distinct subgroups can be identified, which are partially determined by how the dictionaries were developed. First, we find that dictionaries with categorical sentiment labels tend to behave similarly. More interestingly, however, we also identify groups of similar sentiment dictionaries that depend on each other by design. Based on our observations, we argue that newly compiled polarity dictionaries should be based on either diverse extant resources or newly created sentiment annotations to avoid error propagation and the amplification of potentially present biases.

Our paper is structured as follows. First, we describe the dictionaries analysed in this paper and the preprocessing steps that we applied. We then briefly describe the clustering approach applied to our data set and present the results of our analysis. Finally, we discuss and critically assess our findings.

2 Data and Preprocessing

For this research, we thoroughly combed the web for German-language polarity dictionaries that contain polarity ratings and identified 15 resources in total. The search was limited to already existing sentiment dictionaries that were freely available on the internet for academic and non-commercial purposes and contained a polarity rating of words. Databases containing annotated data that could theoretically be used to create a sentiment dictionary were not considered. The dictionaries vary considerably in their development processes and methods. Table 1 gives an overview over the identified resources.

With regard to the scope of this research, we focus on the dimension of polarity to include and analyse as many resources as possible, although we are aware of the relevance of a multidimensional approach to emotions, in particular in psychological research. Consequently, most of the sentiment resources included in our analysis contain more than one sentiment dimension. AffDict [36] and AffMeaning [2] contain sentiment ratings on potency (strong vs. weak) and activity (calm vs. lively) in addition to evaluation (good vs. bad, i.e. polarity). With these dimensions, AffDict and AffMeaning adhere to the dimensional view of Osgood, Suci and Tannenbaum with evaluation, potency and activity constituting an affective space in language [25]. AffNorms [18] includes four psycholinguistic attributes: abtractness/concreteness, arousal, imageability and valence (i.e. polarity). BAWL-R contains ratings on

[1] e.g. `https://sites.google.com/site/iggsahome`

imageability, arousal and valence, as well as linguistic properties of words that may influence their perception [40]. LANG [13] and Wordnorms [19] have ratings on valence, arousal and concreteness.

The authors who created the Morph resource [33] point out that in their experiments on sentiment classification using the full set of dimensions yields higher prediction accuracy than using just one dimension, but if one single dimension is used, polarity ratings are most predictive. Thus it seems that polarity can serve as a reasonable proxy in cases as ours where it is not possible to include the full range of dimensions in the analysis.

The dictionaries also vary considerably in the methods that were used to create them. In five cases, sentiment ratings were collected from human annotators: AffDict [36], AffMeaning [2], BAWL-R [40], LANG [13] and Wordnorms [19]. Other dictionaries rely on already existing resources. The Polarity Clues were created by translating existing English sentiment resources and enriching them with synonyms [41]. SentiMerge was created by simultaneously combining polarity scores from several extant sentiment dictionaries using a Bayesian probabilistic model [11]. AffNorms [18] used three already existing German sentiment resources as training data to automatically infer polarity values for over 350 000 words using a supervised machine learning algorithm following Turney et al. [37]. SentiWS [29] is based on automatically translated entries of the General Inquirer, a German collocation dictionary and collocation analysis, using pointwise mutual information (PMI) to assign polarity weights, following the approach by Turney and Littmann [38]. EmotionDict [16] uses already existing German sentiment resources (among them the Polarity Clues [41] and SentiWS [29]), enriching them with synonyms. SePL [31] was created by extracting opinion-bearing phrases of reviews and using the star ratings to infer opinion values. Morph [33] used the Polart lexicon [15] and added words from various databases to infer the sentiment values of German compound words based on their morphological structure.

ANGST pursues a mixed strategy by using the valence, arousal and imageability ratings of the BAWL-R [40], and supplementing them with ratings in dimensions for additional words and with an additional dimension, dominance.

For the ALPIN dictionary, human annotators rated text snippets from the Austrian Media Corpus [28] as positive or negative in a crowd-sourcing survey. The sentiment value of a word was then determined by the number of negative and positive texts it appears in, as proposed by [1].

As Table 1 shows there is a lot of heterogeneity among the dictionaries. Basic preprocessing steps were applied to normalise the dictionaries in order to merge, analyse and compare them. The steps of the preprocessing depend on the specific dictionary's structure. For a detailed account of the preprocessing applied to each dictionary, refer to Appendix A.

The sentiment values in AffDict [36], AffMeaning [2], AffNorms [18], ANGST [35], BAWL-R [40], LANG [13], SentiMerge [11] and Wordnorms [19] had to be rescaled to the interval [−1,1], i.e., going from maximally negative to maximally positive.

Table 1 Overview of the 15 German polarity resources included in this analysis.

