Reconstruction of historical sentiment-scores and their suitability as an indicator for word-diversification

Christian Obereder (11704936)

August 31, 2023

Supvervisor: MMag. Dr. Andreas Baumann
Co-Supervisor: Univ.Ass. PhD Gabor Recski

1 Abstract

As language changes over time, so does the meaning of individual words[HLJ16a]. For instance, a word may become more ambiguous or less ambiguous, as different meanings for that word become more or less popular. An example for this is the word 'freak', which, until recently, was mostly negatively connoted and referred to abnormal individuals. However, in recent decades, 'freak' is also used in a much more positive manner to refer to individuals that are enthusiastic about a topic.

This work uses valence-information associated with words as an indicator for ambiguity and, for that purpose, adapts a regression-based method for affect-lexicon-expansion for historical language stages. In particular, a regression model is trained to predict valence-scores from word-embeddings and this model is then applied to historical word-embeddings in order to 1. reconstruct valence-scores for historical language stages and 2. observe how these predicted valence-scores and their distribution shift over time and whether or not this is a suitable indicator for word-diversification.[1]

2 Problem statement

Measuring the ambiguity or level of diversification of a word over a long period of time (decades or even centuries) can be a challenging task. Not only is word-ambiguity difficult to define and model, but it is also not possible to rely on crowdsourcing to gather this information for past centuries. A possible approach to assess the level of diversification or ambiguity of a word may be to observe the valence associated with that word over time. For instance, when a word w has mostly negative connotations at a given language stage but negative as well as positive connotations at some later language stage, diversification of w may have occurred. One way of measuring such changes in sentiment is to use word-embeddings trained on texts from consecutive historical language stages, and to observe how the distribution of valence-scores of w's closest neighbors in the word-embedding-space changes over time.

However, while texts from many different time-periods exist and can be used to train word-embeddings, only contemporary valence-lexicons are available on a large scale. This is prob-

[1] Code available at: https://github.com/GitianOberhuber/Interdisciplinary-Project-2022W/

lematic for the procedure described above, as the valence-score associated with w (and its neighbors) e.g two centuries ago may strongly differ from its contemporary valence-score. Because of this, some way of reconstructing historical valence-values for individual words in a target language stage is needed.
[Li+17] describe an approach for automated expansion of an affect-lexicon, by training a regression model that takes word-embeddings as input and predicts their representation in a potentially multi-dimensional affect space (such as the VAD-affect-model [BL94]). In other words such a model learns how different dimensions in word-embeddings contribute to different affect-dimensions, for example to a valence-dimension. This model can then be applied to the word-embedding of a word for which the affect-lexicon has no entry in order to predict the affect-value of that word. While Li et al. use this approach to expand the affect-lexicon that was used for training of the regression model, this work applies the model to historical word-embeddings for the reconstruction of historical valence-scores. In order to measure how well the method proposed by Li et al. generalizes to foreign word-embeddings, the performance of the regression model is evaluated on a contemporary dataset different from the one used for training and on a small historical affect-lexicon.

Therefore, the aim of this project is to reconstruct historical valence-scores and their distributions in order to be able to better measure how the valence-values associated with a word develop over time.

3 Resources

The approach of training a regression model on contemporary word-embeddings and a contemporary valence-lexicon and then applying this model to historical word-embeddings for the reconstruction of historical valence-labels employed in this work and uses multiple external resources:

- Historical word-embeddings
- Contemporary word-embeddings for the training of a regression model
- Contemporary valence-lexion
- Contemporary word-embedding for regression model evaluation
- Historical valence-lexicon for for regression model evaluation

For each the points, pre-existing resources are used. The first three bullet points are used for the creation of the model and prediction of the historical valence-scores. The last two bullet points are resources that are only used for evaluation. This section describes all resources in detail.

3.1 Histwords - Historical Word-Embeddings

[HLJ16a] provide word2vec word-embeddings trained on data from past decades from the Google Ngram Books Dataset, which will be referred to as the "historical" word-embeddings throughout this report. They provide these embeddings for every decade from 1800 to 1990 and have aligned the embedding-spaces such that comparisons between embeddings of different decades become meaningful.
For all available decades, the embeddings are of size 100.000x300, meaning 300 dimensions

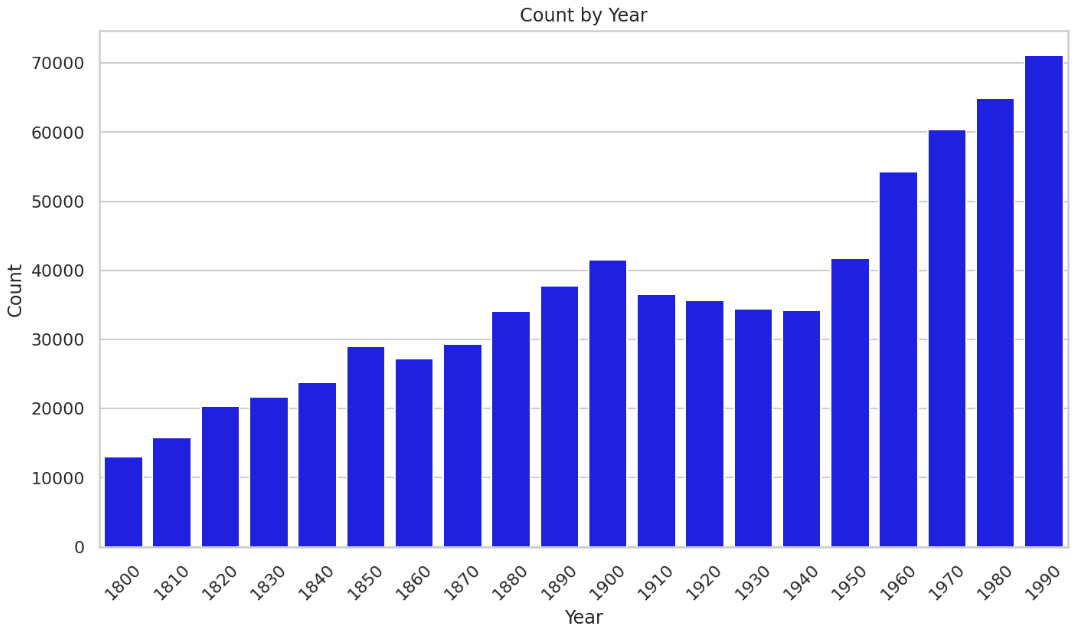

Figure 1: Count of words available throughout the decades depending on start-year for the Histwords word-embeddings

for, theoretically, 100.000 words. However, inspection of the data reveals that a large number of words have word-embeddings that are all zero (meaning they are effectively not present). Further reducing the size of the available data is the fact that for this work, only words that are available for all decades (from a given start-decade onwards) should be considered so that the change of a word over time can be observed without gaps. The number of words available depending on the start-decade is shown in Figure 1.
This resource is used in combination with a regression model that predicts valence-scores based on word-embeddings in order to create valence-scores for a given word in a historical language stage.

