Feelings from the Past—
Adapting Affective Lexicons for Historical Emotion Analysis

Sven Buechel[1] **Johannes Hellrich**[2] **Udo Hahn**[1]

[1]Jena University Language & Information Engineering (JULIE) Lab
http://www.julielab.de

[2]Graduate School 'The Romantic Model'
http://www.modellromantik.uni-jena.de

Friedrich-Schiller-Universität Jena, Jena, Germany

Abstract

We describe a novel method for measuring affective language in historical texts by expanding an affective lexicon and jointly adapting it to prior language stages. We automatically construct a lexicon for word-emotion association of 18th and 19th century German which is then validated against expert ratings. Subsequently, this resource is used to identify distinct emotional patterns and trace long-term emotional trends in different genres of writing spanning several centuries.

1 Introduction

For more than a decade, computational linguists have endeavored to decode affective information[1] from textual documents, such as personal value judgments or emotional tone (Turney and Littman, 2003; Alm et al., 2005). Despite the achievements made so far, the majority of work in this area is limited in at least two ways. First, employing simple positive-negative polarity schemes fails to account for the diversity of affective reactions (Sander and Scherer, 2009) and, second, in contrast to the humanities where numerous contributions focus on emotion expression and elicitation (Corngold, 1998), very little work has been conducted by computational linguists to unravel affective information in historical sources.

Arguably the main problem here relates to the availability of language resources for detecting affect. Algorithms for measuring semantic polarity (positive vs. negative) or emotion typically rely on either annotated corpora, lexical resources (storing the affective meaning of individual words) or a combination of both (Liu, 2015). To ensure proper affect prediction, these resources must accurately represent the target domain but speakers of historical language stages (19th century and earlier) can no longer be recruited for data annotation. Prior work aiming to detect affect in historical text ignored this problem and relied on contemporary language resources instead (Acerbi et al., 2013; Bentley et al., 2014).

Using word embeddings, we tackle this problem by jointly adapting a contemporary affective lexicon to historical language and expanding it in size. Collecting ratings from historical language experts, we successfully validate our method against human judgment. In contrast to previous work based on the categorical notion of polarity (Cook and Stevenson, 2010), we employ the more expressive dimensional Valence-Arousal-Dominance (VAD; Bradley and Lang (1994)) model of affect, instead. As a proof of concept, we apply this method to a collection of historical German texts, the main corpus of the 'Deutsches Textarchiv' (DTA) [*German Text Archive*], in order to demonstrate the adequacy of our approach. Our data indicate that, at least for historical texts, academic writing and belles lettres, as well as respective subgenres, strongly differ in their use of affective language. Furthermore, we find statistically significant affect change patterns between 1740 and 1900 for these genres.

2 Related Work

Prior computational studies analyzing affect in non-contemporary text are very rare. To the best of our knowledge, the work by Acerbi et al. (2013) and Bentley et al. (2014) constitute the first of this kind. They construct a *literary* misery index by comparing frequency of joy-indicating *vs.* sadness-indicating words

[1]We here use *affect* as an umbrella term for both *semantic polarity* and *emotion*.

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 54–61, Osaka, Japan, December 11-17 2016.

in the *Google Books Ngram* corpus (see below) and find correlations with major socio-political events (such as WWII), as well as the annual U.S. *economic* misery index in the 20th century.

As stated above, most prior work focused on the bi-polar notion of semantic polarity, a rather simplified representation scheme given the richness of human affective states (a deficit increasingly recognized in sentiment analysis (Strapparava, 2016)). In contrast to this representationally restricted format, the VAD model of emotion (Bradley and Lang, 1994), which we employ here, is a well-established approach in psychology (Sander and Scherer, 2009) which also increasingly attracts interest in the NLP community (see among others Köper and Schulte im Walde (2016), Yu et al. (2016), and Wang et al. (2016)). It assumes that affective states can be characterized relative to three affective dimensions: *Valence* (corresponding to the concept of polarity), *Arousal* (the degree of calmness or excitement) and *Dominance* (the degree to which one feels in control of a social situation). Formally, the VAD dimensions span a three-dimensional real-valued space which is illustrated in Figure 1, the prediction of such values being a multi-way regression problem (Buechel and Hahn, 2016).