Label	Reference	Description	Method	Size	Scale	Purpose
AffDict	Schröder, 2011 [36]	3 dimensions: Evaluation (good–bad), potency (strong–weak), activity (lively–calm)	1 905 human raters	1 100 total, 376 nouns, 393 verbs, 331 adjectives, 128 combinations	[−4,4]	Modelling impression formation
Aff-Meaning	Ambrasat et al., 2014 [2]	3 dimensions: Evaluation (good–bad), potency (strong–weak), activity (lively–calm); concepts related to authority and community	2 849 human raters	909 words from semantic fields of authority and community	averaged 9-point semantic differential scale	investigate intra-societal consensus and variation in affective meanings
AffNorms	Köper & Schulte im Walde, 2016 [18]	Ratings on 4 psycho-linguistic affective attributes	Supervised Machine Learning	350 000 nouns, verbs and adjectives	[0,10]	Ratings for sentiment analysis
ALPIN	Kolb et al., 2021	Austrian words from the political and media context	Machine learning on human-rated text snippets	3 636 nouns, verbs and adjectives	[−1,1]	Sentiment analysis on Austrian words
ANGST	Schmidtke et al., 2014 [35]	Words of the ANEW rated on 6 dimensions	human raters	1 003 nouns, verbs and adjectives	[−3,3]	Affective word lists for experiments
BAWL-R	Võ et al., 2009 [40]	Extension of BAWL with 3 psycholinguistic indices	200 human raters	>2 900	[−3,3]	Affective word lists for experiments
Emotion-Dict	Klinger, Suliya & Reiter, 2016 [16]	Sentiment labels for 7 basic emotions	3 human raters	4101 nouns, verbs and adjectives	POS, NEG	Emotions detection in literary texts
Polarity Clues	Waltinger, 2010 [41]	Translated English polarity features	Translation of English sentiment resources	10 141 nouns, verbs and adjectives	NEG, NEU, POS	Sentiment analysis
LANG	Kanske & Kotz, 2010 [13]	Ratings for emotional valence, arousal and concreteness	64 human rater, 2 years apart	1 000 nouns	[0,10]	Acquire test-retest reliability for ratings

Table 1 continued from previous page

Label	Reference	Description	Method	Size	Scale	Purpose
Morph	Ruppenhofer & Wiegand, 2017 [33]	Sentiment ratings for rare and complex compounds	Human rated baseline, rule-based classifier	8 400 in total from several samples	NEG, NEU, POS	Approach the problem of coverage in German polarity resources
PolArt	Klenner, Fahrni & Petrakis, 2009 [15]	Lexical resource for German sentiment analysis	Semi-automatic translation approach	10 790	POS, NEU, NEG, SHI, INT	Ratings for sentiment analysis
SePL	Rill et al., 2012 [31]	Opinion bearing words and phrases for German	Infer sentiment value from review rating	Adjectives, nouns, as well as adjective and noun-based phrases	[−1,1]	Inclusion of intensifiers, reducers and negation words
SentiMerge	Emerson & Declerck, 2014 [11]	Combine polarity scores from several data sources and estimate the quality of each source	Normalising and merging with Bayesian framework	98 918 words from Polarity Clues, SentiWS and Polart	[−1.7,1.7]	Ratings for sentiment analysis
SentiWS	Remus, Quasthoff & Heyer, 2010 [29]	Affect dictionary with syntactic category, inflectional forms, polarity and strength	Semi-supervised Machine Learning: Pointwise Mutual Information (PMI)	3468 nouns, verbs, adverbs, adjectives, 1650 negative, 1 818 positive	[−1,1]	Ratings for sentiment analysis
WordNorms	Lahl et al., 2009 [19]	Ratings on 3 psycho-linguistic attributes: concreteness, valence, and arousal	3907 human raters, crowdsourced via webapp	2654 nouns	[0,10]	Ratings for sentiment analysis

The part-of-speech (PoS) tagging in the individual dictionaries is very inconsistent and not all dictionaries provide it. The largest dictionary with 350 000 entries does not provide PoS labels which means that for the vast majority of words our analysis, there is no part-of-speech information is available to begin with. PoS information is consequently not considered and was removed during preprocessing.

In cases of dictionaries with discrete categories ("negative", "positive", "neutral"), labels were replaced with numerical values to allow quantitative analyses on the dictionaries. To this end, the dictionaries with numerical sentiment values were merged first and two separate means were calculated for positive and negative values. The mean of the negative numerical sentiment values ($mean = -0.228$) was then imputed for words that were labeled "negative". The same was done, *mutatis mutandis*, for the words labeled "positive" ($mean = 0.176$).

Figure 1 shows the distribution of mean sentiment values for all words in the merged sentiment dictionary. As can be seen, most words have a sentiment value close to zero, indicating neutral polarity.

Most of the dictionaries range from a few thousand to ten thousand words, as is reflected by the median length of the dictionaries ($median = 3702$). Note that these values relate to the dictionaries *after* preprocessing and cleaning.

As the difference between the smallest with about a hundred and largest dictionary with over 350 000 words is considerable, most words are covered by only one dictionary. Consequently, the data set is relatively sparse. Only around 55 000 words appear in two or more dictionaries, and not a single word is included in all 15 dictionaries. Note that the sparsity of our data set is mainly a reflex of the AffNorms [18] being more than three times larger than the second largest, SentiMerge [11].

The preprocessed dictionaries were finally merged into a large master dictionary comprising 15 dictionaries and polarity information for roughly 400 000 words. The entries consist mostly of single words, but there are also entries that consist of more than one word, since one of the dictionaries, SePL [31] is composed of short phrases and adverb-adjective combinations.

3 Analysis

All analyses as well as the preprocessing were done using `R`, version 4.0.3 [27] and a selection of R packages. In order to compare all dictionaries, we adopted a clustering approach, using the `cluster` package, version 2.1.0 [21].

In order to cluster the dictionaries, the data were arranged in a matrix with 15 rows and roughly 400 000 columns. Next, a distance matrix was calculated based on Euclidean distance. As the sentiment values are already scaled to the interval $[-1,1]$ after the preprocessing, further standardization was not required. We opted for Euclidean distance since this distance measure (as any Minkovski-type distance) is more sensitive to distributional differences than, for example, correlation dissimilarity. Thus, we can more accurately compare and find differences between, for example, dictionaries with a relatively centered distribution of sentiment scores on the one hand and more dispersed dictionaries on the other hand. Importantly, for each pair of dictionaries the distance measure was only based on the set of overlapping words contained in both dictionaries.

