Comparing Gender Bias in Lexical Semantics and World Knowledge: Deep-learning Models Pre-trained on Historical Corpus

Yingqiu Ge[1,2], Jinghang Gu[1]*, Chu-Ren Huang[1], Lifu Li[3]
[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, China
[2]School of Foreign Languages, Yunnan University, China
[3]School of Management, Yunnan Normal University, China
Correspondence: gujinghangnlp@gmail.com

Abstract

This study investigates the impact of continued pre-training transformer-based deep learning models on historical corpus, focusing on BERT, RoBERTa, XLNet, and GPT-2. By extracting word representations from different layers, we compute gender bias embedding scores and analyze their correlation with human bias scores and real-world occupation participation differences. Our results show that BERT, an encoder-only model, achieves the most substantial improvement in capturing human-like lexical semantics and world knowledge, outperforming traditional static word vectors like Word2Vec. Continued pre-training on historical data significantly enhances BERT's performance, especially in the lower-middle layers. When historical human biases are difficult to quantify due to data scarcity, continued pre-training BERT on historical corpora and averaging lexical representations up to the 6th layer provides an accurate reflection of gender-related historical biases and world knowledge.

1 Introduction

The core idea of distributional semantics models (DSMs) is that the context of a word usage can be used to explore its semantics (Harris 1954; Firth 1957). With the rapid advances in deep learning, models based on deep transformer networks (Vaswani et al., 2017) have achieved remarkable performance in many empirical tasks, such as answering questions and engaging in dialogues (Rajpurkar et al. 2016; Adiwardana et al. 2020). Despite this success, how these models acquire and encode linguistic information remains unclear (Avetisyan and Broneske, 2023). These models may reflect human-like gender biases at the semantic level of certain words like humans. However, there is little research on how these biases and semantic information are encoded by the models, and deep-learning-based representations have not engaged rigorously enough with semantic theory. It is still difficult to differentiate whether the model has genuinely progressed in modeling semantics or merely increased its ability to memorize corpus statistics (Pavlick, 2022).

Moreover, deep learning models are typically pre-trained on large contemporary corpora, and it is uncertain whether continued pre-training on historical corpora can help the models learn more human-like semantics and world knowledge related to gender of historical times (Qiu and Xu, 2022). Given that continued pre-training can be computationally expensive, it is necessary to determine which model achieves the most human-like word representation (Vulić et al., 2020). This includes traditional DSMs like Word2Vec (Mikolov et al., 2013), encoder-only models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), or decoder-only models like GPT-2 (Radford, 2019) and XLNet (Yang, et al., 2019).

By probing into the gender biases learned by these models from continued pre-training process, it is possible to study historical societal perceptions of gender bias that may be difficult to measure directly. This research makes several significant contributions: 1. Advancing research in historical sociolinguistics and cognitive bias by bridging the gap between sociolinguistics and deep-learning techniques. 2. Highlights the important role of historical corpora as a treasure trove for studying biases throughout history, allowing researchers to reconstruct historical societal attitudes and analyze biases in a more nuanced and precise manner. 3. Enabling historical bias research in data-scarce environments by demonstrating that models can learn biases from period-specific corpora, enabling historical bias research in contexts where direct data on societal attitudes may be scarce or non-existent. It opens new avenues for studying historical biases in diverse cultural and linguistic contexts.

*Jinghang Gu is the corresponding author with email: gujinghangnlp@gmail.com

2 Related Work

For lexical representations, traditional distributional semantics models (DSMs) include count-based methods like TF-IDF (Jones, 1973) and prediction-based methods such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017). These models produce a single word vector, potentially overlooking semantic differences across different contexts (McLevey et al., 2022).

In contrast, deep learning models aim to obtain sentence representations for real-life applications, with word representations emerging as a byproduct (Pavlick, 2022). Models include BERT (Devlin et al., 2018) and GPT-2/3 (Radford et al., 2019; Brown et al., 2020), which have been extensively studied for their semantic representation capabilities.

Contextualized embeddings have been shown to surpass traditional static word embeddings in capturing word semantics and identifying diachronic semantic shifts. Peters et al. (2018) and Radford et al. (2019) demonstrated that contextualized token embeddings encode word senses even without explicit training. Giulianelli et al. (2020) and Hu et al. (2019) developed contextualized embeddings for historical contexts, examining changes in word meanings over time. However, these studies often rely on models pre-trained on modern corpora, which may bias results towards contemporary language use (Qiu and Xu, 2022).

