GHisBERT – Training BERT from scratch for lexical semantic investigations across historical German language stages

Christin Beck
University of Konstanz
University of Tübingen
christin.beck@uni-konstanz.de

Marisa Köllner
University of Tübingen
marisa.koellner@uni-tuebingen.de

Abstract

While static embeddings have dominated computational approaches to lexical semantic change for quite some time, recent approaches try to leverage the contextualized embeddings generated by the language model BERT for identifying semantic shifts in historical texts. However, despite their usability for detecting changes in the more recent past, it remains unclear how well language models scale to investigations going back further in time, where the language differs substantially from the training data underlying the models. In this paper, we present GHisBERT, a BERT-based language model trained from scratch on historical data covering all attested stages of German (going back to Old High German, c. 750 CE). Given a lack of ground truth data for investigating lexical semantic change across historical German language stages, we evaluate our model via a lexical similarity analysis of ten stable concepts. We show that, in comparison with an unmodified and a fine-tuned German BERT-base model, our model performs best in terms of assessing inter-concept similarity as well as intra-concept similarity over time. This in turn argues for the necessity of pre-training historical language models from scratch when working with historical linguistic data.

1 Introduction

In historical linguistics, studying semantic change and the evolution of word senses has a long-standing tradition (e.g., Paul, 1880; Ullmann, 1942; Stern, 1964; Lehmann, 1992; Bybee, 2015). However, in NLP and computational linguistics, researchers only recently began to take an interest in the topic, focusing on the task of 'shift detection' (cf. Giulianelli et al., 2020), i.e., the identification of changes in word meaning over time. The task has been taken up in a SemEval challenge on identifying lexical semantic change in English, German, Swedish and Latin (SemEval-2020; Schlechtweg et al., 2020), whose success has inspired several follow-up challenges focusing on different sets of languages, e.g., Italian (Basile et al., 2020), Russian (Pivovarova and Kutuzov, 2021), and Spanish (Zamora-Reina et al., 2022). The interest in the topic is fueled by the possibility to address the task of identifying lexical semantic change via pre-trained neural language models. In particular, recent work addresses the task via methodologies based on contextualized embeddings as generated by the state-of-the-art language model BERT (Devlin et al., 2019), exploring methodologies for how to measure, quantify and evaluate semantic change on the basis of these embeddings (see, e.g., Giulianelli et al., 2020; Martinc et al., 2020; Laicher et al., 2021; Kutuzov et al., 2022).

Despite this recent surge of computational methodologies developed for lexical semantic change detection (LSCD), there are still many historical linguistic research questions related to LSCD which have not yet been touched upon computationally. From a historical linguistic perspective, one of the major shortcomings is the lack of temporal depth. That is, most computational studies focus on identifying change in the more recent past, within one language stage, e.g., comparing English data from the 19th century with data from the 20th century CE. While this renders feasible the application of pre-trained language models such as BERT, which have been trained on contemporary data, and might be of interest for information retrieval applications, this is in general not what is of interest to the historical linguist. In historical linguistics, change is usually investigated across longer periods of time of more temporal depth, with change being assessed across language stages, e.g., from Old English (5th-11th century CE) to Middle English (12th-15th century CE), in order to be able to track sense evolutions in more detail (cf. Stern, 1964). Yet, given that prototypically, the language use as well as the orthography in the historical language stages deviate strongly from

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 33–45

the contemporary language, this casts doubt on the applicability of the readily available pre-trained language models to research questions related to significantly older language stages.

In this paper, we address this methodological gap by developing our own historical BERT-based language model for German: GHisBERT. GHisBERT is trained from scratch on corpus data covering all attested stages of historical German, i.e., Old High German (c. 750-1050 CE, OHG), Middle High German (c. 1050-1350 CE, MHG), Early New High German (c. 1350-1650, ENHG), and New High German (from 1650 onwards, NHG) (see, e.g., Nübling et al. (2008) on the German periodization scheme). We illustrate the usability of our model for research questions related to lexical semantics in historical German by conducting a lexical similarity experiment across three language stages, MHG, ENHG, and NHG. Our experiment is based on measuring the cosine similarity between BERT embeddings produced for ten concepts extracted from the Swadesh (1955) list, i.e., culturally stable concepts which should occur frequently in each of the language stages. To test our model, we assess both, the intra-concept similarity over time as well as inter-concept similarites at each of the investigated time periods. In addition, we compare GHisBERT's performance with a fine-tuned German BERT-base model using the same training data and use the unmodified German model for baseline comparisons. We show that GHisBERT performs better than the other models with respect to capturing intra-concept similarities over time as well as capturing lexical semantic interrelations between the investigated concepts. This highlights the usability of BERT-based models for historical linguistic research questions related to lexical semantics, while at the same time emphasizing the necessity of pre-training language models with the relevant historical data.