Clustering is an unsupervised technique used to group objects which are close to each other in a multidimensional feature space to uncover inherent structures in the data [4]. Optimally, the objects in the same cluster show a high degree of similarity while being as dissimilar as possible from objects belonging to different clusters [14]. There are different algorithms to achieve this. One main distinction can be made between partitioning and

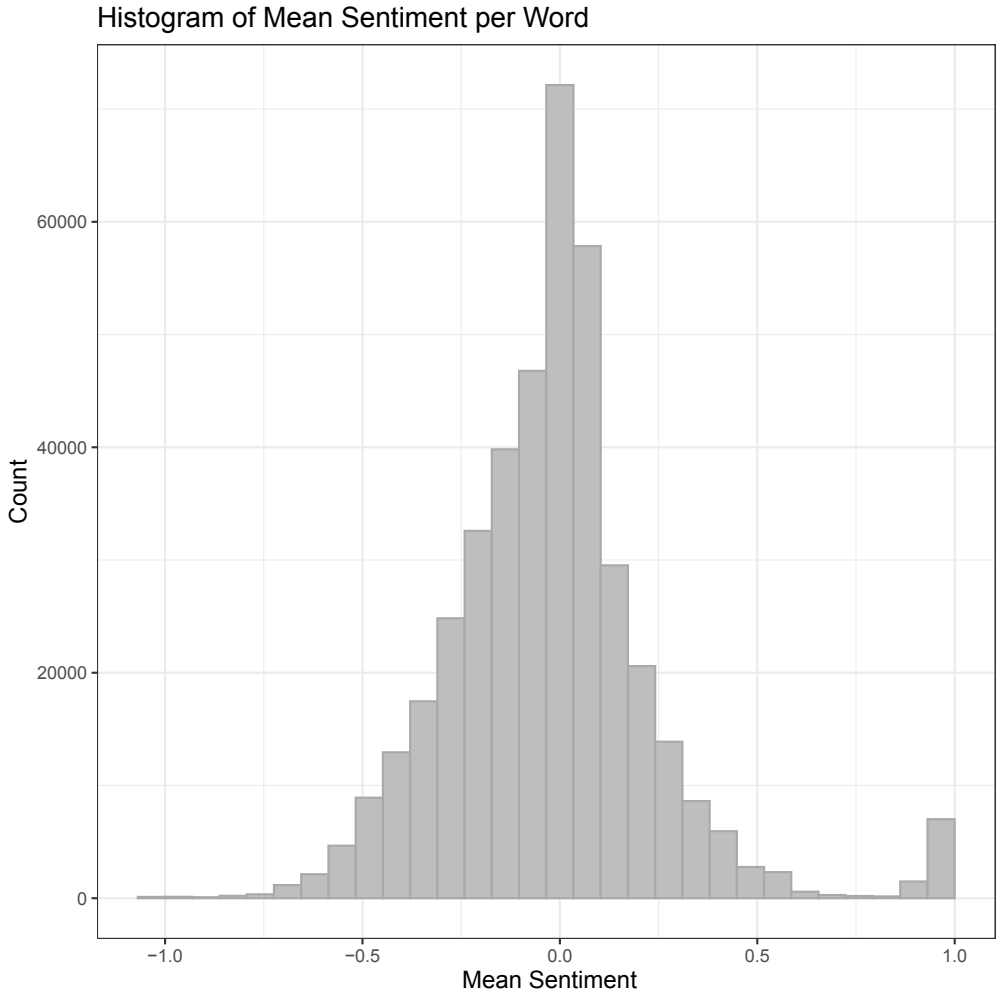

Figure 1 Histogram showing the mean sentiment value per word across all merged dictionaries.

hierarchical methods. Partitioning methods construct a predefined number of k clusters. Hierarchical methods do not construct a single partition with k clusters, but output the situation for $k = 1$ cluster to $k = n$ and all values of k in between [14].

There is no clear consensus about which algorithm is best [3] and cluster validation is a difficult task, as it lacks a common theoretical background and clear-cut best practices or rules [3]. Several cluster validation indices exist, but previous studies have shown that no single index is able to outperform the rest [7, 22, 23]. Further, the performance of the used evaluation criteria depends on the data [23]. All these factors make it difficult to determine the optimal parameters for the cluster analysis at hand.

To identify the parameters that work best for the given data and to assess cluster stability, different cluster algorithms and cluster definition methods were evaluated using the `clValid` package [4]. Cluster evaluation indices can be roughly categorized into external and internal measures. External validation measures rely on an outside data source with known class labels that serve as benchmark data [12]. Such a gold standard does not exist in many cases. In particular, there are no benchmarks concerning a word's "true" sentiment value. We thus relied on internal cluster validation measures that use the clustering and the underlying data set to assess the quality of the clustering. Three internal validation measures concerning the compactness, connectedness and separation of the clusters can be calculated with `clValid` [4].

The package provides a function to facilitate permutating different distance metrics and cluster methods, allowing to assess the robustness of the identified cluster solution. Rank Aggregation, as supported by `RankAggreg` package [26], was used to summarise the results in a super-list with the top three winning cluster algorithms plus optimal k, ranked by how much they maximise connectivity [4] and silhouette width [32] and minimise the Dunn Index [9].