To address this, Hamilton et al. (2016) created HistWords, Word2Vec embeddings trained on historical corpora, to study semantic changes over 100 years of American history (Garg et al., 2018). Yet few works have extended this approach to contextualized language models. Vulić et al. (2020) compared contextualized word embeddings like BERT with traditional static DSMs like FastText, finding that contextualized embeddings generally outperform static ones. Gu et al. (2022) applied lexical semantics in embeddings for practical tasks. Nair et al. (2020) showed that contextualized embeddings have a higher correlation with human judgments. Yet Yenicelik et al. (2020) found that BERT embeddings' organization is "not purely determined by semantics." For world knowledge, previous studies have indicated that climate variations in language and world knowledge are closely linked (Huang and Dong, 2020; Dong and Huang, 2021). However, research comparing grounding and reference is notably absent (Pavlick, 2022). Some studies have explored multimodal variants (Sun et al., 2019; Radford et al., 2021), but they lack the semantic analysis depth of text-only models (Bender and Koller, 2020).

Gender bias is a critical topic across disciplines, with language analysis traditionally used to study it qualitatively (Holmes and Meyerhoff, 2004; Coates, 2015). Gender issues can be studied through machine learning techniques (Lu et al., 2022). If deep learning models can be continue pre-trained to better reflect human-like gender-related bias and world knowledge, they could become powerful tools for sociological and linguistic studies.

Inspired by previous works, this study aims to determine which model—BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-2 (Radford, 2019), or XLNet (Yang, et al., 2019) benefits the most from continued pre-training on historical corpora in terms of capturing more human-like gender-related attitudes, uncover how lexical semantics and world knowledge are encoded across model layers, and evaluate whether these models provide better human-like lexical representations compared to traditional static DSMs. By leveraging deep learning techniques, this research goes beyond previous traditional DSMs studies to explore the distribution of gender-related semantics and world knowledge within deep-learning model architectures, offering new insights into the intersection of language and society.

3 Methodology

The scarcity of historical quantitative data on gender bias in sociolinguistic research underscores the significance of this study. By using word representation as a quantitative tool, we aim to measure biases in historical societal changes. This study employs the 1990 Corpus of Historical American English (COHA) (Davis, 2010) for continued pre-training of BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-2 (Radford, 2019), and XL-Net (Yang et al., 2019).

For lexical semantics, we use the gender rating survey on different adjectives by Williams and Best as a benchmark (1990). For world knowledge, we use 1990 US demographic data on gender occupation participation (Ruggles et al., 2015) to validate the accuracy in the world knowledge dimension. We extract the corresponding word representations from the continue pre-trained models

and conduct linear regression analysis to identify the model that best reflects historical gender bias and gender-related world knowledge. Additionally, we aim to determine the best method for extracting gender-related word representations. The flowchart of the experiment can be seen in Figure 1:

3.1 Model Size

Previous empirical studies (Hu et al., 2020; Warstadt et al., 2020; Radford et al., 2019; Zhang et al., 2021) have shown that larger models tend to improve task performance and capture semantic information more effectively. Therefore, to compare the semantic representation capabilities of different models, it is essential that the models are comparable in size and are pre-trained on the same corpus. This research selects four transformer-based models of comparable sizes: encoder-only models BERT Base (Devlin et al., 2018) and RoBERTa Base (Liu et al., 2019), and decoder-only models GPT-2 small (Radford, 2019), and XLNet Base (Yang et al., 2019).

3.2 Layer Selection

Peters et al. (2018) and Tenney et al. (2019) found that lower hidden layers of BERT-based models tend to capture more syntactic information, while higher layers capture more abstract semantic information. This study plans to extract word representations from different layers of the models to explore the general distribution of semantic information and determine which specific layer best reflects human-like word representation.

3.3 Training Method

The choice between continued pre-training and fine-tuning is crucial. Fine-tuning a pre-trained model tends to yield better results for specific tasks (Wang et al., 2019). However, fine-tuning can alter the parameters of the higher hidden layers, potentially losing some linguistic information (Liu et al., 2019; Merchant et al., 2020; Mosbach et al., 2020). Since this study does not focus on any specific downstream task but aims to explore the general semantic representation of words in the context of a specific historical period, continued pre-training on historical data is more suitable and will be conducted here.

3.4 Evaluation

In terms of evaluation, NLP methods can be broadly divided into three categories (Pavlick, 2022): Extrinsic Task-Based Evaluation, Targeted Task-Based Evaluations (Linzen and Broni, 2020), and Representational or Probing Evaluation (Belinkov and Glass, 2019). This study aligns with the third category, as it investigates the model's understanding of semantic structure by extracting and analyzing word representations. By probing these representations, we aim to reveal how effectively the model captures underlying semantic patterns, including gender biases present in the data.