3.2 Geometry of Culture - Contemporary Word-Embeddings

[KTE19] provide word2vec word-embeddings trained on data from 2000-2012 from the Google Ngram Books Dataset, which will be referred to as the "contemporary" word-embeddings throughout this report. They have a size of 900.000x300 and contain no missing values. As such, almost all words that are available throughout the decades in the historical word-embeddings are also present here, meaning that the number of fully available words for a given start-decade shown in **??** is barely reduced when keeping only words that are also present in the contemporary word-embeddings. However, it is important to note that the contemporary word-embeddings are not aligned with the historical word-embeddings, meaning that in order to make comparisons between the two resources meaningful, alignment of the embedding-space of the contemporary word-embeddings to the embedding-space of the historical word-embeddings had to be performed. For this, the alignment method used by Hamilton et al. to align the historical word-embeddings, which is based on orthogonal Procrustes, is used. More information on this is given in Section 4.1.
This resource is used in combination with a contemporary valence-lexicon (XANEW) in order to create a regression model that predicts valence-scores based on word-embeddings. Li et al.

used word2vec word-embeddings which they trained on Wikipedia for this, however, for this work the Geometry of Culture embeddings were chosen, as the training data used for their creation should be more similar to that of the historical Histwords embeddings.

3.3 XANEW - Contemporary Affect Lexicon

[WKB13] provide an affect lexicon containing 13.915 English words and their representation in the VAD-model [BL94], created in 2013. The data was gathered using the Amazon Mechanical Turk crowdsourcing website, and each dimension was annotated separately (participants were not asked to judge a word on all three VAD-dimensions at once, but only one at a time). As such, it is currently one of the biggest available affect lexicons. For the purpose of this work, only the valence-dimension of the VAD-model is used. The valence-dimension represents the words valence on a 1 to 9 scale, with 1 being the smallest (most negative) and 9 being the highest valency (most positive), with a mean of 5.06 and a standard-deviation of 1.62. Since these valence-scores are used in combination with the contemporary word-embeddings to train a regression model, only those words in the lexicon that are also present in the contemporary word-embeddings are kept, slightly reducing the lexicon size to 13.772.
This resource is used in combination with contemporary word-embeddings (Geometry of Culture) in order to create a regression model that predicts valence-scores based on word-embeddings.

3.4 Wikipedia2Vec - (Contemporary) Word-Embeddings

[Yam+20] provide word2vec word-embeddings trained on texts from Wikipedia, meaning these word-embeddings can also be considered contemporary. However, in this work, they are only used for evaluation. Because of this, "contemporary" word-embeddings refers to the Geometry of Culture word-embeddings throughout this work, as those play a much important role. The Wikipedia2Vec-embeddings contain more than 2.3 million rows and contain no missing values. Similar to the Geometry of Culture word-embeddings, these embeddings need to be aligned, which is described in more detail in Section 4.1.
This resource is used in combination with a regression model that predicts valence-scores based on word-embeddings and the valence-lexicon used to train said model in order to evaluate the model on a different dataset than it was trained on.

3.5 Hellrich Historical Affect-lexicon

Contemporary affect-lexicons are available in relative abundance, as they can be created on a large scale using crowdsourcing. However, this is not so simple for historical language stages. One of the few available resources is a small historical VAD-lexicon for 100 words, provided by [HBH18a]. It was created by two annotators, both of which are doctoral students with experience in working with texts from the 19th century. They were told to annotate words from the perspective of a person living in 1830. Valence is again given on a 1 to 9 scale, with a mean of 4.88 and a standard-deviation of 1.2.
This resource is used in combination with the regression model that predicts valence-scores based on word-embeddings and the predictions that this model generates for historical word-embeddings in order to evaluate the performance of the model for historical language stages.

4 Methods

The method for lexicon expansion described by [Li+17] is adapted to reconstruct historical valence-scores for a target historical language stage. This is achieved by training a regression-model on contemporary word-embeddings and the contemporary XANEW-lexicon and then using that model to predict historical valence-scores from the historical word-embeddings. This approach also requires that the contemporary word-embeddings have their embedding-space aligned with the historical word-embeddings.

4.1 Alignment of embedding-spaces

Observing how the vector-representation of a given word changes in the embedding-spaces of consecutive decades requires those embedding-spaces to be aligned. [HLJ16b] already aligned their historical word-embeddings using orthogonal Procrustes and provide the code used for that purpose[2]. Their method finds the orthogonal matrix Q which minimizes the absolute point-wise differences between a word in different embedding-spaces while preserving cosine-similarities:

$$R^{(t)} = arg_0 \min ||QW(t) - W(t+1)||_F,$$

with $W(t) \in R^{d \times |V|}$ being the matrix of words for year t and $R^{(t)} \in \mathbb{R}^{d \times d}$ [HLJ16b]. After adapting this code for Python 3, it was used to align the contemporary word-embeddings to the historical word-ebmeddings of the year 1990 (the latest available decade). Also, the Wikipedia2Vec-embeddings which are used for evaluation purposes needed to be aligned. In particular, they were aligned to the previously aligned contemporary word-embeddings (i.e the contemporary word-embeddings that are the result of alignment to the historical word-embeddings of the year 1990). Table 1 shows the sum of squared residuals (RSS) for all non-zero rows between the word-embeddings of two consecutive language stages. For the 2000-2012 contemporary word-embeddings, the table shows the RSS to the 1990 word-embeddings before and after alignment. For the Wikipedia2Vec word-embeddings, it shows the RSS to the aligned 2000-2012 contemporary word-embeddings before and after alignment. That the RSS for the aligned versions is significantly smaller than that of the unaligned versions, indicates that the alignment process was successful.