2 Related Work

2.1 Lexical semantic change detection

By now, it has become standard to use semantic vector space approaches based on pre-trained neural language models for detecting lexical semantic change (see, e.g., Tahmasebi et al., 2018; Kutuzov et al., 2018; Schlechtweg et al., 2020; Montanelli and Periti, 2023). These approaches can be grouped into (i) type-based approaches (e.g., Hamilton et al., 2016; Hellrich and Hahn, 2016; Schlechtweg et al., 2019), i.e., approaches which use static word embeddings, e.g., word2vec/SGNS (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) embeddings, generating one global vector for each word in a corpus, and (ii) token-based approaches (Hu et al., 2019; Beck, 2020; Giulianelli et al., 2020; Martinc et al., 2020; Montariol et al., 2021; Kurtyigit et al., 2021; Montanelli and Periti, 2023), i.e., approaches based on contextualized word embeddings, e.g., BERT embeddings, which provide one separate context-dependent vector for each occurrence of a word in a corpus.

While LSCD has been previously dominated by type-based approaches and static embeddings (see, e.g., Kaiser et al., 2020; Laicher et al., 2020), recent research efforts move towards producing state-of-the-art results for LSCD based on contextualized BERT embeddings (see, e.g., Kurtyigit et al., 2021; Kutuzov et al., 2022). Several different metrics have been proposed to assess change on the basis of contextualized embeddings and we introduce the most relevant ones in the following.

2.2 Distance-based metrics

Prototypically, for assessing change (and stability) with contextualized word embeddings, distance-based metrics are used which compare the token embeddings computed for a target word across two (or more) corpora from different time periods. Currently, average pair-wise distance (APD) and inverted cosine-similarity over prototypes (PRT) are standardly employed (see, e.g., Giulianelli et al., 2020; Laicher et al., 2020; Kutuzov et al., 2022).

APD Given two corpora C_1 and C_2 representing two different time periods t_1 and t_2, APD represents the average of the distances between all possible pairs of token embeddings, with one embedding per pair representing a target word occurrence in C_1 and the other embedding corresponding to a target word occurrence in C_2. With $U_w^{t_1}$ and $U_w^{t_2}$ referring to the usage matrices of a target word w in t_1 and t_2 respectively,

$$\text{APD}(U_w^{t_1}, U_w^{t_2}) = \frac{1}{N_w^{t_1} \cdot N_w^{t_2}} \sum_{x_i \in U_w^{t_1}, x_j \in U_w^{t_2}} d(x_i, x_j) \quad (1)$$

N corresponds to the number of occurrences of w in each time period, and d is the cosine distance (1-cosine similarity). High APD values are taken

to be indicative of strong semantic change, and low values are to be interpreted as weak change.

PRT Based on the same definitions, but using cosine similarity c instead of d, PRT is the inverted cosine similarity between the average token embedding of all target word occurrences (i.e., the protoype embedding) in C_1 and the protoype embedding in C_2:

$$\text{PRT}(U_w^{t_1}, U_w^{t_2}) = \frac{1}{c(\frac{\sum_{x_i \in U_w^{t1}} x_i}{N_w^{t1}}, \frac{\sum_{x_{ij} \in U_w^{t2}} x_j}{N_w^{t2}})} \quad (2)$$

Inverted cosine similarity is used instead of cosine similarity to produce higher values for stronger changes (see Kutuzov and Giulianelli, 2020). Accordingly, higher values indicate stronger semantic change, lower values indicate weaker changes.

The distance-based estimates are generally evaluated against a human-annotated gold standard, usually with respect to a gold rank where target words are ordered according to their degree of change (see, e.g., subtask 2 of SemEval-2020). In a systematic comparison, Kutuzov et al. (2022) show that averaging the APD and PRT estimates (ensemble method) provides for robust results with respect to predicting the correct rank of target words in terms of change degrees, performing better than using just individual strategies.

In rare cases, the metrics are used for binary change classification, i.e., to classify whether target words are changing over time or not (cf. subtask 1 of SemEval-2020), which requires additional mechanisms. For example, Kurtyigit et al. (2021) propose to use a thresholding technique based on mean and standard deviation values of cosine distances between embeddings and Liu et al. (2021) introduce an approach using permutation-based statistical testing in combination with cosine distances for binary change detection.

2.3 Historical language models

Despite the increasing success of using BERT for LSCD, it remains unclear whether a model trained mostly on contemporary data, e.g., the original BERT-base model is trained on the Google BooksCorpus (800M words) and English Wikipedia (2,500M words), can be readily applied to historical texts. Without having seen any of the relevant historical data during training, the language model might not be able to represent the historical usages of a word adequately.

Addressing this issue, Qiu and Xu (2022) present HistBERT, a BERT-based model which is pre-trained further (i.e., fine-tuned) on the balanced Corpus of Historical American English (COHA; Davies, 2012), adding high-quality balanced historical data going back to the 1820s. They show that, in comparison with the original BERT model, HistBERT provides for improved performances in word similarity tasks and a semantic shift analysis where the underlying data stems from the historical periods covered by the COHA data. Likewise, in earlier work, Martinc et al. (2020) successfully used fine-tuning of a BERT model on the historical corpora under investigation for performance improvement. In addition to further pre-training, Rosin and Radinsky (2022) propose to use a time-aware self-attention mechanism, which encloses temporal information about the text sequences during the extended learning process.