4 Experimental Settings

4.1 Pre-processing

This experiment selects the Corpus of Historical American English (COHA) (Davis, 2010) as the dataset due to its relatively large and balanced historical corpus. COHA contains over 475 million words from various genres, including fiction, nonfiction, newspapers, and magazines, spanning from the 1820s to the 2010s. Given the lack of systematic quantitative data on stereotypes in social science, this study utilizes the historical survey on gender stereotypes from 1990 (Williams and Best), so texts from 1990-1999 in COHA are selected as the training corpus for subsequent experiments. This subset contains 30,622,378 words and 2,374,121 sentences, with an average sentence length of 13 words.

For data processing, all text in COHA is converted to lowercase, and all punctuation marks are removed. Abbreviations are appropriately handled. The study uses the model's default tokenizer. Another important step is addressing possible mismatches between COHA and the tokenizer. For example, in COHA, words like "don't" are separated into "do" and "n't," which may cause issues during tokenization(Qiu and Xu, 2022). These are substituted back into their original forms. As suggested by Gulordava and Baroni (2011), lemmatization has little effect on the detection of semantic change, so it is not performed in the pre-processing process.

4.2 Continued Pre-training on Models and Representation Extraction

This study aims to continue pre-training BERT, RoBERTa, GPT-2, and XLNet using the COHA (1990) corpus and compare the results with a traditional Word2Vec model trained from scratch. The training starts from the last checkpoint of the original base models, following the official guidelines of

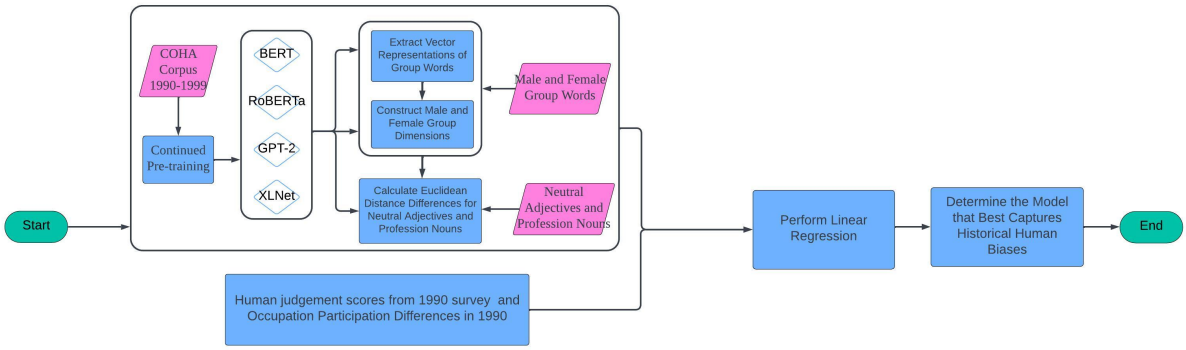


Figure 1: Flowchart of Continued Pre-training and Extraction of Gender-related Word Representation

Hugging Face. Since the average sentence length in the 1990s COHA is just 13 words, much shorter than the default 512, this research follows Qiu and Xu (2022) and limits the maximum sequence length to 128.

To thoroughly train the models, the number of epochs is set to be 300. However, if the loss becomes stable for a long time or the evaluation score stabilizes or starts to drop, early stopping will be applied. This research saves every checkpoint at the end of each epoch to analyze the dynamic changes in word representations. All training processes are completed on Alibaba Cloud using NVIDIA A10 cards with 24GB memory. The output models are stored in Alibaba Cloud OSS buckets.

For Word2Vec training, this research follows Hamilton et al. (2016). The symmetric context window size is set to 4 (on each side) with embeddings of size 300. The Word2Vec model is trained using the CBOW method with a smoothing parameter of 0.75. The negative sample prior is set to log(5), and the context is simply the same vocabulary as the target words.

Details of the models and configurations used are shown in Table 1 and Tale 2, which provides the necessary information to train the models using the openly available source code.

Model	Word2Vec
Parameters	Vocabulary Size*300
Embedding Size	300
Window Size	4
Smoothing Parameter	0.75
Negative Sample Prior	log(5)
sg	0

Table 1: Descriptions and Hyper-parameters of Word2Vec Training

This study aims to explore the semantic representation of words as a more "abstract" concept, rather than their representation in specific sentence contexts. Traditional static word embedding methods intuitively use distributional semantics to represent words, but deep learning models may differ from static word vectors. Firstly, words may be tokenized into sub-word tokens, which are influenced by the context and position within the sentence (Mickus et al., 2019). However, research by Vulić et al. (2020) demonstrates that pre-trained encoders still retain lexical semantics despite various contexts. This research adopts Vulić et al. (2020)'s unsupervised word-level representation strategies and configurations to probe the lexical semantics of words related to gender.