4.2 Regression model for predicting historical valence-scores

As proposed by [Li+17], a Ridge-Regression model is trained using word-embeddings as input and valence-scores as the target. Ridge Regression extends least-squares regression by introducing a penalty on the model weights:

$$\min_{\vec{a}} \sum_{s \in V} \| f_j\big(\vec{w}^s\big) - y_j^s\big) \|_2^2 + \alpha R\big(\vec{a}^j\big),$$

with V being the words that the model is trained on, α being the regularization weight and $R\big(\vec{a}^j\big)$ the regularization vector that is applied on the model weights $\vec{a}^j = \big[a_1^j, a_2^j, \ldots, a_n^j\big]$ [Li+17]. The Geometry of Culture / contemporary word-embeddings introduced in Section 3.2 and the XANEW-lexicon introduced in Section 3.3 are used for model training. Partly, these two resources were chosen because they cover roughly the same time period (2000-2012 and 2013 respectively) and are trained with the same method (word2vec / SGNS) and similar data (Google Books Ngram Dataset) as the historical word-embeddings.

[2] https://github.com/williamleif/histwords

Year	Count
1800–1810	7133.27
1810–1820	8786.49
1820–1830	9558.01
1830–1840	9123.23
1840–1850	8810.37
1850–1860	8739.76
1860–1870	8727.94
1870–1880	8614.50
1880–1890	8706.18
1890–1900	8713.47
1900–1910	9109.31
1910–1920	9277.74
1920–1930	9470.97
1930–1940	9525.75
1940–1950	9614.07
1950–1960	9695.19
1960–1970	10018.94
1970–1980	10436.41
1980–1990	10656.92
1990–2012_aligned	11992.38
2012_aligned–w2v_aligned	9799.33
2012_aligned–w2v_unaligned	277093.20
1990–2012_unaligned	406800.13

Table 1: RSS between non-zero rows of word-embeddings of consecutive language stages, with the last four rows showing the difference between the 1990 historical word-embeddings and the aligned and unaligned 2000-2012 contemporary word-embeddings as well as the difference between the aligned 2000-2012 contemporary word-embeddings and the aligned and unaligned Wikipedia2Vec embeddings

4.3 Application of regression model on historical word-embeddings

The regression model described in 4.2 is used to predict valence-scores for all words in the historical word-embeddings. Since the model is trained to predict valence from word-embeddings, and the historical word-embeddings result from historical texts for a given language stage, this approach should results in valence-scores that are more accurate for a given language stage than simply using the valence-scores of a static, modern lexicon (e.g applying the XANEW lexicon from 2013 to embeddings from 1800).

Once valence-scores for all words in the historical word-embeddings haven been predicted, the neighborhood for a given word over the decades can be analyzed. In particular, the k-nearest-neighbors based on cosine-similarity are selected with k = 30, and the probability distribution of their associated valence-scores is calculated. Initially, a k of 100 was also considered, however, this was abandoned as results were largely similar to k = 30 and no further experiment were conducted for determening an optimal value for k. With one such probability distribution of the 30 nearest neighbors for each decade of the historical word-embeddings, how the modes of these distributions change can be seen as an indicator for word-diversification / level of ambiguity. For example when a word has rather unimodal distributions for some number of decades, but this then slowly changes to e.g a bimodal distribution for later decades, word-diversification may have occurred.

5 Results & Analysis

5.1 Evaluation of predicted valence-scores on contemporary data

The regression model which predicts valence-scores based on word-embeddings and that was trained using Geometry of Culture word-embeddings and the XANEW-lexicon described in Section 4.2 is evaluated on contemporary data in two ways. First, performance on the word-embeddings used for the training of the model is evaluated by creating a 80/20 train-test split and calculating the RMSE of the model on the test-set. This means the model is evaluated on an unseen part of the same dataset which was used for training. Secondly, all words that were part of the test-set in the previous setup are again used but this time for the Wikipedia2Vec word-embeddings. This means evaluation is performed on the same unseen words but now also on a different dataset than the one used for training. However, it is important to note that the Wikipedia2Vec word-embeddings were aligned to the Geometry of Culture word-embeddings, as described in Section 4.1. Both setups yielded similar results, with the former achieving an RMSE of 0.87 and the latter of 0.79. Also, these results are very similar to the RMSE of 0.83 that [Li+17] achieved in their work for the XANEW-lexicon (on an unseen part of the dataset they used for training).

5.2 Evaluation of predicted valence-scores on historical data

The predictions of valence-scores that the model produces for the Histwords historical word-embeddings are evaluated using the historical gold-labels described in Section 3.5. Results are compared to [HBH18b], as they not only provide the historical gold-label dataset, but also compare three neighborhood-based approaches for reconstruction of historical affect-scores based on historical word-embeddings and a contemporary affect-lexicon. They also use the XANEW-lexicon, however, their evaluation setup is different from the one used in this work in two ways:

- They use different historical word-embeddings (COHA [Dav02] as opposed to Histwords)
- They report results over all three dimensions of the VAD-model (as opposed to using only the valence-dimension)

Also, they use Pearson's r as their evaluation metric, which is why a different evaluation metric from the one in Section 5.1 is used here. Lastly, Hellrich et al. experiment with two languages (English and German), two types of word-embeddings (SVD and SGNS) and two types of seed-word selection (full and limited), where seed-words refer to the set of words that are chosen from a contemporary affect-lexicon and that the predicted historical affect-values are therefore based on. The results produced in this work are only compared to their English-, SGNS-, full-seed-words-results, as this work only uses English resources, all used word-embeddings are SGNS-word-embeddings and all available seed-words were used (as Hellrich et al. have found this to be superior to a limited, supposedly historically stable subset of seedwords).

Results are shown in Table 2. Pearson's r for the regression model is noticeably higher, however, since Hellrich et al. predict all three dimensions of the VAD-model while this work only uses the valence-dimension, it can be argued that latter task is easier. Also, the difference in the datasets used may be problematic for direct comparisons. Additionally training the regression model on the same dataset that Hellrich et al. and using all VAD-dimensions would make this comparison more meaningful.

Method	Pearson's r
Regression Model	0.501
kNN (Hellrich)	0.365
ParaSimNum (Hellrich)	0.361
RandomWalkNum (Hellrich)	0.361

Table 2: Comparison of Pearson's r of the regression model used in this work (trained on the Histwords word-embeddings and only using the valence-dimension of the VAD-model) and different, neighborhood based approaches (trained on the COHA word-embeddings and using all three dimensions of the VAD-model) by Hellrich et al.