In the present analysis, three different cluster algorithms are compared: Agglomerative nesting, partitioning around medoids and divisive analysis. The influence of different linkage methods is also evaluated, as well as different values of k from 1 to 5.

Divisive analysis is a hierarchical clustering method that starts out with a single cluster containing all objects and works bottom-up. In each step, the object that is most dissimilar to all other objects is identified and separated into a splinter group. All other objects are either assigned to the new splinter group or remain in their original cluster, depending on their similarity. In each iteration, the cluster with the largest diameter is selected and one object is separated until there are $k = n$ clusters.

Agglomerative nesting works on a reverse logic. At the beginning, each objects starts out as an individual cluster. In the first step, the two most similar objects are fused into one cluster. All distances are recalculated, and the process is repeated until all objects form a single, large cluster. An important parameter is the linkage method that determines the similarity between two objects. Several linkage methods exist, three commonly used ones are average linkage, complete linkage and Ward linkage. Complete linkage merges two clusters with the smallest maximum distance between them. Average linkage fuses clusters with the smallest average distance between them. Ward's method merges the two clusters that provide the smallest increase in within-cluster variance.

Partitioning around medoids clustering requires a predefined number k of clusters that the user wants to extract. The algorithm then selects k representative objects in the data. The clusters are formed by assigning each remaining object to the nearest representative object, the medoid [14]. The average distance (or dissimilarity) of the representative object to all objects of the same cluster is minimised. The principle is similar to k-means clustering which aims to minimise the average distance, making it susceptible to outliers. In this regard, partitioning around medoids is the more robust method.

The divisive analysis algorithm with three clusters yielded the best outcome for the three quality metrics. Consequently, we conducted divisive clustering with $k = 3$ clusters. A more detailed account of the employed methods and cluster evaluation can be found in the supplementary materials (`https://phaidra.univie.ac.at/o:1169856`).

4 Discussion

The cluster dendrogram reveals interesting insights. First of all, the cluster analysis suggests that most dictionaries are relatively similar to each other, as they form a single, large group in the dendrogram (left-most cluster in Figure 2). We will take a closer look at the internal structure of this group before we discuss the two dictionaries representing outliers (SePL and Polart on the right in Figure 2).

First, it is noteworthy that some dictionaries in the large cluster were created by extending or building on already existing ones. ANGST [35] uses the ratings of valence, arousal, and imageability of BAWL-R [40] as a basis and extends them with ratings on dominance and potency, additional words and new arousal ratings.

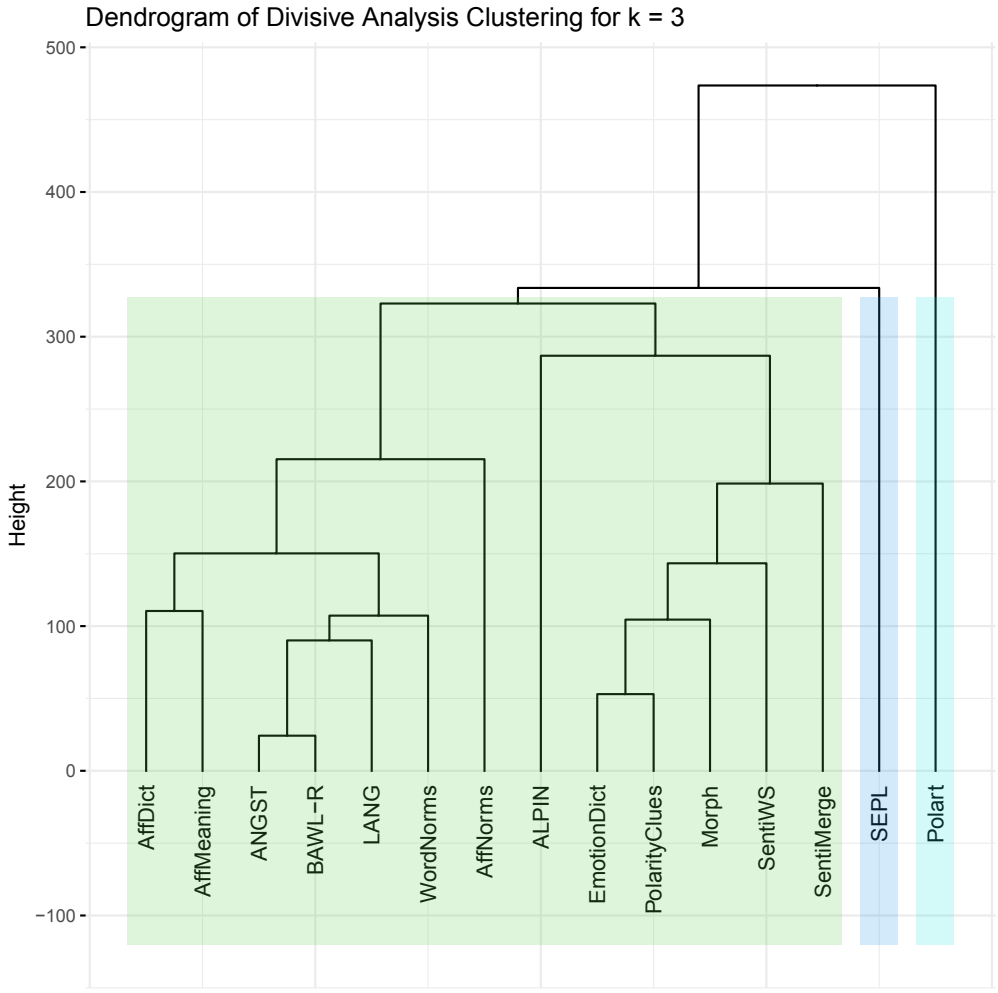

Figure 2 Dendrogram of divisive clustering using Euclidean distances. The $k = 3$ clusters are highlighted.