Thanks to the popularity of the VAD scheme in psychology, plenty of resources have already been developed for different languages. For English, the *Affective Norms of English Words* (ANEW; Bradley and Lang (1999)) incorporate 1,034 words paired with experimentally determined affective ratings using a 9-point scale for Valence, Arousal and Dominance, respectively (see Table 1 for an illustration of the structure of such a lexicon). Warriner et al. (2013) provided an extended version of this resource (14k entries) employing crowdsourcing. As far as German-language emotion lexicons are concerned, ANGST (Schmidtke et al., 2014) is arguably the most important one for NLP purposes—it was only recently constructed (comprising 1,003 lexical entries) and replicates ANEW's methodology very closely (see Köper and Schulte im Walde (2016) for a more complete overview of German VAD resources).

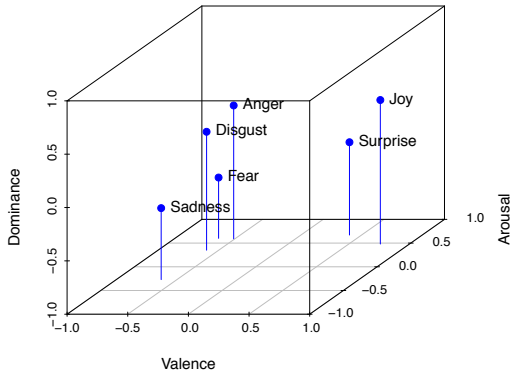

Figure 1: The three-dimensional space spanned by the VAD dimensions. For a more intuitive explanation, we display the position of six *basic emotions* (Ekman, 1992; Russell and Mehrabian, 1977).

As the manual creation of affective lexicons (polarity or VAD) is expensive, their automatic extension is an active field of research since many years (Turney and Littman, 2003; Rosenthal et al., 2015). Typically, unlabeled words are attributed affective values given a set of seed words with known affect association, as well as similarity scores between seed and unlabeled words. Concerning emotions in VAD representation, Bestgen (2008) presented an algorithm based upon a k-Nearest-Neighbor methodology which expands the original lexicon by a factor of 17 (Bestgen and Vincze, 2012). Cook and Stevenson (2010) were the first to induce a polarity lexicon for non-contemporary language from historical corpora by employing a pointwise mutual information (PMI) metric to determine word similarity and the much received algorithm by Turney and Littman (2003) for polarity induction. PMI is, like latent semantic analysis (LSA; Deerwester et al. (1990)), an early form of distributional semantics which, in the meantime, has been replaced by singular value decomposition with positive pointwise mutual information (SVD_{PPMI}; Levy et al. (2015)) and skip-gram negative sampling (SGNS; Mikolov et al. (2013)). Quite recently, evidence is available that the latter behaves more robust than the former (Hamilton et al., 2016).

Most prior studies covering long time spans (e.g., Acerbi et al. (2013)) rely on the *Google Books Ngram* corpus (GBN; Michel et al. (2011), Lin et al. (2012)). However, this corpus might be problematic for Digital Humanities research because of digitization artifacts and its opaque and unbalanced sampling (Pechenick et al., 2015; Koplenig, 2016). For German, we use the DTA[2] (Geyken, 2013; Jurish, 2013), which consists of books transcribed with double-keying and selected for their representativeness. The DTA aims for genre balance and provides a range of metadata for each document, e.g., authors, year, classification (like belles lettres and academic texts) and sub-classification (e.g., poem, biology, medicine).

[2]TCF version from May 11, 2016, available via `www.deutschestextarchiv.de/download`

3 Methods

Our methodology consists of two main parts. First, we adapt a contemporary VAD emotion lexicon to historical language and expand it jointly in size, and, second, we use this expanded lexicon to analyze emotions in historical language stages.

3.1 Inducing Historical VAD Lexicons

One of the most commonly used algorithms for affect lexicon induction was proposed a decade ago by Turney and Littman (2003) and put into practice for historical language by Cook and Stevenson (2010). Unfortunately, this procedure expects seed words of discrete polarity classes, a format we consider less informative for affective language analysis. For VAD vectors, we here employ the induction algorithm introduced by Bestgen (2008) instead. Bestgen's algorithm computes the affective score of the word w, $\bar{e}(w)$, given the set of the k nearest neighboring words to w from a seed lexicon, NEAREST(k, w), as

$$\bar{e}(w) := \frac{1}{k} \sum_{v \in \text{NEAREST}(k,w)} e(v) \tag{1}$$

where $e(v)$ is the emotion value of the word v, a three-dimensional VAD vector (see Table 1 and Figure 1 for illustration). We modify Bestgen's method by replacing LSA with SGNS for determining word similarity. In order to account for word-emotion association as present in historical language stages, we use word embeddings derived directly from the target language stage instead of contemporary ones. Seed values for the induction are taken from the contemporary ANGST lexicon (Schmidtke et al., 2014). This method results in a *hybrid* lexicon whose seed VAD values are empirically determined by contemporary speakers, whereas the similarity of words (and therefore the set of words taken into account when computing emotion values for words not in the seed lexicon) is determined from historical corpora. Although the emotion values computed in this way might be somewhat biased towards the contemporary language stage, such a hybrid lexicon should be more suitable for a historical analysis than lexicons with contemporary information only.