Yet, while fine-tuning on historical data improves lexical semantic change detection, the strong prevalence of the contemporary data used for training BERT might still skew the fine-tuned model towards modern-day language use. Manjavacas and Fonteyn (2022a) show that for historical English (with data going back to 1473 CE), pre-training a BERT model from scratch on the relevant historical data provides for a stronger background model than just fine-tuning the original BERT model with respect to a variety of downstream tasks. In addition, Manjavacas and Fonteyn (2022b) show that historically pre-trained models, i.e., MacBERTh for historical English (1450-1950 CE) and GysBERT for historical Dutch (1500-1950 CE), perform significantly better with respect to non-parametric word sense disambiguation than the corresponding modern models.[1] [2]

Addressing the task of Named Entity Recognition in historical texts, Schweter et al. (2022) pre-train a historical multilingual BERT model (hmBERT) with historical data from German (1683-1949), French (1814-1944), English (1800-1899), Finnish and Swedish (each 1900-1910), establishing a new state-of-the-art via their model.[3]

However, while these models highlight the usefulness of pre-training historical language models, the training data of these models does not support investigations of data exceeding the most recent

[1] https://macberth.netlify.app/

[2] GysBERT and GHisBERT are accidental namesakes.

[3] https://huggingface.co/dbmdz/bert-base-historic-multilingual-cased

historical language stages. It is unclear how well these models scale to data going back further in time, i.e., to data stemming from another historical language stage, where the language differs even more substantially. To our knowledge, there exists no contextualized language model which covers the historical stages of German which we investigate in the present study.

3 GHisBERT: A historical German language model

In this paper, we present **GHisBERT** (**G**erman **His**torical **BERT**), a BERT-based model trained from scratch on historical German data, covering all attested stages of the language, i.e., OHG, MHG, ENHG, and NHG, with data going back to 750 CE.[4]

3.1 Training data

The training data for our model stems from two different sources. More precisely, we extracted all sentences from the *Referenzkorpora zur deutschen Sprachgeschichte* 'Reference Corpora of Historical German', which contain subcorpora for OHG (Referenzkorpus Altdeutsch, ReA, 750-1050 CE; Zeige et al., 2022), MHG (Referenzkorpus Mittelhochdeutsch, ReM, 1050-1350 CE; Klein et al., 2016), and ENHG (Referenzkorpus Frühneuhochdeutsch, ReF, 1350-1650 CE; Herbers et al., 2021).[5] This resulted in 3,227 sentences for OHG, 245,880 sentences for MHG , and 106,988 sentences for ENHG. Sentence splitting was performed based on the presence of modern punctuation markers indicating sentence boundaries (*!.?*) as well as specific historical sentence boundaries, e.g., the middle dot (·), following the respective corpus guidelines.[6] To further balance the training data and to extend the data with contemporary German data, we added data from the *Deutsches Textarchiv* (DTA, Textarchiv, 2023), which is already split into sentences, extracting 100,000 randomly sampled sentences for each of the following periods: 1400-1599, 1600-1799, and 1800-1999. An overview of the data is given in Table 1.

Corpus	Period	Time Span	Sentences	Words
ReA	OHG	750-1050	3 227	18 424
ReM	MHG	1050-1350	245 880	2.3M
ReF	ENHG	1350-1650	106 988	3.7M
DTA1	ENHG	1400-1599	100 000	2.6M
DTA2	NHG	1600-1799	100 000	2.1M
DTA3	NHG	1800-1999	100 000	1.6M
Total	All	750-1999	656 095	12.3M

Table 1: Overview of the training data for GHisBERT.

3.2 Model training

Following Manjavacas and Fonteyn's (2022a) work on historical English, we use the hyperparameterization of the BERT-base configuration and the HuggingFace implementation for training GHisBERT from scratch on historical German data.[7] This corresponds to 12 hidden layers with a hidden size of 768, 12 attention heads, a maximum length of 512 for position embeddings and a vocabulary size of 32,000 tokens. Likewise, we use the masked language modeling (MLM) objective for optimization during training. We trained over 10 epochs, using small batches of size 8 (to avoid memory issues) and gradient accumulation.

For comparison, we further pre-train a modern German BERT-base model via MLM with the same data used for GHisBERT, i.e., we continue training from the last checkpoint of dbmdz/BERT-base-german-cased (henceforth BERT-german), fine-tuning the pre-trained model with historical data.[8] BERT-german was originally trained on over 2 billion words extracted from contemporary texts, e.g., Wikipedia dumps and the EU Bookshop Corpus (Skadiņš et al., 2014).[9] Fine-tuning was performed using the same parameters, but only over 4 epochs as per the recommendations of the original BERT paper (Devlin et al., 2019). We refer to the historically fine-tuned version of BERT-german as BERT-fine. We did not use the multilingual historical model developed by Schweter et al. (2022), i.e.,

[4] GHisBERT is available as a huggingface repository under https://huggingface.co/christinbeck/GHisBERT.

[5] https://www.deutschdiachrondigital.de/

[6] We are aware that identifying sentence boundaries based on punctuation might not always be correct in historical German. Nonetheless, this approximation gives us the relevant context which is needed for training a BERT model.

[7] https://huggingface.co/docs/transformers/model_doc/bert

[8] https://huggingface.co/dbmdz/bert-base-german-cased

[9] Alternatively, we could have used the German BERT variant provided by deepset (https://www.deepset.ai/german-bert). Our choice between the two variants was arbitrary. The dbmdz model is trained on a larger variety of text sources, but whether this presents an advantage over the deepset model still needs to be experimentally defined. We plan to experiment with further model variants and architectures in the future.

hmBERT, which also contains historical German data, in our experiment, because for one, the historical German data used for training hmBERT still only represents the NHG language stage and for another, having multiple training languages renders a direct comparison with our model more difficult.