5.3 Analysis of valence-neighborhood over time and its suitability as an indicator for word-diversification

With a regression model being able to predict the valence-score of a word w for a historical language stage based on word-embeddings of that language stage, it is possible to analyze the distribution of valence-scores of w's closest neighbors in the corresponding embedding-space. Inspection of those neighborhood distributions revealed that the vast majority of them have only a single mode and resemble a normal distribution.
This work is interest in word-diversification in the sense that it could be indicated by changes in a words valence neighborhood. More precisely, a word may gain a new meaning that is, in regards to its valence, different enough from its former meaning(s) so that a new mode can be observed in the probability distribution of the valence neighborhood. As it is reasonable to expect that a diversifying word will have the popularity of its different meanings change gradually over time and not completely shift within a decade, it begs the question of whether or not this gradual change can be observed in a word's valence-neighborhood. The simplest form of this would be a shift from a unimodal distribution to a more bimodal distribution.

While searching for words that show such behaviour, two strategies were employed:

- Inspect words where such diversification in a valence-sense is known to likely have occured.
- Quantitatively measure shifts over time from unimodal to bimodal distributions of valence-neighborhoods and vice versa.

5.3.1 Inspection of likely candidates for diversification

When thinking of words which have a different contemporary meaning compared to its meaning in historical language stages, 'gay' might be a prominent example. In modern language, it is used to describe homosexuality and can also be used in offensive manner. In historical language on the other hand, it was used much more positively as 'bright' or 'happy'[3]. Another example would be freak, which diversified from primarily referring to abnormal individuals in a negative way to also being used to positively refer to a person that is

[3] https://www.oxfordlearnersdictionaries.com/definition/english/gay_1

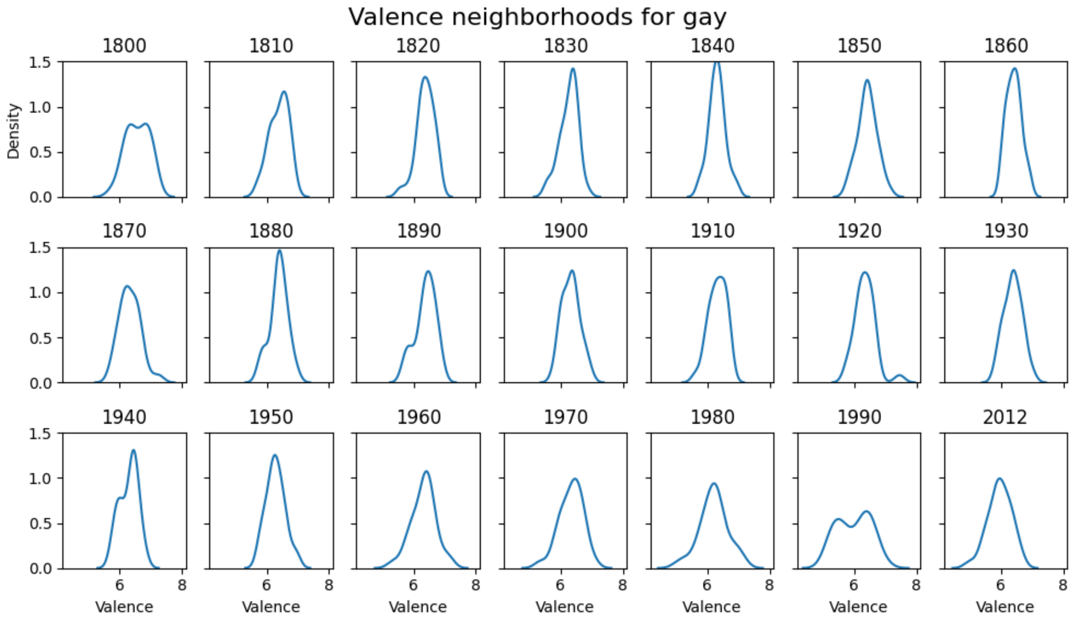

Figure 2: Probability distribution of valence-scores for the closest neighbors of the word 'gay' in the embedding-spaces of consecutive decades from 1800 to 2012

passionate about some topic[4]. The neighborhood valence-distributions of these two words can be seen in Figure 2 and Figure 3. For the word 'freak', the starting year is 1850, as this is the first decade for which it is present in the historical word-embeddings. Both figures show noticeable shifts in distribution over time, however, no clear trend can be seen. For the word "gay" and year 1990, the distribution is distinctly bimodal, however, the preceding and succeeding years 1980 and 2012 are clearly unimodal, so this is likely not the word-diversification we are looking for, as it does not seem plausible for a word to have gained and then lost a new meaning both within a single decade. For the word 'freak', no clear multimodal distributions can be seen and all shifts toward what may be less unimodal distributions in a given decade (e.g 1890 or 1980) quickly shift back to unimodal in subsequent decades.

5.3.2 Quantitative search using Hartigan's Dip-Statistic

Hartigan's Dip-Statistic for Multimodality[5] [HH85] is used to measure the multimodality of a given distribution. For a word w and each decade t of the set of its available decades T, this measure is calculated and its derivative is approximated and summed up using the formula

$$\sum_{t=1}^{|V|-1} (f(x_{w,t+1}) - f(x_{w,t-1}))/2,$$

where $f(x_{w,t})$ denotes Hartigan's Dip-Statistic for year t and word w, and $x_{w,t+1}$ and $x_{w,t-1}$ denote the preceding and succeeding decades. Using this method, the words whose neighborhood distributions changed the most in terms of uni- and mutlimodality are found. For the data starting from 1800, the five words with the highest values for this measure are: 'contempt', 'heresy', 'flew', 'stones' and 'melt'. Figure 4 and Figure 5 show the neighborhood valence-distributions of the 'contempt' and 'flew'. Figures for the other three words are provided in the

[4] https://www.oxfordlearnersdictionaries.com/definition/english/freak_1?q=freak