Model	BERT-Base	RoBERTa-Base
Parameters	110 million	125 million
Layers	12	12
Embedding Size	768	768
Max Sequence Length	128	128
Train Batch Size	64	64
Learning Rate	5e-05	5e-05
Optimizer	AdamW	AdamW
Gradient Clipping	1.0	1.0
Random Seed	42	42
Model	**XLNET-Base**	**GPT-2-Small**
Parameters	110 million	117 million
Layers	12	12
Embedding Size	768	768
Max Sequence Length	128	128
Train Batch Size	64	64
Learning Rate	5e-05	5e-05
Optimizer	AdamW	AdamW
Gradient Clipping	1.0	1.0
Random Seed	42	42

Table 2: Descriptions and Hyperparameters of Deep-learning Model Training

For all models used in this research—BERT Base (Devlin et al., 2018), RoBERTa Base (Liu

et al., 2019), GPT-2 Small (Radford et al., 2019), and XLNet Base(Yang et al., 2019)—each word representation is extracted in isolation without any external context. Special tokens [CLS] and [SEP] are excluded from sub-word embedding averaging. Two strategies are used for comparison: one is to extract only the representations from layer Ln, and the other is to average representations over all layers up to the n-th layer (including Ln).

4.3 Evaluation Metrics

This study utilizes historical survey data and objective records as benchmarks to evaluate the models' abilities to capture lexical semantics and reflect world knowledge.

To assess lexical semantics, we draw on the survey conducted by Williams and Best (1990), which measured people's perceptions of gender stereotypes using a list of adjectives. Participants provided scores indicating whether each adjective was perceived as more feminine or more masculine. For our study, we retained only the adjectives that appear in the Word2Vec vocabulary. The complete list of adjectives and their corresponding human-elicited scores are provided in the appendix.

For evaluating world knowledge, we use data from the 1990 U.S. Census (Ruggles et al., 2015) to calculate the gender disparity in occupational participation. This data serves as the "ground truth" or "objective metric" for societal gender roles at that time, reflecting historical realities. The full occupational participation data, broken down by gender, is also included in the appendix.

Building on the approach of Garg et al. (2018), we have constructed two "gender" dimensions: one for female-associated terms (e.g., "she," "her") and another for male-associated terms (e.g., "he," "his"). These lists, along with the lists of adjectives and occupations, are also available in the appendix. Words from the adjective and occupation lists are referred to as "neutral words" in this study.

To measure the association strength between neutral words and gender groups, we first create "gender group vectors" by averaging the vectors of words within the female and male groups. We then compute the Euclidean Distance between each neutral word's vector and the gender group vectors. This allows us to determine the relative norm distance of each neutral word concerning the male and female groups, from which the gender embedding bias of each word from each model is calculated. The bias score is defined as follows:

$$\text{Bias Score} = \|\vec{\mathbf{v}}_{\text{neutral}} - \vec{\mathbf{v}}_{\text{female}}\|_2 - \|\vec{\mathbf{v}}_{\text{neutral}} - \vec{\mathbf{v}}_{\text{male}}\|_2$$

The neutral vector represents the vector of a neutral word, and female and male vector denote the average vectors for the female and male groups respectively.

To measure embedding bias against historical data, this research follows Garg et al. (2018) that Ordinary Least Squares (OLS) linear regression analysis is conducted between the survey (or census) data and the gender embedding bias scores from the models to examine the correlation between the model's gender bias and the gender stereotypes as reflected in human.

R^2 (coefficient of determination) is used as an evaluation metric in this context because it quantifies the proportion of variance in the human survey data or census data that can be explained by the model's embedding biases (Montgomery et al., 2021). A higher R^2 value indicates a stronger correlation between the embedding bias captured by the models and the actual societal biases reflected in human data. The OLS linear regression analysis is conducted using the Python library statsmodels, which provides a robust framework for such statistical evaluations. This method allows us to precisely quantify the alignment between model-inferred biases and historical human biases, thus providing an objective measure of model performance.