In order to be able to deal with the historical orthography and word forms present in our data, we train our own BERT tokenizer on our historical data. This tokenizer is used for tokenization before feeding the historical data into any of the models.[10]

4 Lexical similarity and stability across language stages

To test the applicability of our model to investigations of lexical semantic change in historical language stages, we conduct a case study which investigates whether the lexical semantic stability of ten Swadesh concepts is captured adequately over time, i.e., across three consecutive historical language stages: MHG, ENHG, and NHG. To do so, we compare GHisBERT with BERT-fine and BERT-german via a lexical similarity analysis, assessing the inter-concept similarity at each time stage as well as the intra-concept similarity of each concept across time.

4.1 Target concepts

Most existing computational studies on LSC in German base their investigations on the 48 German target words which were part of the SemEval-2020 challenge (see, e.g., Kurtyigit et al., 2021). However, only very few of these NHG target words can be found in the historically older language stages. We therefore selected ten target words which occur in all three language stages from the 200-word list of basic vocabulary introduced by Swadesh (1955). These concepts are well distributed throughout the list according to Swadesh's stability ranking: VOGEL 'bird', HUND 'dog', EI 'egg', FISCH 'fish' (among the first 50 most stable concepts); BERG 'mountain', FUSS 'foot, KOPF 'head' (among the 50-100 most stable concepts); FRAU 'woman', BAUM 'tree', SONNE 'sun' (among the 100-200 most stable concepts). The basic vocabulary list was both narrowed and extended in recent studies in the course of the establishment of different databases (see, e.g., Dellert and Buch, 2018; Holman et al., 2008), but since the estimation of a concept stability ranking is highly data-dependent, it differs with regard to the languages under investigation. We therefore use the stability ranking of the well-established 200-word Swadesh list, provided by Dellert and Buch (2018), for the selection of the target words. While the concepts themselves are expected to be stable across languages and time, the corresponding word forms are not excluded from undergoing lexical semantic change. However, given their concept stability, we expect the word forms to be relatively stable within one language and within our examined time range.

4.2 Data

For our investigation, we extract all sentences from the 'Reference Corpora of Historical German' in which one of our targets occurs, using the same sentence generation principles as given in Section 3.1. This proportion of the data covers the MHG and ENHG period in our study (via the ReM and ReF corpora). To cover the NHG period, we extract all sentences from the DTA in which the target concepts occur in the time span 1700-1999. Overall, this results in 148,306 sentences, with 3,942 MHG sentences, 6,009 ENHG sentences, and 138,355 NHG sentences.[11]

While the concepts are assumed to be stable parts of the language, occurring in all three stages, the word forms themselves are subject to change over time, undergoing phonological and morphological changes (see, e.g., Nübling et al., 2008). To be able to track the concepts as target words over the language stages, we first had to identify the relevant historical lemmas of our concepts, forming 'etymological chains' assigning the historical word forms of each concept to their contemporary counterparts.

4.3 Etymological chains

We build our etymological chains based on information extracted from Kluge (2012), an etymological dictionary which provides OHG and MHG correspondences of NHG words. For example, *fuoz* is given as the OHG form and *vuoz* as the MHG form of NHG *Fuß* 'foot' in Kluge (2012) (see Table 3 in Appendix A for a full list of lemma correspondences and the respective occurrence frequencies across stages). We searched for all possible correspondences of our target concepts in the lemmatized versions of each of the corpora under inves-

[10] The source code used for tokenization, model training and fine-tuning is available at https://github.com/christinschaetzle/GHisBERT.

[11] We excluded the OHG period from our investigation since our target concepts only rarely occurred in this stage, see Table 3 in Appendix A for the relevant occurrence frequencies.

tigation, and extracted the respective sentences in their non-lemmatized form.[12]

4.4 Concept embeddings

For each of the sentences in which our target concepts occur, we generate target word embeddings using our different model versions in the following way. First, we replace the target word representing one of our concepts with the NHG lemma version of the concept, e.g., *vuoz* is replaced with *Fuß*, to mitigate the word form bias reported by Laicher et al. (2021). That is, we use the non-lemmatized, original, sentences to produce concept embeddings, but replace the target word form by its modern concept lemma. After tokenization, we pass the sentences to the model and extract the corresponding sentence embeddings at the second-to-last layer. We use the second-to-last layer, since this layer has been shown to provide the most context-specific embeddings (Ethayarajh, 2019).[13] Next, we compute the word embeddings of each target concept occurrence by averaging over the respective word-piece embeddings, as is standard procedure.

4.5 Lexical similarity analysis

In order to generate insights into whether our model is able to be used for systems investigating lexical semantic change across language stages, we investigate whether GHisBERT is able to produce adequate results in a lexical similarity analysis of our stable target concepts. That is, we assess the inter-concept similarity of each concept at each language stage, by computing the cosine similarity (COS) between the average embedding of a concept to the average embeddings of all other concepts at a given stage. Additionally, we measure the intra-concept similarity of each concept over time, by comparing the average embeddings of each concept separately between language stages via COS. Ideally, a concept should show significantly greater simliarities to itself over time than to other concepts at each language stage. In addition, the best model should show the largest differences (i.e., lowest similarities) across concepts, capturing the lexical semantic interrelations between the target concepts. To test this, we compute paired t-tests testing for significant differences between the inter-similarity distribution of a concept and the intra-similarity of a concept over time.