3.2 Measuring Textual Emotion

Building on an adapted lexicon, it is possible to (more) accurately determine the emotion values of historical texts. For this task, we use the Jena Emotion Analysis System[3] (JEMAS; Buechel and Hahn (2016)) since it has been (as one of the first tools for VAD prediction) thoroughly evaluated and is, to the best of our knowledge, currently the only tool for this purpose freely available. The lexicon-based approach it employs yields reasonable performance (Staiano and Guerini, 2014; Buechel and Hahn, 2016) and is easily adaptable to other domains by replacing the lexicon—a feature most valuable for historical applications as well. Basically,[4] it calculates the emotion value of a document d (a bag of words), $\bar{e}(d)$, as the weighted average of the emotion values of the words in d, $\bar{e}(w)$, as computed by Equation 1:

$$\bar{e}(d) := \frac{\sum_{w \in d} \lambda(w, d) \times \bar{e}(w)}{\sum_{w \in d} \lambda(w, d)} \tag{2}$$

where $\bar{e}(w)$ is defined as the vector representing a neutral emotion, if w is not covered by the lexicon, and λ denotes some term weighting function. Here, we use absolute term frequency as the resulting performance is among the best for automatically expanded lexicons (Buechel and Hahn, 2016).

4 Experiments

4.1 Gold Standard

One considerable difficulty concerning lexicons for historical language stages relates to their proper validation, since we lack native speakers for data annotation. Hence, to assess the quality of our results

[3] https://github.com/JULIELab/JEmAS

[4] For brevity, we only give a loose formal specification. See Buechel and Hahn (2016) for a more elaborated definition.

we constructed a small gold standard of 20 words annotated by seven doctoral students from various humanities fields. Their areas of expertise strongly overlap with the time periods covered by the slice of the DTA we are investigating. The instructions and rating scales follow the design of Warriner et al. (2013), yet with one crucial exception—subjects were requested to put themselves in the position of a person living between 1741 and 1900. We used such a wide temporal range since we expected different raters to ground their rating decisions on different time spans, varying with their historic expertise and acquaintance with a specific period. When averaging the different ratings, these biases should level off resulting in valid ratings relative to the entire time span. Our 20 stimulus words were randomly selected from words present within both the ANGST seed lexicon and the subset of the 1741–1900 DTA corpus, thus avoiding any "noisy" words such as annual figures. Table 1 provides some sample entries.

For comparison with existing resources, we measure inter-annotator agreement (IAA) by calculating the standard deviation between all given ratings for each word and dimension and then averaging these values for every VAD dimension (Average Standard Deviation; ASD). Our raters achieved an ASD of 1.61, 1.85, and 1.83 for Valence, Arousal, and Dominance, respectively. These IAA ratings are better than the ASDs reported by Warriner et al. (2013)—1.68, 2.30, and 2.16—suggesting that our experts are able to consistently rate non-contemporary word emotions.

4.2 Lexicon Expansion and Historical Adaptation

As mentioned before, we operate on the 1741–1900 part of the DTA. One text from this period written in Latin was excluded, leaving us with 1,022 texts. To ensure matches between this corpus and our VAD seed lexicon, we preprocessed ANGST (Schmidtke et al., 2014) with the CAB[5] lemmatization system used by the DTA (Jurish, 2013), without further filtering or modification of these entries. We then trained 200 dimensional SGNS embeddings[6] on this corpus.

Lemma	Valence	Arousal	Dominance
"Mutter" *(mother)*	2.00	-1.14	-1.29
"Erholung" *(recovery)*	0.86	-2.29	0.57
"giftig" *(poisonous)*	-2.29	1.86	-0.71
"Krise" *(crisis)*	-2.00	2.00	-0.86

Table 1: Sample entries from the historical gold standard relative to their empirically determined Valence-Arousal-Dominance (VAD) values.