[5] https://pypi.org/project/diptest/

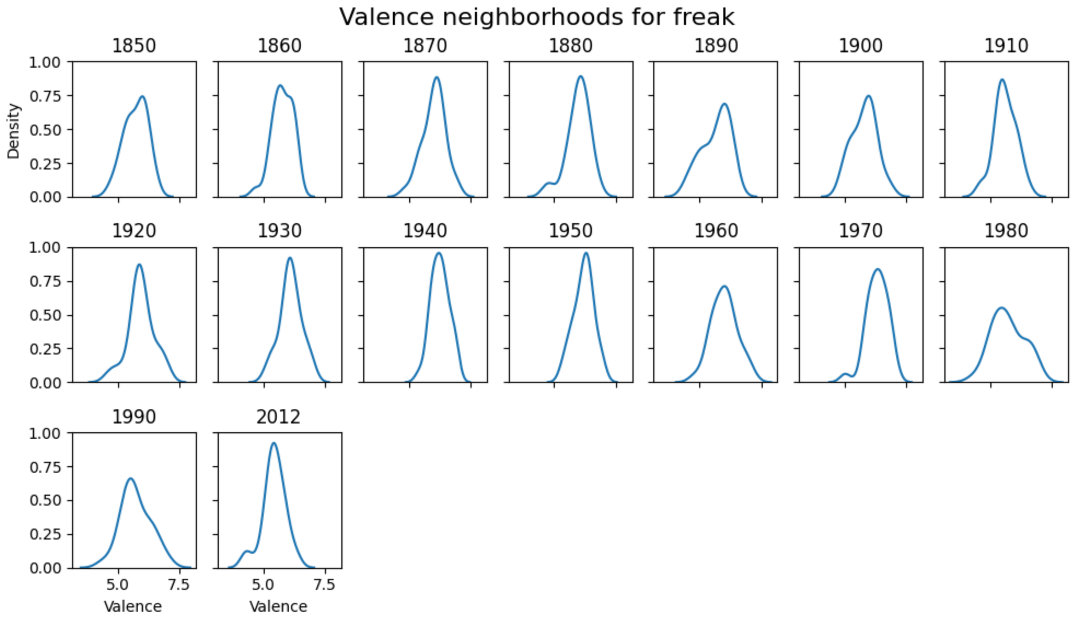

Figure 3: Probability distribution of valence-scores for the closest neighbors of the word 'freak' in the embedding-spaces of consecutive decades from 1850 to 2012

appendix. The distributions for 'contempt' contain some clearly bimodal distributions and a few rather unimodal distributions. However, like with 'gay' and 'freak', no real trend that stays consistent over multiple decades is recognizable. This is even more clearly the case for the Figure showing the neighborhood distributions for 'flew', which constantly fluctuate between uni- and bimodal distributions. Lastly, the appendix contains line plots displaying the value of Hartigan's D and the approximated derivative over time for all five words.

6 Conclusion

The method for affect lexicon expansion based on a regression model trained on word-embeddings introduced in [Li+17] has been successfully adapted for the creation of historical valence scores. Evaluation of this model on the contemporary XANEW lexicon yielded very similar results to those of Li et al., even though a different set of contemporary word-embeddings (Geometry of Culture) was used for training. Furthermore, the model performed similarly well when applied to the Wikipedia2Vec word-embedding (i.e a different set of word-embeddings than it was trained on), suggesting that such a model generalizes well to contemporary data. The results of the evaluation on the historical lexicon are not as easily comparable to other works, the method used in this report theoretically outperforms the methods employed by the creators of the historical lexicons, however, they used the COHA historical word-embeddings, while this work uses the Histword historical word-embeddings. In future work, the experiments described in this report could be done based on the COHA dataset.

In regards to word-diversification, words where a consistent trend in that word's valence neighborhood probability distribution from a uni- to bimodal distribution (or vice versa) were searched for. This was done by looking at words where diversification in a valence-sense is known to likely have occurred (e.g gay or freak) and also by inspecting words where a bimodality-statistic changed the most out of all words over the decades. However, no words that show the described behaviour were found. These results suggest that the reconstructed

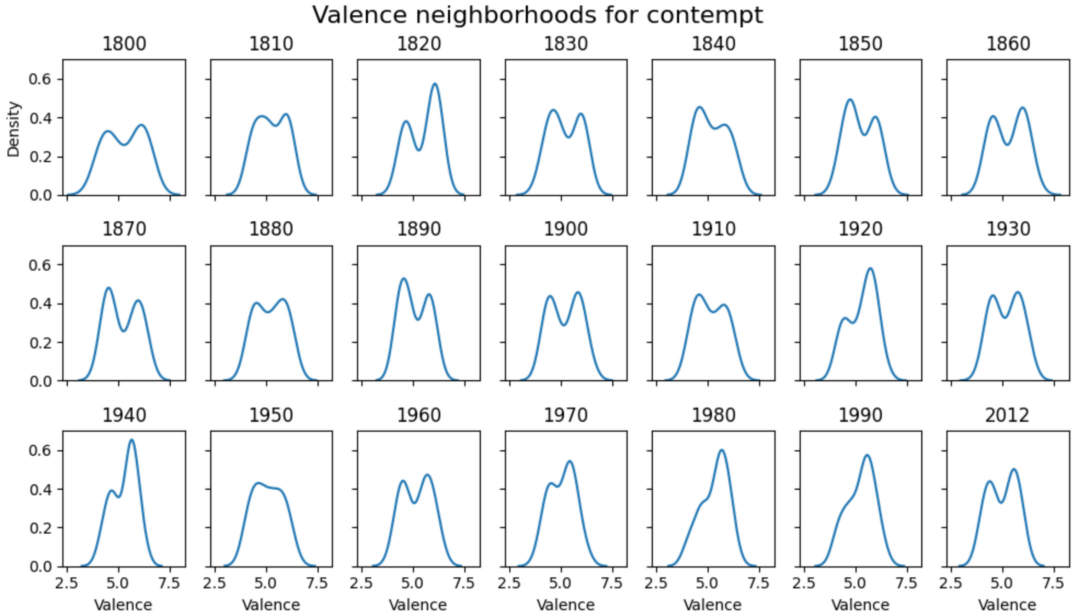

Figure 4: Probability distribution of valence-scores for the closest neighbors of the word ’gay’ in the embedding-spaces of consecutive decades from 1800 to 2012