5 Results and Discussion

5.1 Distribution of Gender-related Lexical Semantics and World Knowledge in Proto-models

To investigate the distribution and fundamental state of lexical semantics and world knowledge of gender-related words in different deep learning models, word representations for each neutral adjective and occupation noun were extracted from each layer of the original open-source models. The gender bias embedding score for each word in these models was calculated, followed by an analysis of the correlation strength between the model's gender bias scores, human bias scores, and occupation participation differences. The R^2 values from the OLS (Ordinary Least Squares) analysis for each layer across different models are presented in the line charts in Figure 2:

Figure 2 presents the R^2 values comparing the word representation bias scores of neutral adjectives at each layer of various proto-models with

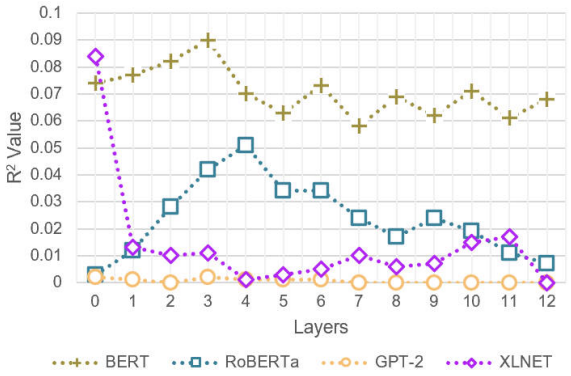


Figure 2: Lexical Semantic Representation of Adjectives in Each Layer of the Original Models (The R^2 value for Word2Vec is 0.099)

human bias scores obtained from the survey. Our analysis indicates that deep learning models, when used without fine-tuning or continued pretraining, do not perform adequately in analyzing diachronic semantic distributions. Specifically, the R^2 values for each layer of all deep learning models fell short of those obtained from traditional Word2Vec embeddings. Among the models evaluated, the original BERT model outperformed the others, followed by RoBERTa, XLNet, and GPT-2. Additionally, Figure 2 suggests that encoder-only models generally outperform decoder models in this task.

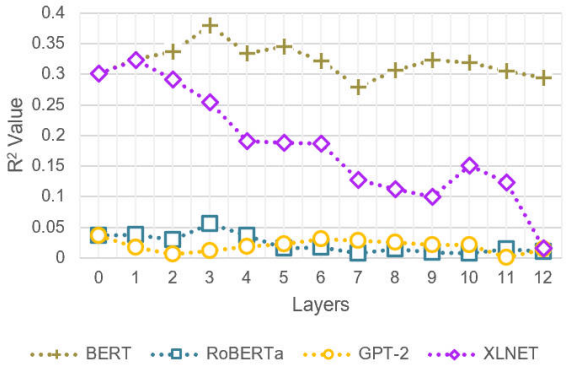


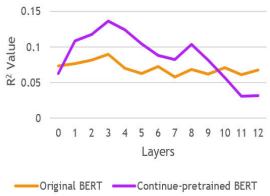
Figure 3: World Knowledge Representation of Occupation Nouns in Each Layer of the Original Models (The R^2 value for Word2Vec is 0.284)

Figure 3 illustrates the R^2 values between the word representation bias scores for occupation nouns at each layer of proto-models and the gender occupation participation data from official census. The results show that the performance of proto-models remains unsatisfactory when compared to the ground truth of gender occupation participation. Among the models, only BERT and XLNet exhibit slightly better performance.

Despite the presence of semantic information in each layer of deep learning models, this information tends to be dispersed across all layers. Generally, the core lexical semantics of words are predominantly concentrated in the lower-middle layers of most models.

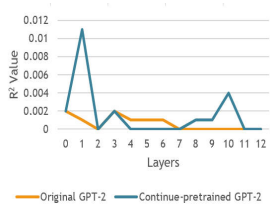
5.2 Effects of Continued Pre-training on Historical Corpus on Deep-learning Models

After continued pretraining of the afore-mentioned models using the 1990s COHA corpus, we extracted word representations from each layer of the trained models to assess their alignment with human similarity judgments and census data. We also evaluated whether the gender-related semantic representation abilities of the models were enhanced compared to their original versions.

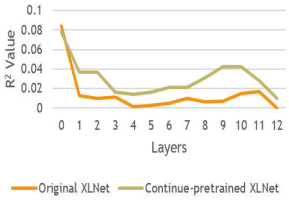


(a) R^2 Values of Adjectives for BERT

(b) R^2 Values of Adjectives for RoBERTa



(c) R^2 Values of Adjectives for GPT-2

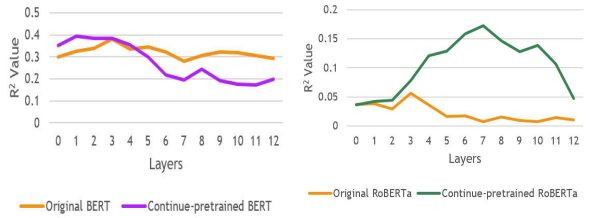


(d) R^2 Values of Adjectives for XLNet

Figure 4: Summary of R^2 Values of Adjectives of Each Layer from BERT, RoBERTa, GPT-2 and XLNet (before and after Continued Pretraining)