We ran the modified version of Bestgen's expansion algorithm (see above) on these word embeddings using ANGST as seed lexicon. The k-parameter was determined by running the process for each integer $k \in [1, 50]$ measuring Pearson's r between original and induced values at each step. The correlation was highest for $k = 16$ ($r = 0.681$; average correlation over all three dimensions) which was thus employed to induce the final lexicon.

Our expanded and historically adapted lexicon comprises 143,677 word-emotion pairs. The correlation between these induced values and our historical gold standard amounts to $r = 0.75$, 0.64 and 0.56 for Valence, Arousal, and Dominance, respectively (the differences between the dimensions are consistent with prior work (Bestgen and Vincze, 2012)). Hence, our performance on historical data is even higher than the performance Bestgen and Vincze (2012) reported when using Bestgen's original algorithm to predict *contemporary* word emotions. We take this as a hint that our modifications (e.g., using SGNS instead of LSA) more than compensate for the additional difficulty of inducing *historical* word emotions.

4.3 Application to the DTA Historical Corpus

In order to demonstrate the potential of our approach for the Digital Humanities, we now examine the distribution of emotions in the DTA corpus relative to different categories of metadata. First, we take into account the genres of a document considering the whole study period. Second, we look at changes in emotions over time (also taking genre differences into account). Furthermore, of the three main genres distinguished within the DTA— belles lettres, academic texts and functional texts—we focus on the first

[5]Available via `www.deutschestextarchiv.de/demo/cab/`

[6]We used the PYTHON-based GENSIM implementation (accessible from `https://radimrehurek.com/gensim/`), with the following parameters: context window of up to 10 neighboring words, minimum word frequency of 10, negative sampling with 5 noise words and downsampling for words with a frequency of 10^{-3} or higher. We trained for 5 epochs, decreasing the learning rate from initial 0.025 down to 0.0001 in each of them.

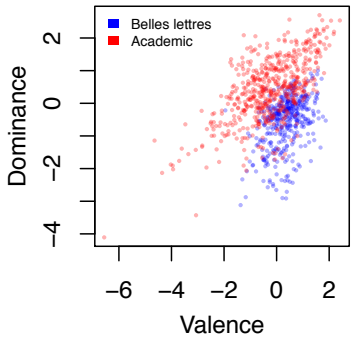

Figure 2: Distribution of two of the main document classes relative to Valence and Dominance.

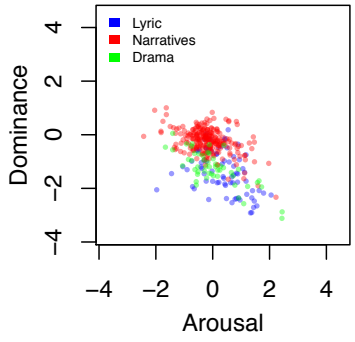

Figure 3: Distribution of subclasses of belles lettres relative to Arousal and Dominance.

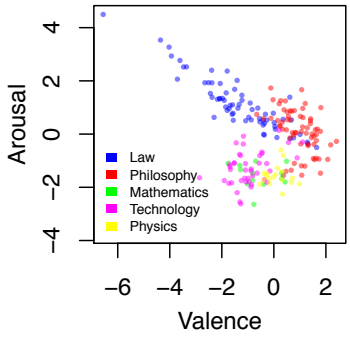

Figure 4: Distribution of five subclasses of academic texts relative to Arousal and Valence.

two because, upon inspection of the texts present in each category, they tend to be much more distinctively defined than the latter which, in our view, remains quite opaque.

For the following experiments, we processed the documents of the study period via JEMAS (see Section 3.2) employing the newly constructed historically adapted emotion lexicon. The VAD output of this system was standardized so that mean $M = 0$ and standard deviation $SD = 1$ for each dimension. Subsequently, we visualize our data with 2-D scatterplots each displaying two of the three VAD dimensions (Figures 2–4). Of the three possible plots (one for each pair of VAD dimensions) we here include the most illustrative ones for each comparison.