Overall, there is still no consensus on which metrics to use for identifying lexical semantic change (and stability in turn) based on BERT embeddings. We experimented with several of the distance-based metrics introduced in Section 2.2, including APD, COS, PRT, and the ensemble method, which averages APD and PRT. Overall, we found that COS, with its value boundedness between 0 and 1, provides for the most interpretative measure with respect to both, intra-concept similiarity over time as well as inter-concept similiarity.[14]

4.6 Evaluation

Most existing work on LSCD ranks the target words under investigation with respect to a quantitative estimate indicating the degree of change of a word between two time periods. This ranking is then usually evaluated against a gold dataset, where the same target words have been ranked on the basis of a detailed, extensive manual annotation process. As this is the first research enterprise setting out to track lexical semantic change based on contextualized embeddings across historical language stages of German, there exists yet no gold data that goes back far enough in time to be compatible with our investigated data. Therefore, we were not able to perform a comparable ground truth evaluation. Instead, we calculate t-tests for assessing similarities and differences between the results produced by the individual models. In addition, we perform qualitative cross-checks of the underlying data via a manual inspection of 50 randomly sampled sentences per concept and language stage.

5 Results

Inter-concept similarity The inter-concept similarity at each time stage shows how similar the average embedding of each concept is to the average embeddings of all other concepts. In terms of the inter-concept similarity at each stage, GHisBERT provides for the best results, presenting similarities that range between 0.18 and 1, representing low and high similarities between concepts adequately, while BERT-fine and BERT-german provide much

[12] In addition, we considered further spelling variants to cover as much data as possible, e.g., *bërg* is used for BERG 'mountain' in ReM, while Kluge (2012) gives *berc* for MHG.

[13] We also experimented with concatenation of the embeddings of the last four layers, averaging over the embeddings of the last four layers, and summing the embeddings of the last four layers. The differences between those approaches are marginal, but concatenation and the second-to-last layer approach produce slightly stronger similarity values.

[14] We provide the code for our experiment under https://github.com/christinschaetzle/GHisBERT.

	GHisBERT			BERT-fine			BERT-german		
Concept	COS_{ME}	COS_{EN}	COS_{avg}	COS_{ME}	COS_{EN}	COS_{avg}	COS_{ME}	COS_{EN}	COS_{avg}
BAUM	0.90	0.96	0.93***	0.95	0.94	0.95***	0.98	0.99	0.99***
BERG	0.92	0.97	0.95***	0.94	0.93	0.94***	0.98	0.99	0.99***
EI	**0.70**	**0.92**	**0.81*** **	**0.73**	**0.92**	**0.82***	0.94	0.98	0.96***
FISCH	0.85	0.87	0.86***	0.94	0.93	0.93***	0.99	0.99	0.99***
FRAU	0.95	0.95	0.95***	0.95	0.91	0.93***	0.99	0.99	0.99***
FUSS	0.88	0.89	0.89***	0.94	0.89	0.92**	0.97	0.99	0.98***
HUND	0.87	0.95	0.91***	0.91	0.93	0.92***	0.98	0.98	0.98***
KOPF	**0.85**	**0.94**	**0.90*** **	0.93	0.94	0.93***	0.98	0.99	0.98***
SONNE	0.89	0.95	0.92***	0.93	0.93	0.93***	0.98	0.99	0.98***
VOGEL	0.92	0.96	0.94***	0.94	0.94	0.94***	0.97	0.98	0.97***

Table 2: Cosine similarities between average concept embeddings from MHG and ENHG (COS_{ME}) and ENHG and NHG (COS_{EN}), as well as the average of these similarities (COS_{avg}). Statistically significant differences between inter- and intra-concept similarity are calculated via t-tests (p<0.001***, p<0.01**, p<0.05*).

higher similarities, see the heatmaps in Figure 1. In particular, BERT-german produces very high simliarity values between concepts at each stage, i.e., values ranging between 0.87 and 1, not being able to capture the differences between the concepts.

Overall, GHisBERT gives the most pronounced representation of synchronic inter-concept similarities. At the MHG stage, EI 'egg' shows the lowest similarity to all other concepts with all three models. This is an interesting text effect which is borne out in particularly by the GHisBERT embeddings: in the MHG proportion of the data, EI only occurs in Latin texts, referring to the 3rd person masculine pronoun *ei* 'he', and not to 'egg'. As such, it is no surprise that it differs from all other concepts. Other lexical semantic similarities which are neatly captured by GHisBERT at all stages are the relationship between animal concepts, e.g., FISCH 'fish', VOGEL 'bird', and HUND 'dog' show high simliarities to one another, and the interrelation between body parts, e.g., KOPF 'head' and FUSS 'foot'. In addition, FRAU 'woman', which is the only human, sentient concept, shows lower similarities than the other concepts to one another (with EI being an exception here).