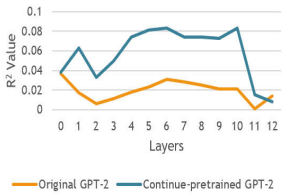
Figure 4 summarizes the R^2 values comparing adjectives' word representations from BERT, RoBERTa, GPT-2, and XLNet to human-elicited survey scores, evaluated across each model layer before and after continued pretraining on historical data. The results demonstrate that continued pretraining on a historical corpus generally enhances the models' ability to represent gender-related semantics, aligning them more closely with human judgments. Notably, BERT shows the most significant improvement in capturing nuanced gender associations, particularly in the lower-middle layers. RoBERTa and XLNet also exhibit enhanced perfor-

mance, though the gains are less consistent across all layers. GPT-2 shows the least improvement, reflecting the challenges faced by decoder-only architectures in modeling fine-grained gender biases. Overall, these findings underscore the importance of continued pretraining on domain-specific corpora to enrich the models' semantic representations and better reflect the complexities of human language understanding.

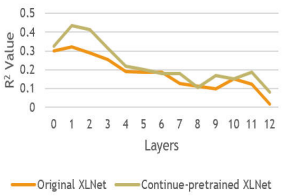


(a) R² Values of Occupation Nouns for BERT

(b) R² Values of Occupation Nouns for RoBERTa



(c) R² Values of Occupation Nouns for GPT-2



(d) R² Values of Occupation Nouns for XLNet

Figure 5: Summary of R² Values of Occupation Nouns of Each Layer from BERT, RoBERTa, GPT-2 and XLNet (before and after Continued Pretraining)

Figure 5 summarizes the R² values comparing occupation nouns' word representations from BERT, RoBERTa, GPT-2, and XLNet to real-world occupation participation differences across each model layer, both before and after continued pretraining on historical data. The results reveal that pretraining on a historical corpus also significantly enhances the models' ability to capture gender-related world knowledge. This improvement is especially evident in the lower-middle layers. BERT, in particular, show marked gains in representing gendered associations related to occupation nouns, indicating that the model benefit from integrating historical context to develop a deeper understanding of how gender roles have been encoded in language over time. Overall, these findings highlight the potential of continued pretraining on specific corpora to strengthen the semantic representation capabilities of deep-learning models, especially in areas that reflect societal attitudes and biases.

5.3 Comparison of Models' Human-Likeness in Lexical Semantics and World Knowledge

To determine which type of model best represents human-like semantic representations, we extract word vectors from each trained model and evaluate their correlation with human similarity judgments and census data. Two strategies are employed for word vector extraction: the first involves using only the representations from layer Ln, and the second is averaging representations across all layers up to the n-th layer (including Ln). Figure 6 presents the results for adjectives using only the representations from layer Ln, while Figure 7 shows the results for occupation nouns only the representations from layer Ln.

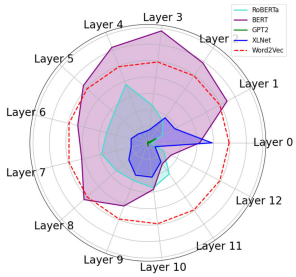


Figure 6: R² Values of Adjectives in Single Layers

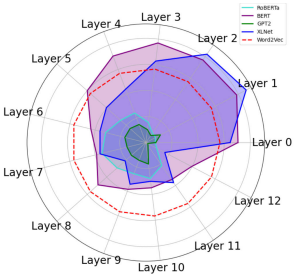


Figure 7: R² Values of Occupation Nouns in Single Layers

As shown in Figure 6 and Figure 7, after continued pretraining on historical data, only the lexical representations from layers 1-5 of BERT surpassed those of Word2Vec for both adjectives and occupation nouns. In contrast, individual layers from other models did not surpass Word2Vec. Additionally, the trend observed indicates that type-level lexical information is more concentrated in the lower layers, approximately layers 1-5.

Figure 8 and Figure 9 present the results for adjectives and occupation nouns by averaging representations across all layers up to the n-th layer (including Ln).

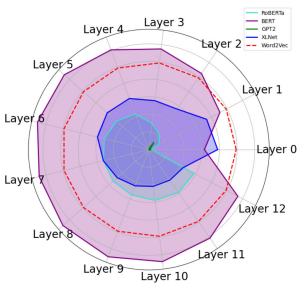


Figure 8: Average Layers of Adjectives Representation

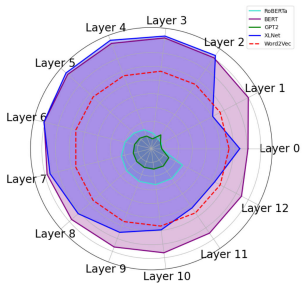


Figure 9: Average Layers of Occupation Representation

Since the original BERT and XLNet models already outperform Word2Vec in terms of world knowledge for occupation nouns, it is expected that these models continue to surpass Word2Vec even after training. However, the distribution of information in BERT and XLNet has shifted from being relatively uniform to being more concentrated in the lower layers, with information primarily concentrated in layers 0-6. In contrast, GPT-2 and RoBERTa perform relatively poorly and do not exceed Word2Vec.