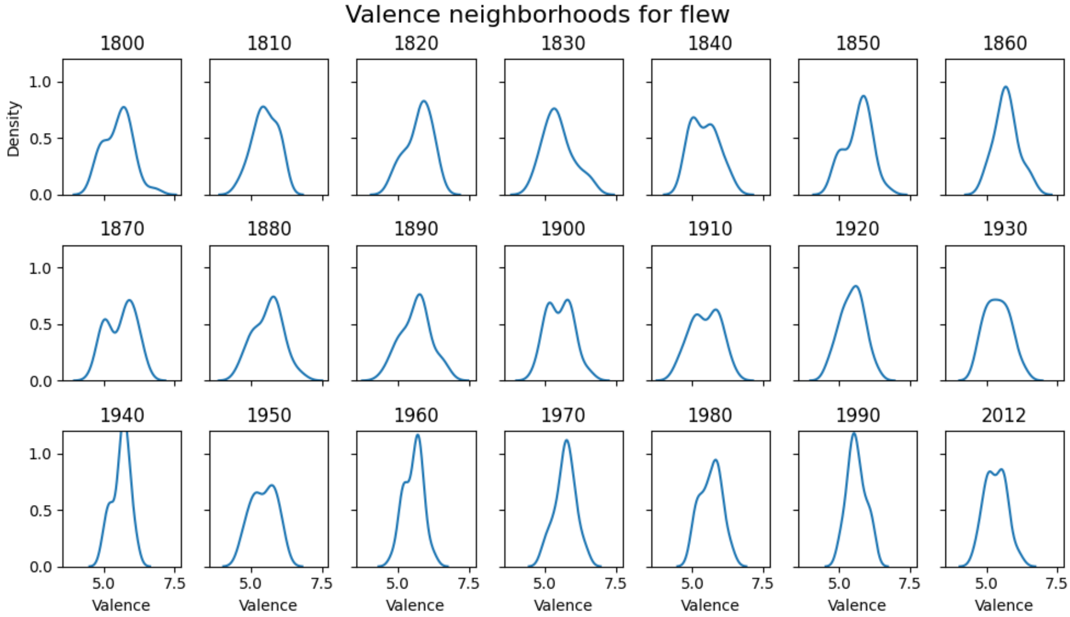

Figure 5: Probability distribution of valence-scores for the closest neighbors of the word ’freak’ in the embedding-spaces of consecutive decades from 1800 to 2012

valence-scores and the information provided by their neighborhoods in a historical word-embedding-space do not seem to be suitable for identifying word-diversification.

References

[HH85] John A Hartigan and Pamela M Hartigan. "The dip test of unimodality." In: *The annals of Statistics* (1985), pp. 70–84.

[BL94] Margaret M Bradley and Peter J Lang. "Measuring emotion: the self-assessment manikin and the semantic differential." In: *Journal of behavior therapy and experimental psychiatry* 25.1 (1994), pp. 49–59.

[Dav02] Mark Davies. *The corpus of historical American English (COHA): 400 million words, 1810-2009*. Brigham Young University, 2002.

[WKB13] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. "Norms of valence, arousal, and dominance for 13,915 English lemmas." In: *Behavior research methods* 45 (2013), pp. 1191–1207.

[HLJ16a] William L Hamilton, Jure Leskovec, and Dan Jurafsky. "Cultural shift or linguistic drift? comparing two computational measures of semantic change." In: *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*. Vol. 2016. NIH Public Access. 2016, p. 2116.

[HLJ16b] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1489–1501. DOI: 10.18653/v1/P16-1141. URL: https://aclanthology.org/P16-1141.

[Li+17] Minglei Li et al. "Inferring Affective Meanings of Words from Word Embedding." In: *IEEE Transactions on Affective Computing* PP (July 2017), pp. 1–1. DOI: 10.1109/TAFFC.2017.2723012.

[HBH18a] Johannes Hellrich, Sven Buechel, and Udo Hahn. *Inducing and Tracking Affective Lexical Semantics in Historical Language*. arxiv.org/abs/1806.08115. 2018.

[HBH18b] Johannes Hellrich, Sven Buechel, and Udo Hahn. "Modeling Word Emotion in Historical Language: Quantity Beats Supposed Stability in Seed Word Selection." In: *LaTeCH@NAACL-HLT*. 2018.

[KTE19] Austin C Kozlowski, Matt Taddy, and James A Evans. "The geometry of culture: Analyzing the meanings of class through word embeddings." In: *American Sociological Review* 84.5 (2019), pp. 905–949.

[Yam+20] Ikuya Yamada et al. "Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 23–30.

Appendix

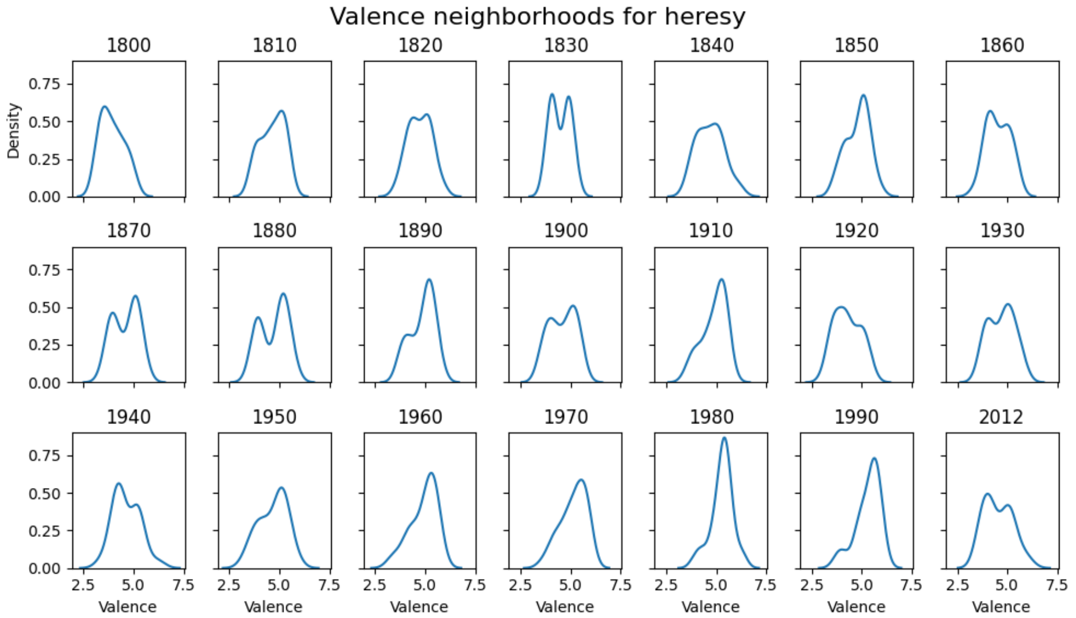

Figure 6: Probability distribution of valence-scores for the closest neighbors of the word 'heresy' in the embedding-spaces of consecutive decades from 1800 to 2012