Intra-concept similarity across time Table 2 shows the cosine similarities between the average concept embeddings across language stages, i.e., between MHG and ENHG (COS_{ME}), between ENHG and NHG (COS_{EN}), and the average across the two distributions (COS_{avg}) for each of the three models. Despite the high inter-concept similarities reported for BERT-german, all three models show highly statistically significant differences between the average inter-concept similarity distributions and the average intra-concept similarity over time (COS_{avg}), see Table 2. Yet again, for BERT-fine and BERT-german, the similarity values are less nuanced than for GHisBERT. In particular, the comparably large change for EI, which is due to the Latin influence in MHG that is not present in the ENHG and NHG data for EI, is most pronounced for GHisBERT. However, the similarity values for EI are similar with GHisBERT and BERT-fine, despite a lower significance in terms of the difference between EI's inter- and intra-concept similarity for BERT-fine. Overall, the COS_{avg} distributions of GHisBERT and BERT-fine do not show a statistically significant difference, whereas the difference between GHisBERT and BERT-german is significant (as is the difference between BERT-fine and BERT-german, both with p<0.001).

Yet, what is striking, is that GHisBERT adequately estimates a larger difference, i.e., a lower similarity, between MHG and ENHG than between ENHG and NHG, while this is not the case for the other two models, with a significant difference between the COS_{ME} distributions of GHisBERT to the other models (p<0.001). GHisBERT's results are in line with broader linguistic developments in historical German (see, e.g., Nübling et al., 2008; Fleischer and Schallert, 2011), for which MHG can be characterized as a major period of change, with a considerably freer word order and several strong phonological and morphological changes (e.g., vowel reduction), leading to the ENHG period, thus reflecting a stronger change between MHG and ENHG than between ENHG and NHG.

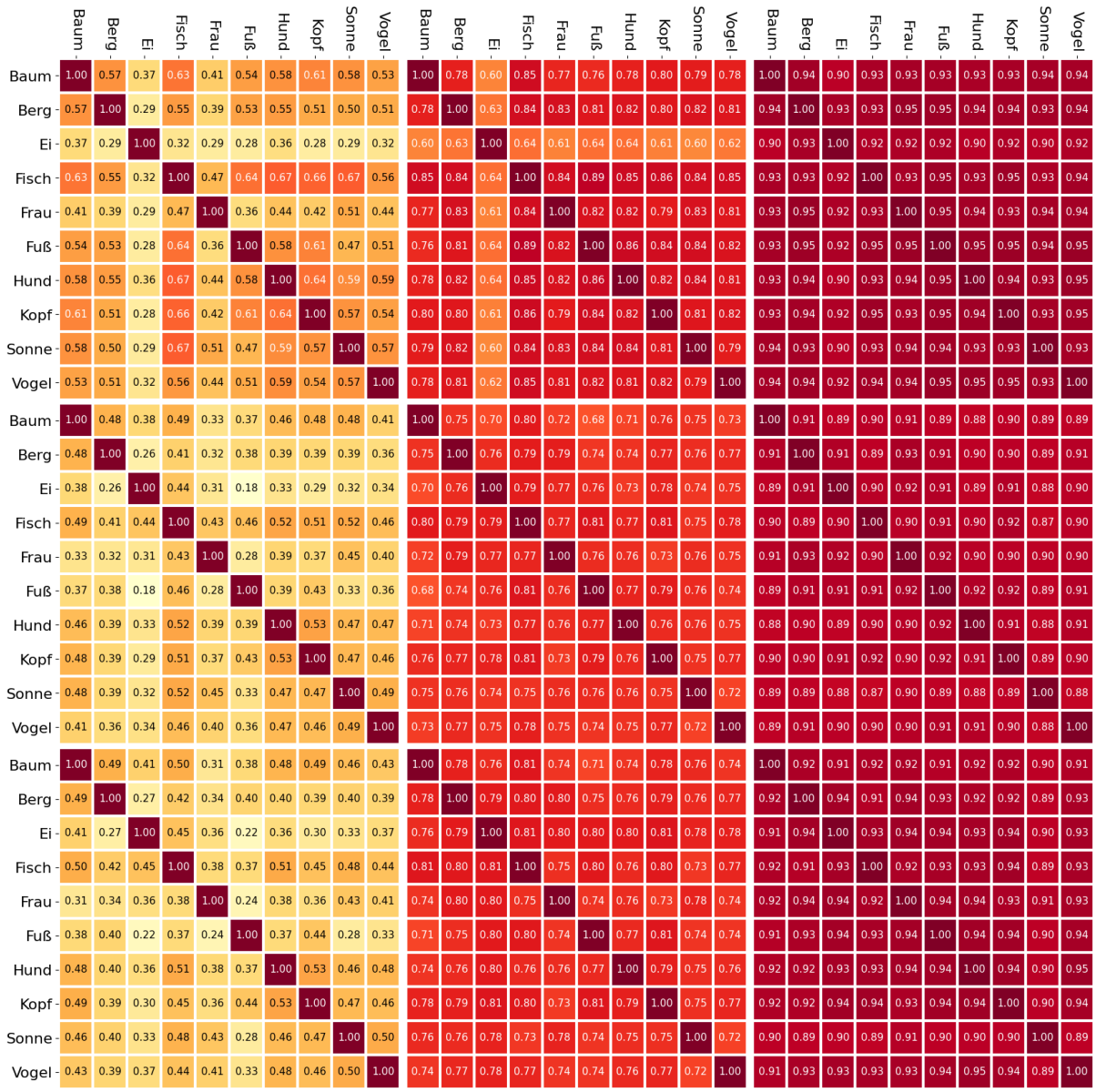

Figure 1: Heatmaps showing the inter-concept similarities at the MHG (top), ENHG (middle), NHG (bottom) stage as calculated via embeddings from GHisBERT (left), BERT-fine (center), and BERT-german (right).