4.3.1 Distinction of Text Genres and Domains

Comparing belles lettres to academic texts, Figure 2 depicts their distribution relative to Valence and Dominance so that each data point relates to one document. The two genres are clearly separated[7] and their clusters show only little overlap. These observations suggest that the VAD values our system generates reflect the membership of a text for a certain genre. It may also indicate that our method is valid insofar as it catches some relevant intrinsic characteristics of the processed documents. To further illustrate the usefulness of our work for, e.g, literary studies or history of mind, we statistically tested whether these classes differ relative to the three emotional dimensions (using non-parametric tests; the data are, in general, not normally distributed) and give median (Md) values. In this setting, belles lettres display significantly higher Valence (Md $= 0.39$) and lower Dominance (Md $= -0.40$) than academic texts (Md $= -0.22$ and 0.42, respectively; $p < .05$ using a Mann-Whitney U test). This may reflect the technical nature of academic writing, e.g., explaining certain methodologies, and therefore expressing more control (which is closely related to Dominance).[8] Differences in Arousal were not significant.

Concerning the subgenres of belles lettres, we compared the predefined classes *lyric* and *drama*, and also *narratives*, a subclass we defined for this experiment subsuming different fine-grained distinctions between German terms for novels, novellas and tales. Again, a visual examination (see the Dominance-Arousal plot in Figure 3) reveals good separability. Dramas show lower Valence (Md = 0.00) than lyric and narratives (Md = 0.43 and 0.45, respectively), whereas lyric excels with high Arousal (Md = 0.43) contrary to narratives (Md = -0.22) and drama (Md $= -0.13$). Furthermore, narratives have a markedly higher Dominance (Md $= -0.19$) in contrast to lyric (Md $= -1.44$) and drama (Md $= -1.23$). The differences between the groups are significant relative to each dimension ($p < .05$; using the Kruskal-Wallis test, since we compare more than two groups).

Another striking distribution is depicted in Figure 4 which displays the relative positioning of five subclasses of academic texts, namely law, philosophy, mathematics, technology, and physics. Apparently, we come up with a clear (almost linear) separation between philosophy and law, on the one hand, and

[7] The notion of separability can be quantified as the performance of a classifier predicting the genre of a document given its VAD values. We ran these experiments in a pilot study finding good separability (almost 90% accuracy in this case) but exclude the details for brevity.

[8] The interpretations we offer in this section are meant as an illustration of how our quantitative data could be utilized within the (Digital) Humanities. We currently do not claim that these results can be taken for granted given our experimental data.

mathematics, physics and technology, on the other hand (thus empirically substantiating intuitions of different academic cultures dividing the sciences from the humanities (Kagan, 2009) in emotional terms). Also, the plot reveals more fine-grained features in line with common-sense intuitions about these study fields, e.g., parts of the philosophical texts are indistinguishable from law texts, while others show pronounced overlap with physics (possibly reflecting the impact of different subdisciplines, such as philosophy of law and philosophy of science). Also, physics and technology are fairly well set apart from each other, while mathematics seems to be equally similar to both. The qualitative fields display higher Valence and Arousal (Md = 0.26 and 0.54, respectively) than the quantitative ones (Md = -0.70 and -1.47). However, the sciences show higher Dominance than law and philosophy (Md = 1.12 as opposed to 0.53; all differences significant: $p < .05$ using a Mann-Whitney U test). Extending our interpretation concerning Figure 2, this may reflect the more technical nature of writings in the quantitative fields as opposed to the language-centered disciplines.

4.3.2 Shifts in Emotion over Time

We now turn to the question whether shifts in emotion can be traced in the texts of the DTA corpus over time. Again, we considered, first, all texts of the corpus, second, texts of the major academic class and, third, texts of the major class belles lettres. Due to data sparsity we did not take into account subclasses. We found clear evidence for long-duration shifts in emotion values considering the different groups. Performing linear regression, our data (quantified as the β-coefficient of linear regression models, i.e., the steepness of the regression line) suggest a specifically strong increase in Dominance concerning academic texts (possibly reflecting the establishment of a more technical style in scientific writing) and in the corpus as a whole, as well as a decrease of Arousal in belles lettres (possibly reflecting the shift from highly emotional sentimentalism via romanticism to rather descriptive realism (Watanabe-O'Kelly, 1997)). We summarize our findings concerning long-duration shifts in Table 2. These figures might seem rather small; recall, however, that the VAD values are normalized (given in SD) and that the documents we consider span 160 years so that, e.g., Arousal in belles lettres decreased by almost one SD (-0.96).

Lemma	Valence	Arousal	Dominance
Academic	-0.002*	0.000	0.003***
Belles lettres	0.001	-0.006***	0.001
All	-0.002*	-0.003***	0.004***

Table 2: β-coefficients of linear models predicting Valence, Arousal and Dominance (VAD), respectively, given a year. Levels of significance: * $p < .05$; ** $p < .01$; ***$p < .001$.