Thus, it is not surprising that these two dictionaries are highly similar to each other and were separated into different clusters only in the last iteration of the divisive algorithm.

Interestingly, ANGST [35], LANG [13] and BAWL-R [40] share a common methodological feature: they all used self-assessment manikins [20] for collecting the sentiment ratings. As can be seen in Figure 2, the three dictionaries have a relatively high similarity and were separated at a late step in the divisive clustering process. This highlights the role of the data collection procedure in compiling sentiment dictionaries.

In a similar way, there are dependencies between other dictionaries as well. AffDict [36] and AffMeaning [2] were not created for the purpose of conducting sentiment analyses, they were the result of two studies in social psychology. In both instances, survey participants were asked to rate words on the same three dimensions: on evaluation (good vs. bad), potency (strong vs. weak), and activity (lively vs. calm). AffDict [36] was used to model impression formation. AffMeaning [2] was used to examine intra-societal consensus and variation in affective meanings of concepts related to authority and community. The author of the AffDict paper [36] also collaborated on the paper on AffMeaning [2], the other authors on that paper

appear to be lab colleagues. While the research topics are different, the general methodology seems be rather similar and may explain why these two dictionaries ended up in the same cluster.

The very extensive AffNorms (with over 350 000 words) made use of BAWL-R [40], Wordnorms [19] and LANG [13] as training data for a supervised machine learning algorithm. This allows to automatically generate sentiment values for large amounts of words. AffNorms appears in the same cluster as its three seed dictionaries, but somewhat removed from them.

AffDict [36], AffMeaning [2], BAWL-R [40], LANG [13] and Wordnorms [19] are clustered quite closely to each other. As pointed out above, they share a similar development process that involved data collection from human annotators. EmotionDict [16] used the Polarity Clues [41] as a resource to build upon and was semi-automatically enriched with synonyms. Consequently, the words in these two dictionaries are expected to have largely identical sentiment labels. Since a constant was used to impute numerical sentiment values for the sentiment labels, this similarity persists in the cluster analysis.

Morph was created in an attempt to model the polarity of low-frequency complex German compound words based on their morphological composition. Its base data set is sampled out of the Polart lexicon [15]. In addition, the authors added words from the CELEX database and used additional compound words from Wegwarte, an online collection of German neologisms and Wiktionary. All these words are included in our analysis, as well as the *test-train* set and *dev* set that were used by Ruppenhofer, Steiner and Wiegand [33] to evaluate their approach.

ALPIN (Austrian Language Polarity in Newspapers) [17] was developed in the framework of the DYSEN project[2]. It is the only dictionary that is specific to Austria. The labeled text data that was used for creating it stems from Austrian newspapers and contains, *inter alia*, German words specific to Austria. Its relation to Austria sets it apart, and it is was developed independently of already existing resources. Unsurprisingly, it was separated rather early in the clustering process, indicating that is rather different from the other dictionaries within the large cluster on the left-hand side of the dendrogram in Figure 2.

SentiMerge [11] was created based on a Bayesian probabilistic model and combines polarity values from the Polart lexicon [15], SentiWS [29], Polarity Clues [41] and the German SentiSpin dictionary [42]. The latter resource was not included in the analysis at hand as it was not accessible and the author was not reachable. In the cluster dendrogram, SentiMerge appears in close proximity to two of its constituents, SentiWS [29] and Polarity Clues [41]. Interestingly, the Polart lexicon is very distant from all other German polarity resources and forms a cluster of its own (see below).

The Sentiment Phrase List (SePL) [31] was the second dictionary in the clustering process to initiate a splinter group and is thus quite dissimilar from the remaining dictionaries. This does not come as a surprise, as it possesses some unique features that set it apart. First, it was created based on product reviews accompanied with one to five-star ratings. Second, it contains not only single words, but short phrases like *absoluter Mist* ("absolute rubbish"). It can thus be expected to have a small overlap with the other dictionaries.

The Polart lexicon forms its own cluster and was the first object to be separated into a splinter group during the iterative clustering process. This is surprising, as some dictionaries, like Morph, used the Polart lexicon as a seed dictionary. The dissimilarity to the other dictionaries might be attributed to the interesting structure of Polart. It provides categorical labels that indicate sentiment orientation as well as numerical values that indicate sentiment

[2] *Dynamic Sentiment Analysis as Emotional Compass for the Digital Media Landscape*; more information on the DYSEN project can be found here: `https://www.oeaw.ac.at/acdh/projects/dysen/`.

intensity. The sentiment intensity, however, can take seven different, discrete values: 0, 0.3, 0.5 and 0.7, as well as their negatives. This sets it apart from the other dictionaries: It is less fine-grained than the dictionaries that contain continuous sentiment values, but more fine-grained than the dictionaries that only provide sentiment orientation in two or three categories and that had to be imputed with the positive and negative mean sentiment value calculated from the numerical sentiment dictionaries. Thus, Polart being an outgroup in the dendrogram may be a reflex of this scaling in combination with the use of (distributionally sensitive) Euclidean distance for clustering dictionaries.