In addition, several of our qualitative observations fit well with the results produced by GHisBERT. For one, concepts which show lower similarities with respect to both COS_{ME} and COS_{EN}, i.e., FISCH 'fish' and FUSS 'foot', show polysemy in the corpora from all three language stages, with FUSS 'foot' referring to the body part, the 'foot' (bottom) of a mountain, and its usage as a measure of length. FISCH 'fish' in turn is found in its biological as well as astrological usage and additionally occurs often in biblical contexts. For another, KOPF 'head', which shows a comparably low COS_{ME} but a large COS_{EN} similarity, seems to be undergoing change between MHG and ENHG: in MHG, KOPF is still mainly found in its historically older use as 'drinking vessel, cup' (cf. Kluge, 2012; Pfeifer et al., 1993), which differs strikingly from the usage in ENHG and NHG as 'head', and is no longer found in modern German. While this development stands out with GHisBERT, it is less evident with the other models, see Table 2.

In sum, our lexical similarity analysis shows that GHisBERT provides for the best results in terms of capturing the lexical semantic relationships between our ten target concepts in the historical language stages. The results produced by GHisBERT present a more nuanced picture of synchronic as well as diachronic interrelations between target con-

cepts than the results achieved via the unmodified and the fine-tuned BERT-german models. Overall, these findings are in line with our manual qualitative cross-checks of the underlying data.

6 Conclusion

This paper provides evidence for the usability of BERT-based models for investigations of lexical semantic change going beyond the contemporary language stage. More precisely, we show via a lexical similarity analysis that BERT embeddings can be used for assessing inter- and intra-concept similarities across three historical German language stages, Middle High German, Early New High German, and New High German. In a systematic comparison, we show that pre-training a BERT-based model from scratch with the relevant historical data provides for more adequate results than fine-tuning alone. This in turn highlights the relevance of pre-training neural language models with language-specific data for lexical semantic investigations.

Limitations

While our paper presents the first research endeavor (that we know of) which investigates lexical semantics in historical German going beyond the NHG stage using BERT embeddings, it also points out the necessity of more ground truth data for evaluation. The lack of a gold standard for evaluation is the strongest limitation of our paper, leading to a lack of a true quantitative evaluation. Annotating data from historical language stages is notoriously difficult and time-consuming, requiring expert knowledge of the language stages (see, e.g., Beck et al., 2020). Therefore, we first set out to investigate whether GHisBERT potentially is a useful tool for investigating lexical semantic change across language stages in this paper before manually annotating data, but definitely plan to do so in the future (together with expert annotators). Along with this, we intend to evaluate our model with respect to further lexical semantic tasks in the future.

A further limitation is the large computational power and time which is generally needed for training a BERT model from scratch: this might not always be feasible for researchers with a more historical linguistic background, which might be lacking the necessary infrastructure. It is thus unclear how well our methodology is transferable to studies seeking to understand lexical semantic developments in the history of other languages and with respect to different datasets. A related issue is that most studies on LSC focus on using BERT embeddings, but it remains unclear how well more recent large language models, e.g., GPT-4 (OpenAI, 2023), and different model architectures scale to the task of investigating LSC across language stages, and, in turn, how these play out the computational issues.

We moreover leave frequency effects and extra-linguistic factors, such as different text genres and dialects, aside in this paper, but intend to look further into this as part of future work.

Ethics Statement

This paper does not present a risk or produce any liabilities for individuals and/or groups of individuals and we do not expect any negative consequences. We provide full transparency of our methodology by open-sourcing our source codes and models. The corpus data underlying our approach in this paper is available as open source and we do not select data based on any ethnic restrictions. We either use all the available data or randomly select individual data points. Having stated this, historical data is generally not equally distributed with respect to regional and social aspects (as well as many other factors), and it is to be expected that certain varieties of speakers will be underrepresented in the data. This is a general problem of historical linguistic work, which we were not able to mitigate in this paper.

Acknowledgements

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding our research within DFG FOR 2237: Project "Words, Bones, Genes, Tools: Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past" (both authors) and within Project D02 "Evaluation Metrics for Visual Analytics in Linguistics" – Project-ID 251654672 – TRR 161 (Christin Beck). In addition, we would like to thank Fadhl Al-Eryani for helping us with the corpus work and Chundra Cathcart for setting up the Kluge chains and discussing various aspects of this work with us.

References

Pierpaolo Basile, A. Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 diachronic lexical semantics (DIACR-Ita) task.

EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020.

Christin Beck. 2020. DiaSense at SemEval-2020 task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 50–58, Barcelona (online). International Committee for Computational Linguistics.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain. Association for Computational Linguistics.