5 Conclusion

In this paper, we introduced a novel methodology for measuring emotion in non-contemporary texts by linking neural word embeddings derived from historical corpora, an adapted expansion algorithm for affective lexicons, and a lexicon-based method for emotion analysis. To demonstrate the potential of our approach for the Digital Humanities, we then conducted a study on emotional patterns within the DTA, a high-quality collection of historical German texts, using the multidimensional VAD model of emotion. This is the first application study of this kind, since prior studies on affect in historical texts were conducted with lexicons that were both non-specific for historical texts and less informative in terms of their affect representation scheme (Acerbi et al., 2013; Bentley et al., 2014).

We found evidence that different genres and subgenres of belles lettres and academic texts in the DTA show contrasting patterns in their emotional characteristics. Moreover, we identified pronounced long-term trends in textual emotions between 1741 and 1900. Both these observations can, though cautiously, be linked to explanatory patterns as discussed in the humanities (thus granting face validity to our findings).

We are interested in transferring these results to other languages, as well as conducting more fine grained temporal modeling, i.e., using multiple, temporally more specific lexicons for tracking emotional change. Future methodological work will focus on broadening the coverage of our gold standard as well as on the quality of induction algorithms for affective lexicons of historical language. Our induced historical word emotion lexicon in a format compatible with JEMAS and the gold standard are publicly available[9].

[9] https://github.com/JULIELab/HistEmo

Acknowledgements

This research was partly conducted within the Graduate School 'The Romantic Model' which is funded by grant GRK 2041/1 from the *Deutsche Forschungsgemeinschaft (DFG)*.

References

Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. The expression of emotions in 20th century books. *PLoS ONE*, 8(3):e59030.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *HLT-EMNLP 2005 — Proceedings of the Human Language Technology Conference & 2005 Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, 6-8 October 2005*, pages 579—586.

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLoS ONE*, 9(1):e83147.

Yves Bestgen and Nadja Vincze. 2012. Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, 44(4):998–1006.

Yves Bestgen. 2008. Building affective lexicons from specific corpora for automatic sentiment analysis. In *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, 26 May - June 1, 2008*, pages 496–500.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, University of Florida, Gainesville, FL.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Vol. 2: Long Papers. The Hague, The Netherlands, August 29 - September 2, 2016*, number 285 in Frontiers in Artificial Intelligence and Applications, pages 1114–1122.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation. La Valletta, Malta, May 17-23, 2010*, pages 28–34.

Stanley Corngold. 1998. *Complex Pleasure: Forms of Feeling in German Literature*. Stanford University Press.

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Alexander Geyken. 2013. Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften. Berlin, Germany, December 12-13, 2011*, pages 221–234.

William L. Hamilton, Jure Leskovec, and Daniel Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 1: Long Papers, pages 1489–1501.

Bryan Jurish. 2013. Canonicalizing the Deutsches Textarchiv. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften. Berlin, Germany, December 12-13, 2011*, pages 235–244.

Jerome Kagan. 2009. *The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century*. Cambridge University Press.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 2595–2598.

Alexander Koplenig. 2016. The impact of lacking metadata for the measurement of cultural and linguistic change using the GOOGLE NGRAM data sets: Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, and Slav Petrov. 2012. Syntactic annotations for the GOOGLE BOOKS NGRAM corpus. In *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, July 10, 2012*, volume System Demonstrations, pages 169–174.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR 2013 — Workshop Proceedings of the International Conference on Learning Representations. Scottsdale, Arizona, USA, May 2-4, 2013*.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the GOOGLE BOOKS corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One*, 10(10):e0137041.

Sara Rosenthal, Preslav I. Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SEMEVAL-2015 Task 10: Sentiment analysis in Twitter. In *SemEval-2015 — Proceedings of the 9th Workshop on Semantic Evaluation @ NAACL-HLT 2015. Denver, Colorado, USA, June 4-5, 2015*, pages 451–463.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

David Sander and Klaus R. Scherer, editors. 2009. *The Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press.

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.

Jacopo Staiano and Marco Guerini. 2014. DEPECHE MOOD: A lexicon for emotion analysis from crowd annotated news. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, June 22-27, 2014*, volume 2: Short Papers, pages 427–433.

Carlo Strapparava. 2016. Emotions and NLP: Future directions. In *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, page 180.

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

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 2: Short Papers, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Helen Watanabe-O'Kelly, editor. 1997. *The Cambridge History of German Literature*. Cambridge Univ. Press.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 540–545.