The results indicate that only BERT achieved notably positive outcomes, demonstrating the best correlation with human annotations. According to Vulić et al. (2020), the performance of individual layers can be task and language dependent, while averaging across all layers might sometimes reduce performance, averaging across the bottom-most layers is generally beneficial. For this study, which aims to reflect historical human gender bias more accurately, averaging up to the 6th layer (inclusive) is recommended.

To verify the stability of these results, we further analyzed R^2 values using 10 different random seeds to check whether they remained consistent. Figure 10 shows the Mean Absolute Deviation (MAD) error of R^2 values for each layer. The R^2 values are generally stable, with the standard deviation for the average 6th layer result being approximately 0.01, indicating that the training is relatively robust and stable.

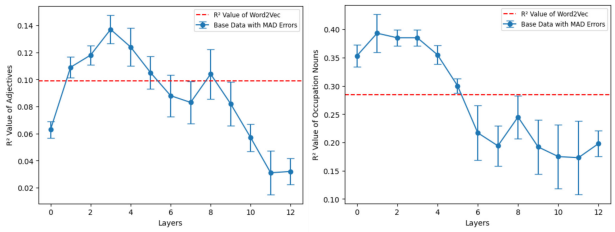


Figure 10: R^2 Values with Adjectives and Occupation Nouns of MAD Errors from Random Seed Training Models

6 Conclusion

These findings demonstrate that continued pre-training on historical data is effective for historical semantic analysis, particularly for examining historical gender bias. Our findings reveal that gender-related type-level semantic information is primarily concentrated in the lower-middle layers of deep-learning models, with an optimal strategy being to average representations up to the 6th layer. This approach allows for a more accurate reflection of historical human biases, as evidenced by BERT's performance, which outperformed both static word embeddings like Word2Vec and other transformer-based models in generating human-like semantic representations.

Overall, while each layer of a deep-learning model contributes to capturing different aspects of semantic information, type-level lexical information is predominantly concentrated in the lower-middle layers. The optimal performance of specific layer can vary based on the language and task, however, averaging representations across all layers up to a certain point generally proves to be a more robust approach.

6.1 Significance and Implications

This study has several key implications. First, it shows that continued pre-training on historical corpora can enhance deep learning models' ability to represent gender biases in ways that align with human understanding, supporting sociocultural research.

Second, the findings highlight the strengths of models like BERT in capturing linguistic nuances that simpler models, like Word2Vec, might miss. This study offers practical guidance for optimizing model design and application across linguistic tasks.

Lastly, it demonstrates that deep-learning models can reveal hidden patterns of bias even in data-scarce environments, enhancing historical analysis.

6.2 Limitations and Future Directions

The study focuses on a specific period (the 1990s) and one type of bias (gender-related). Future research could explore other time periods, biases, and cultural settings to broaden our understanding.

Additionally, the study did not examine other fine-tuning strategies or larger models. Future work could investigate domain-specific fine-tuning strategies and assess if larger models provide more precise representations of historical biases.

6.3 Summary

In summary, our research suggests that deep-learning models pre-trained on historical data are powerful tools for semantic analysis. By understanding how these models distribute information across layers, researchers can better explore the evolution of language and bias. This work lays the groundwork for refining model training tech-

niques, expanding linguistic corpora, and uncovering deeper insights into the relationship between language, culture, and society.

Acknowledgements

This study was funded by The Hong Kong Polytechnic University Projects (#P0048932, #P0051089).

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., & Le, Q. V. (2020). Towards a human-like open-domain chatbot. https://doi.org/10.48550/arXiv.2001.09977

Avetisyan, H., & Broneske, D. (2023). Decoding the encoded–linguistic secrets of language models: A systematic literature review. *CS & IT Conference Proceedings*, *13*(16). https://doi.org/10.5121/csit.2023.131606

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72. https://doi.org/10.48550/arXiv.1812.08951

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.48550/arXiv.1607.04606

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge. https://doi.org/10.4324/9781315645612