Joan Bybee. 2015. *Language Change*. Cambridge University Press.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

Johannes Dellert and Armin Buch. 2018. A new approach to concept basicness and stability as a window to the robustness of concept list rankings. *Language Dynamics and Change*, 8(2):157 – 181.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Jürg Fleischer and Oliver Schallert. 2011. *Historische Syntax des Deutschen. Eine Einführung*. Narr Verlag, Tübingen.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Birgit Herbers, Sylwia Kösser, Ilka Lemke, Ulrich Wenner, Juliane Berger, Sarah Kwekkeboom, and Frauke Thielert. 2021. Dokumentation zum Referenzkorpus Frühneuhochdeutsch und Referenzkorpus Deutsche Inschriften. *Bochumer Linguistische Arbeitsberichte*, 24.

Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, Dik Bakker, et al. 2008. Advances in Automated Language Classification. *Quantitative Investigations in Theoretical Linguistics*.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Thomas Klein, Klaus-Peter Wegera, Stefanie Dipper, and Claudia Wich-Reif. 2016. Referenzkorpus Mittelhochdeutsch (1050-1350). Version 1.0. https://www.linguistics.ruhr-uni-bochum.de/rem/.

Friedrich Kluge. 2012. *Etymologisches Wörterbuch der deutschen Sprache*. De Gruyter.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Severin Laicher, Gioia Baldissin, Enrique Castaneda, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Winfried P. Lehmann. 1992. *Historical Linguistics: An Introduction*. Holt. 3rd edition, first published in 1962.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Statistically significant detection of semantic shifts using contextual word embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Enrique Manjavacas and Lauren Fonteyn. 2022a. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Enrique Manjavacas and Lauren Fonteyn. 2022b. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. ArXiv:2304.01666.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Damaris Nübling, Antje Dammel, Janet Duke, and Renata Szczepaniak. 2008. *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*, 2nd edition. Gunter Narr Verlag, Tübingen.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hermann Paul. 1880. *Principien der Sprachgeschichte*. Niemeyer, Tübingen.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Wolfgang Pfeifer, Wilhelm Braun, Gunhild Ginschel, Gustav Hagen, Anna Huber, Heinrich Petermann Klaus Müller, Gerlinde Pfeifer, Dorothee Schröter, and Ulrich Schröter. 1993. *Etymologisches Wörterbuch des Deutschen*. Digitalisierte und von Wolfgang Pfeifer überarbeitete Version im Digitalen Wörterbuch der deutschen Sprache, `https://www.dwds.de/d/wb-etymwb`.

Lidia Pivovarova and Andrey Kutuzov. 2021. RuShiftEval: a shared task on semantic shift detection for Russian. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Wenjun Qiu and Yang Xu. 2022. HistBERT: A pretrained language model for diachronic lexical semantic analysis. *CoRR*, abs/2202.03612.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *To appear in Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Stefan Schweter, Luisa März, and Erion Çano. 2022. hmBERT: Historical multilingual language models for named entity recognition. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2022)*.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gustaf Stern. 1964. Meaning and change of meaning, with special reference to the English language. In *Indiana University Studies in the History and Theory of Linguistics*. Indiana University Press, Bloomington. First published in 1931.

Morris Swadesh. 1955. Towards greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21(2):121–137.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Deutsches Textarchiv. 2023. *Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*. Berlin-Brandenburgischen Akademie der Wissenschaften, Berlin. https://www.deutschestextarchiv.de/.

Stephen Ullmann. 1942. The range and mechanism of changes of meaning. *The Journal of English and German Philology*, 61:46–52.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

Lars Erik Zeige, Gohar Schnelle, Martin Klotz, Karin Donhauser, Jost Gippert, and Rosemarie Lühr. 2022. Deutsch Diachron Digital. Referenzkorpus Altdeutsch. Humboldt-Universität zu Berlin. http://www.deutschdiachrondigital.de/rea/.

A Appendix A: Target concepts

	NHG		ENHG		MHG		OHG		
Concept	lemma	*n*	lemma	*n*	lemma	*n*	lemma	*n*	Total *n*
BAUM ‘tree’	*Baum*	181	*Baum*	300	*boum*	8 845	*boum*	0	9 326
BERG ‘mountain’	*Berg*	423	*Berg*	647	*berc*	14 020	*berg*	0	15 090
EI ‘egg’	*Ei*	5	*Ei*	129	*ei*	7 385	*ei*	0	7 519
FISCH ‘fish’	*Fisch*	110	*Fisch*	344	*visch*	5 331	*fisc*	1	5 786
FRAU ‘woman’	*Frau*	2 050	*Frau*	2 935	*vro(u)we*	3 7702	*frouwa*	0	42 687
FUSS ‘foot’	*Fuß*	487	*Fuß*	65	*vuoz*	21 999	*fuoz*	3	22 554
HUND ‘dog’	*Hund*	110	*Hund*	269	*hunt*	8 070	*hunt*	0	8 449
KOPF ‘head’	*Kopf*	29	*Kopf*	223	*kopf*	16 067	*kopf, kupf*	0	16 319
SONNE ‘sun’	*Sonne*	396	*Sonne*	914	*sunne*	11 293	*sunna*	3	12 606
VOGEL ‘bird’	*Vogel*	151	*Vogel*	183	*vogel*	7 643	*fogal*	0	7 977
All		3 942		6 009		138 355		7	148 313

Table 3: Target concepts and the occurrence frequencies of the corresponding lemmas at each language stage.