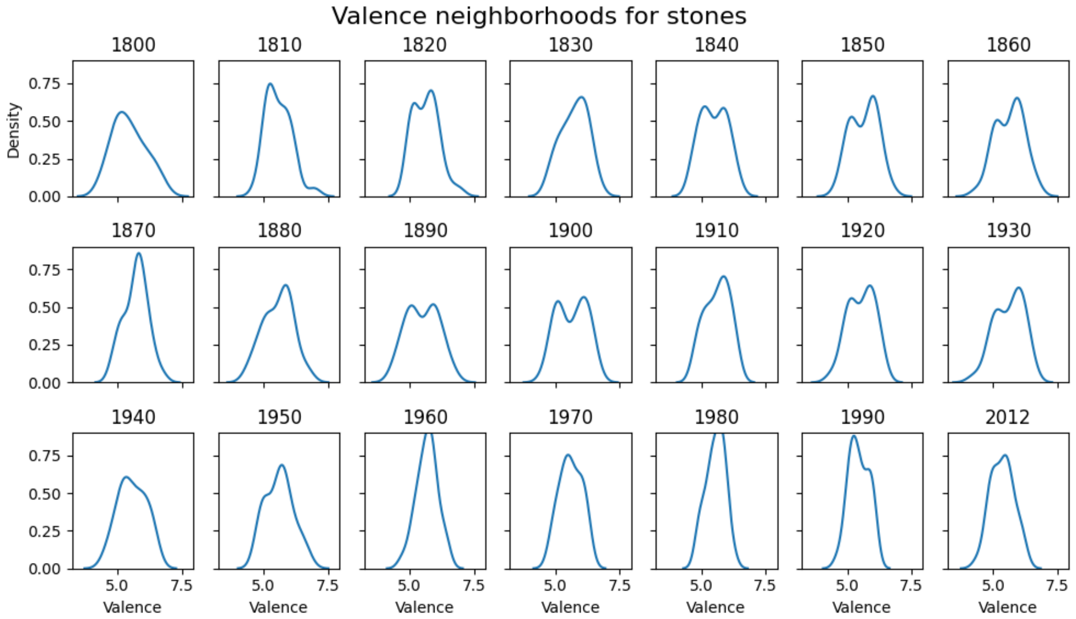

Figure 7: Probability distribution of valence-scores for the closest neighbors of the word 'stones' in the embedding-spaces of consecutive decades from 1800 to 2012

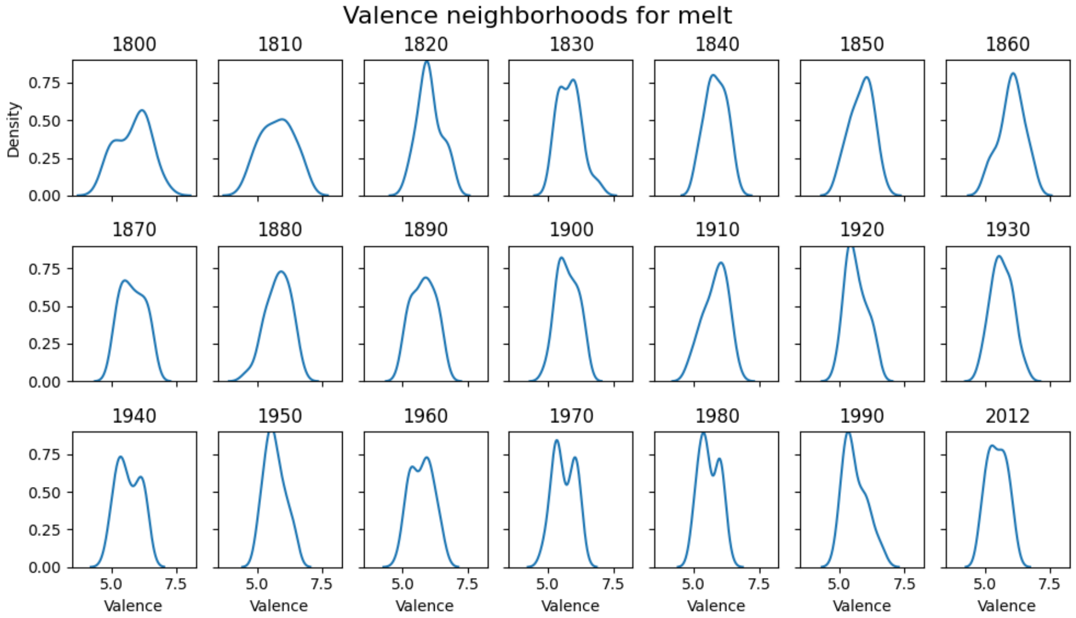

Figure 8: Probability distribution of valence-scores for the closest neighbors of the word 'melt' in the embedding-spaces of consecutive decades from 1800 to 2012

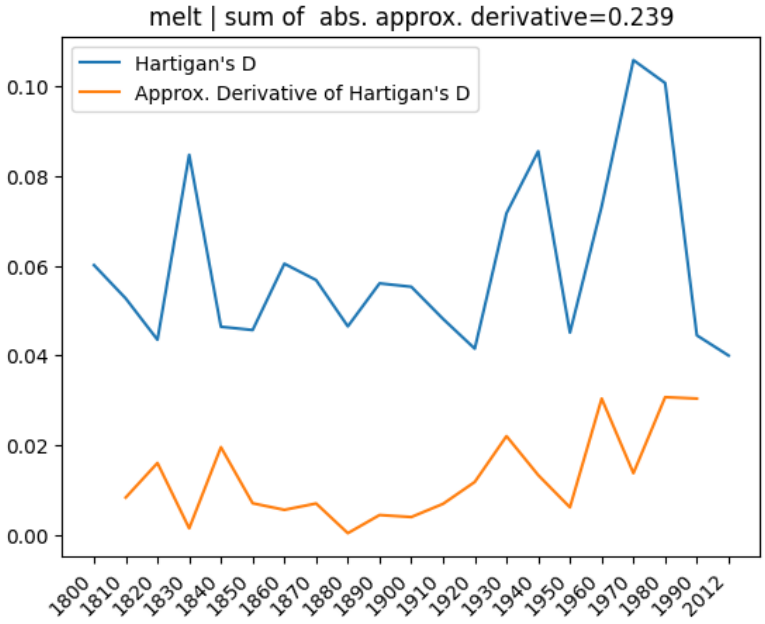

Figure 9: Hartigan's D and its approximated derivative over time for the world 'melt'