5 Conclusion and Outlook

In this paper, we present an overview of a lion's share of the German sentiment dictionaries that are currently available. It becomes evident that polarity resources are very heterogeneous both in terms of how they are generated and their structure. Although it is reassuring to see that most of them share similarities as to how words are rated, we also see that some of them form subgroups consisting of dictionaries that depend on each other.

These dependencies are an immediate consequence of the compilation procedure. First, some dictionaries are direct extensions of others. Second, extant dictionaries are often used to evaluate new dictionaries. This can be potentially problematic: if new dictionaries are only tested against extant resources that are already related, this may in the worst case amplify built-in biases and propagate labeling errors. We thus recommend using diverse polarity resources both for the evaluation of new sentiment dictionaries as well as, more generally, for testing and evaluating sentiment-analysis algorithms.

The "dictionary of dictionaries" we assembled during the research process is publicly available for further research and can be accessed on the Gitlab repository for this paper[3] or on Github[4].

This resource is not meant as a ready-to-use tool for sentiment analysis. It is rather a by-product of our research process and made available to encourage and facilitate further research on German polarity resources for sentiment analysis. Numerous compelling research question might be investigated with the help of our dictionary of dictionaries that the scope of our paper did not touch upon.

For one, we did not compare the performance of individual or subgroups of resources with each other. Recent research has shown that side-by-side performance comparisons of off-the-shelf sentiment resources can give fruitful insights into their reliability and validity [5].

Secondly, we focused on polarity ratings and discarded other sentiment dimensions if they were available. This was done to facilitate comparisons of the highly heterogeneous resources. It may be worthwhile for future research to evaluate the benefits of a multi-dimensional approach in sentiment analysis. Thirdly, our aim was not to create an integrated dictionary, but to bring the available dictionaries into a comparable format. The authors of SentiMerge [11] propose a bayesian framework for dictionary integration to deal with differences between individual dictionaries via statistical modeling. The dictionary assembled by us opens up a convenient and comprehensive framework to test and apply such and similar approaches in future research.

And finally, while we note that the dependencies among German polarity resources may be problematic with regard to bias propagation, we do not evaluate or quantify potential

[3] https://gitlab.com/acdh-oeaw/dysen/dysen-ldk2021
[4] https://github.com/bettina-mj-kern/LDK_2021

bias in any way, as this is beyond the scope of this paper. The detection of bias is, however, undoubtedly an important issue in sentiment analysis and requires further research with regard to how to identify and remedy biases in sentiment tools. Recent research indicates that the validity of sentiment resources is in many cases questionable [39]. Moreover, there are reasonable doubts about whether sentiment dictionaries should be applied outside the domain or even the intended use case for which they were developed [30]. During the initial emergence of sentiment analysis, the development focus was primarily on the scalability of the tools, on their ability to harness large amounts of text in an automated fashion and draw information from them. As the field advances and matures, validity, reliability and risk of bias emerge as relevant areas of research with the aim to attain more robust and more fine-grained results that accurately and reliably capture the sentiment content in a text. Our study provides only a piece of the mosaic and hopefully gives rise to further research.

References

1 Sattam Almatarneh and Pablo Gamallo. Automatic construction of domain-specific sentiment lexicons for polarity classification. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017*, pages 175–182. Springer International Publishing, 2018. doi:10.1007/978-3-319-61578-3_17.

2 Jens Ambrasat, Christian von Scheve, Markus Conrad, Gesche Schauenburg, and Tobias Schröder. Consensus and stratification in the affective meaning of human sociality. *Proceedings of the National Academy of Sciences*, 111(22):8001–8006, 2014.

3 Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Inigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013. doi:10.1016/j.patcog.2012.07.021.

4 Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta. clValid: An R package for cluster validation. *Journal of Statistical Software*, 25(4):1–22, 2008. URL: http://www.jstatsoft.org/v25/i04/.

5 Chung-hong Chan, Joseph Bajjalieh, Loretta Auvil, Hartmut Wessler, Scott Althaus, Kasper Welbers, Wouter van Atteveldt, and Marc Jungblut. Four best practices for measuring news sentiment using 'off-the-shelf'dictionaries: a large-scale p-hacking experiment. *Computational Communication Research*, 3(1):1–27, 2021. doi:10.5117/CCR2021.1.001.CHAN.

6 Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, 2017.

7 Evgenia Dimitriadou, Sara Dolničar, and Andreas Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1):137–159, 2002.

8 Elizabeth Duffy. Emotion: an example of the need for reorientation in psychology. *Psychological Review*, 41(2):184, 1934. doi:10.1037/h0074603.

9 Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974. doi:10.1080/01969727408546059.

10 Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

11 Guy Emerson and Thierry Declerck. Sentimerge: Combining sentiment lexicons in a bayesian framework. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 30–38, 2014.

12 Julia Handl, Joshua Knowles, and Douglas B Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005. doi:10.1093/bioinformatics/bti517.

13 Philipp Kanske and Sonja A Kotz. Leipzig affective norms for german: A reliability study. *Behavior research methods*, 42(4):987–991, 2010. doi:10.3758/BRM.42.4.987.

14 Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

15 Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. Polart: A robust tool for sentiment analysis. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 235–238, 2009.

16 Roman Klinger, Surayya Samat Suliya, and Nils Reiter. Automatic emotion detection for quantitative literary studies. a case study based on franz kafka's "das schloss" and "amerika". *Proceedings of the Digital Humanities*, 2016.

17 Thomas Kolb, Katharina Sekanina, Andreas Baumann, and Julia Neidhardt. Austrian language polarity in newspapers (ALPIN). Dataset, v1.0. URL: `https://phaidra.univie.ac.at/o:1169855`.

18 Maximilian Köper and Sabine Schulte Im Walde. 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, 2016.

19 Olaf Lahl, Anja S Göritz, Reinhard Pietrowsky, and Jessica Rosenberg. 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, 2009. `doi:10.3758/BRM.41.1.13`.

20 Peter Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. *Technology in mental health care delivery systems*, pages 119–137, 1980.

21 Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2019.

22 Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 24(12):1650–1654, 2002. `doi:10.1109/TPAMI.2002.1114856`.

23 Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985. `doi:10.1007/BF02294245`.

24 Charles E Osgood. Dimensionality of the semantic space for communication via facial expressions. *Scandinavian journal of psychology*, 7(1):1–30, 1966. `doi:10.1111/j.1467-9450.1966.tb01334.x`.

25 Charles E Osgood, George J Suci, and Percy H Tannenbaum. 1957the measurement of meaning. *Urbana: University of Illinois Press*, 47, 1957.

26 Vasyl Pihur, Somnath Datta, and Susmita Datta. *RankAggreg: Weighted Rank Aggregation*, 2020. R package version 0.6.6. URL: `https://CRAN.R-project.org/package=RankAggreg`.

27 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL: `https://www.R-project.org/`.

28 Jutta Ransmayr, Karlheinz Mörth, and Matej Ďurčo. Ii. amc (austrian media corpus) – korpusbasierte forschungen zum österreichischen deutsch, 2017.

29 Robert Remus, Uwe Quasthoff, and Gerhard Heyer. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171, 2010.

30 Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016. `doi:10.1140/epjds/s13688-016-0085-1`.

31 Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V Zicari, and Nikolaos Korfiatis. A phrase-based opinion list for the german language. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS'2012)*, pages 305–313, 2012.

32 Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. `doi:10.1016/0377-0427(87)90125-7`.

33 Josef Ruppenhofer, Petra Steiner, and Michael Wiegand. Evaluating the morphological compositionality of polarity. In *Proceedings of the 11th international conference on Recent Advances in Natural Language Processing (RANLP'2017)*, pages 625—633, 2017.

34 James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

35 David S Schmidtke, Tobias Schröder, Arthur M Jacobs, and Markus Conrad. Angst: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior research methods*, 46(4):1108–1118, 2014. doi:10.3758/s13428-013-0426-y.

36 Tobias Schröder. A model of language-based impression formation and attribution among germans. *Journal of Language and Social Psychology*, 30(1):82–102, 2011. doi:10.1177/0261927X10387103.

37 Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, 2011.

38 Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003. doi:10.1145/944012.944013.

39 Wouter van Atteveldt, Mariken ACG van der Velden, and Mark Boukes. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, pages 1–20, 2021. doi:10.1080/19312458.2020.1869198.

40 Melissa LH Vo, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538, 2009. doi:10.3758/BRM.41.2.534.

41 Ulli Waltinger. German polarity clues: A lexical resource for german sentiment analysis. In *LREC*, pages 1638–1642. Citeseer, 2010.

42 Ulli Waltinger. Sentiment analysis reloaded-a comparative study on sentiment polarity identification combining machine learning and subjectivity features. In *WEBIST (1)*, pages 203–210, 2010.

43 Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2019. R package version 1.4.0. URL: https://CRAN.R-project.org/package=stringr.

A Preprocessing of the Sentiment Dictionaries

AffDict [36]. This dictionary contains word ratings three dimensions on social concepts, separately for men and women and summarised for both genders. For our purposes, only the word column and evaluation column with averaged ratings by both genders were identified as relevant. The other columns were dropped. The values are scaled [−4,4] and thus scaled to [−1,1]. The German umlauts (ae → ä, oe → ö and ue → ü) were changed manually inside the csv file, as there are words that contain these letter combinations, but are not umlauts (e.g. homosexuell) which makes it difficult to find a regular expression pattern. Placeholders like "versprechen *(etwas)*", "zanken *mit*", "lernen *von*" were removed with regular expressions to match the other dictionaries.

AffMeaning [2]. The AffMeaning was created while investigating the affective meaning of authority- / community-related concepts. 2849 participants rated 909 words on three dimensions. The ratings are averaged separately for men and women and for both genders. For our purposes, only the evaluation columns with ratings by both genders were identified as relevant, other columns were dropped. A 9-point semantic differential scale was used and averaged over participants. The values thus range from [1, 8] and were scaled to [−1,1] for

the purpose of our analysis. Placeholders like *jmd. auszeichnen* were removed with regular expressions.

AffNorms [18]. The resource represents the most extensive sentiment dictionary for German at the current time. It consists of 350 000 German lemmas with four psycho-linguistic attributes: abstractness, arousal, imageability and valence. For the purpose of this paper, only the valence scale was used. It reflects polarity scaled [0,10] and was rescaled to [−1,1].

ALPIN [17]. This dictionary is based on roughly 5000 labeled text snippets taken from the Austrian Media Corpus [28] that were labeled in a crowd-sourcing survey. Sentiment scores range in the interval [−1, 1]. Only words that surface in more than one snippet were included resulting in 4600 words in total before further cleaning. During preprocessing, abbreviations, urls, numbers and symbols were removed. For duplicate words with different sentiment score, the mean was taken.