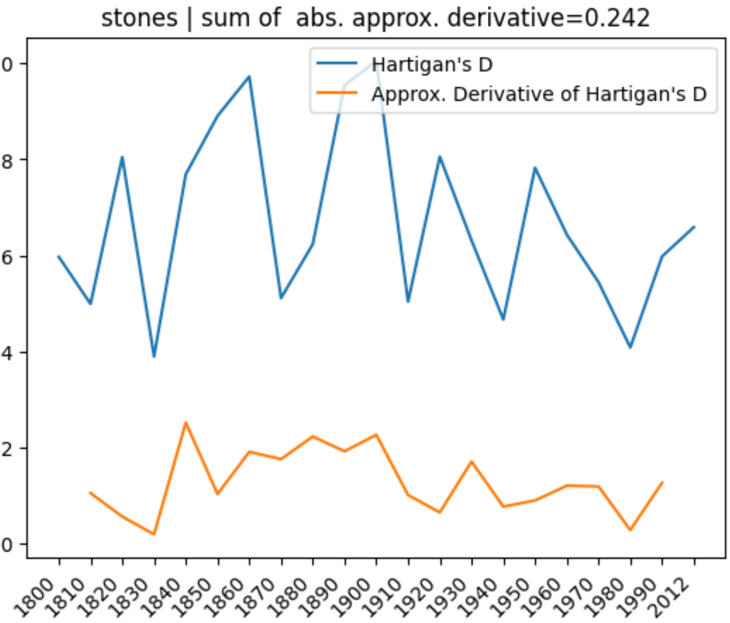

Figure 10: Hartigan's D and its approximated derivative over time for the world 'stones'

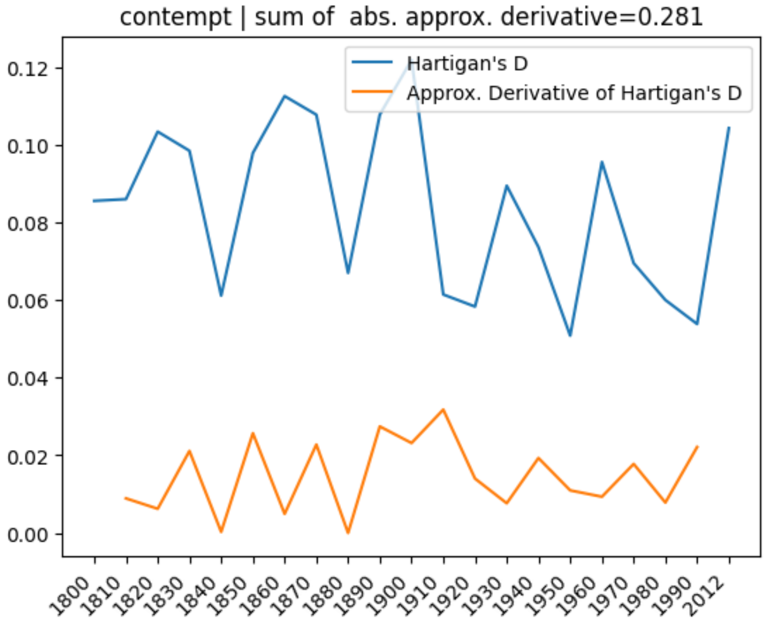

Figure 11: Hartigan's D and its approximated derivative over time for the world 'contempt'

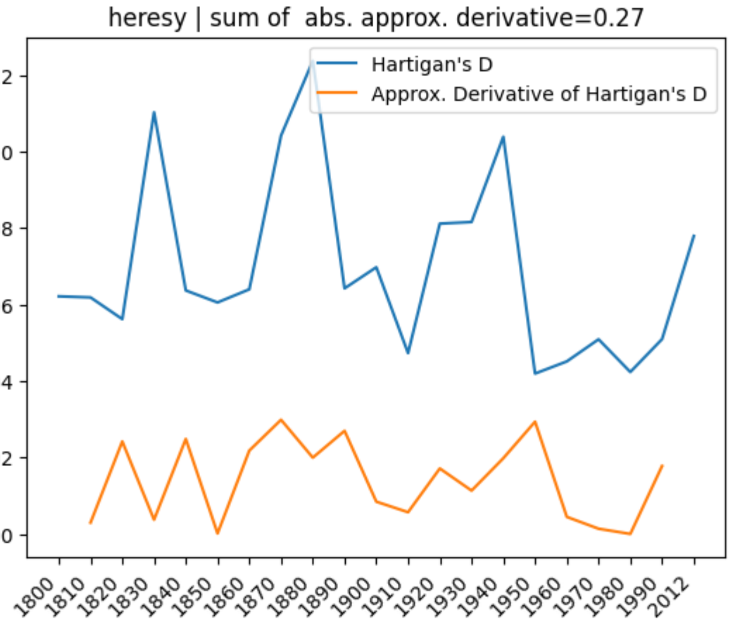

Figure 12: Hartigan’s D and its approximated derivative over time for the world ’heresy’

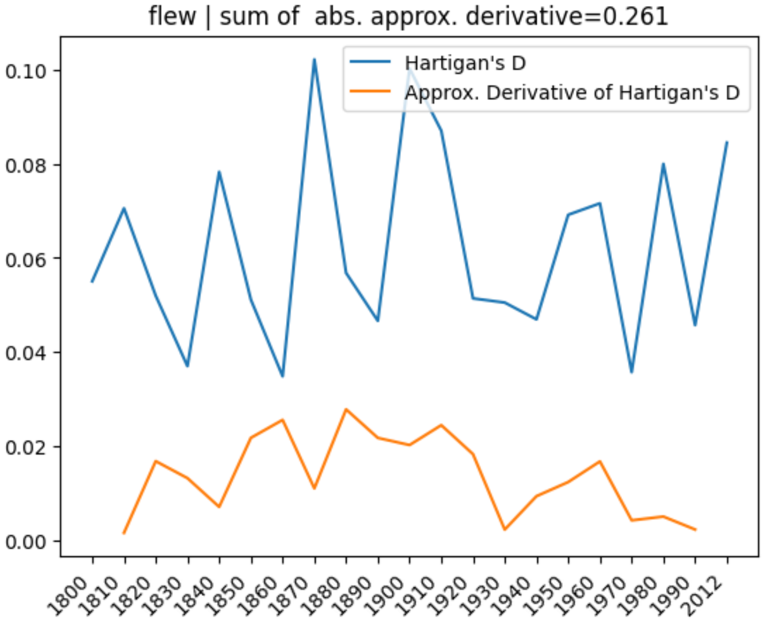

Figure 13: Hartigan’s D and its approximated derivative over time for the world ’flew’