ANGST [35]. The aim of the Affective Norms for German Sentiment Terms (ANGST) was to provide a German adaptation of the ANEW, the Affective Norms for English Words [20]). This corpus provides normative emotional ratings of pleasure, arousal and dominance for a large number of English words. BAWL-R was used as a starting groundwork. In the case of the valence, the ANEW words were translated into German and received ratings from BAWL-R if available. For words not in the BAWL-R, ratings on a bipolar scale ranging $[-3, 3]$ were collected from 65 participants. For our study, the word column and the valence column were extracted from the data set. The ratings were scaled to [−1,1]. Further preprocessing was not necessary.

BAWL-R [40]. The aim of BAWL-R is to help create stimulus material for experiments on affective verbal processing. It is a revision and extension of the previous version, BAWL: 700 new words and arousal ratings were added by surveying 200 Psychology students. BAWL-R contains ratings for imageability, arousal and valence for 2900 words, as well as standard deviations and meta data such as the number of letters, syllables and phonemes, bigram frequencies, number of orthographic neighbours and so on. For our purposes, the word column and the corresponding valence values were extracted. Nouns were capitalised using the `str_to_title()` function from the `stringr` package [43]. The valence ratings are ranged [−3,3] and were thus rescaled to [−1,1].

EmotionDict. Extension of sentiment analysis to literary texts. EmotionDict follows Ekman's definition of fundamental emotions [10]. This is of note because most other sentiment dictionaries follow a dimensional approach. Ekman's theory, in contrast, describes emotions as discrete categories, distinguishable by an individual's facial expression and physiological patterns, for example of the autonomic nervous system. EmotionDict consists of seven text files, containing words that reflect the seven basic emotions: anger (*Wut*), fear (*Angst*), enjoyment (*Freude*), sadness (*Trauer*), disgust (*Ekel*), contempt (*Verachtung*) and surprise (*Überraschung*). The surprise text file had to be excluded as it contains a mixture of words with different polarities. The remaining six text files were merged into a single word list. Words in the joy text file were assigned a positive numerical value and words in the other five text files received the negative numerical value, as described in the main text.

LANG. The Leipzig Affective Norms for German [13] was created to provide researchers with norms for experimental studies on verbal emotional processing. 1 000 short German nouns were rated twice by two independent samples two years apart to assess the retest reliability

across samples and time. The rating was done on a 9-point scale using self-assessment manikins. Only the ratings on valence were used for this paper. Originally ranging from 1 to 9, the ratings were rescaled to values between −1 and 1.

Morph [33]. This dictionary was created researching coverage problems of German sentiment dictionaries.The authors attempt to estimate the polarity of complex German compound words based on the polarity of their morphological composition. This resource was not meant as a tool for sentiment analysis, but merely presents the result of modelling a word's sentiment based on its constituents. It is therefore questionable how well Morph will perform as a sentiment prediction tool. The dictionary consists of five text files containing words and three sentiment labels (NEG, NEU and POS). One of them contains only affixes and was not included in the analysis. Additional information like word type or inflections were removed with regular expressions. Words with a negative sentiment label were assigned a numerical value as described in the text, and words with a neutral one 0.

PolArt [15]. The Polart lexicon contains negative, neutral and positive sentiment labels and fixed numerical values (0, 0.3, 0.5 and 0.7) that encode sentiment intensity. For negative labels, the sentiment value was reversed to negative by multiplying it with −1. The resource further includes shift words that reverse the polarity if neighbouring words, and intensifiers. For these words, we assigned a neutral sentiment value of 0. The label column was then dropped along with other unneeded columns.

Polarity Clues [41]. The Polarity Clues consist of three text files with negative, positive and neutral lemmas. Three other text files contain their inflected forms, but for our purpose, lemmas are sufficient. The word column and the sentiment labels were extracted for the analysis. The labels were then transformed into numerical values as described in the main text. The first 19 rows were dropped as they contained numbers and symbols that were not relevant for us.

SentiMerge [11]. The authors propose a framework for merging sentiment resources of different lengths and different scales. The authors demonstrate their method by merging the Polart lexicon, SentiWS, Polarity Clues and the German SentiSpin [42] that was also used to build the Polarity Clues. The words in the dictionary were all lowercase and thus transformed to title case in all instances that that had the "noun" part-of-speech tag, again making use of Hadley Wickham's `stringr` package [43]. 878 words, however, were tagged as "XY" and remained lowercase. 1805 words appeared between two and four times in the dictionary because they exhibited different part-of-speech tags which probably originated from the merging of different sentiment resources. We resolved this issue by taking the mean sentiment of these words and discarding the superfluous entries.The values were rescaled to [−1, 1].

SentiWS [29]. SentiWS contains 3 471 negative and positive words, their inflections, part-of-speech tags and a sentiment value between −1 and 1. The negative and positive words come in two different text files. After some light data cleaning and removal of unneeded columns, the two sets are combined into one. Further preprocessing was not necessary.

SePL [31]. The Sentiment Phrase List provides opinion values ranging [−1, 1] for words and short phrases, as well as standard deviations and standard errors. The relevant column

containing the opinion value was extracted from the data in the text file alongside with the corresponding words. Further preprocessing steps were not necessary.

Wordnorms [19]. The Wordnorms consist of 2654 nouns that were rated on concreteness, valence and arousal by a sizable sample of 3 907 participants via web application. The resulting sentiment dictionary contains standard deviations for the mean ratings as well as metadata, like the number of ratings each word received, the number of letters and the results of the cluster analysis. For our research interest, only the words and mean valence ratings were relevant. Words without a valence rating were dropped. The ratings ranged $[0, 5]$ and were consequently scaled to $[-1,1]$.