Davies, M. (2010). The corpus of historical american english (version 3.0) [[Accessed: 1,27,2024]]. https://corpus.byu.edu/coha/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Dong, S., & Huang, C.-R. (2021). From falling to hitting: Diachronic change and synchronic distribution of frost verbs in chinese. *Workshop on Chinese Lexical Semantics*, 22–30. https://doi.org/10.1007/978-3-031-06703-7_2

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: Computer applications* (pp. 231–243). Springer.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973. https://doi.org/10.18653/v1/2020.acl-main.365

Gu, J., Xiang, R., Wang, X., Li, J., Li, W., Qian, L., Zhou, G., & Huang, C.-R. (2022a). Multi-probe attention neural network for covid-19 semantic indexing. *BMC bioinformatics*, *23*(1), 259. https://doi.org/10.1186/s12859-022-04803-x

Gulordava, K., & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, 67–71. https://aclanthology.org/W11-2508

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*. https://doi.org/10.48550/arXiv.1605.09096

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Holmes, J., & Meyerhoff, M. (2004). *The handbook of language and gender*. John Wiley & Sons. https://doi.org/10.1002/9780470756942

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*. https://doi.org/10.18653/v1/2020.acl-main.158

Hu, R., Li, S., & Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 3899–3908. https://doi.org/10.18653/v1/P19-1379

Huang, C.-R., & Dong, S. (2020). From lexical semantics to traditional ecological knowledge: On precipitation, condensation and suspension expressions in chinese. *Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers 20*, 255–264. https://doi.org/10.1007/978-3-030-38189-9_27

Jones, K. S. (1973). Index term weighting. *Information storage and retrieval*, *9*(11), 619–633. https://doi.org/10.1016/0020-0271(73)90043-0

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*. https://doi.org/10.48550/arXiv.1903.08855

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Lu, L., Gu, J., & Huang, C.-R. (2022). Inclusion in csr reports: The lens from a data-driven machine learning model. *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, 46–51. https://aclanthology.org/2022.csrnlp-1.7

McLevey, J. V., Crick, T., Browne, P., & Durant, D. (2022). A new method for computational cultural cartography: From neural word embeddings to transformers and bayesian mixture models. *Canadian Review of Sociology/Revue canadienne de sociologie*, *59*(2), 228–250. https://doi.org/10.1111/cars.12378

Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020). What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*. https://doi.org/10.48550/arXiv.2004.14448

Mickus, T., Paperno, D., Constant, M., & Van Deemter, K. (2019). What do you mean, bert? assessing bert as a distributional semantics model. *arXiv preprint arXiv:1911.05758*. https://doi.org/10.48550/arXiv.1911.05758

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*. https://doi.org/10.48550/arXiv.1310.4546

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons. https://doi.org/10.1111/biom.12129

Mosbach, M., Khokhlova, A., Hedderich, M. A., & Klakow, D. (2020). On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. *arXiv preprint arXiv:2010.02616*. https://doi.org/10.48550/arXiv.2010.02616

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized word embeddings encode aspects of human-like word sense knowledge. *arXiv preprint arXiv:2010.13057*. https://doi.org/10.48550/arXiv.2010.13057

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*(1), 447–471. https://doi.org/10.1146/annurev-linguistics-031120-122924

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://doi.org/10.48550/arXiv.1802.05365

Qiu, W., & Xu, Y. (2022). Histbert: A pre-trained language model for diachronic lexical semantic analysis. *arXiv preprint arXiv:2202.03612*. https://doi.org/10.13140/RG.2.2.14905.44649

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell,

A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9. https://api.semanticscholar.org/CorpusID:160025533

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. https://doi.org/10.48550/arXiv.1606.05250

Ruggles, S., Genadek, K., Goeken, R., Grover, J., Sobek, M., et al. (2015). Integrated public use microdata series: Version 6.0 [dataset]. *Minneapolis: University of Minnesota*, *23*, 56.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. *Proceedings of the IEEE/CVF international conference on computer vision*, 7464–7473. https://doi.org/10.48550/arXiv.1904.01766

Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*. https://doi.org/10.18653/v1/P19-1452

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing pretrained language models for lexical semantics. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7222–7240. https://doi.org/10.18653/v1/2020.emnlp-main.586

Wang, A., Hula, J., Xia, P., Pappagari, R., McCoy, R. T., Patel, R., Kim, N., Tenney, I., Huang, Y., Yu, K., et al. (2019). Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. *arXiv preprint arXiv:1812.10860*. https://doi.org/10.48550/arXiv.1812.10860

Warstadt, A., Zhang, Y., Li, H.-S., Liu, H., & Bowman, S. R. (2020). Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2010.05358*. https://doi.org/10.18653/v1/2020.emnlp-main.16

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study, rev*. Sage Publications, Inc. https://api.semanticscholar.org/CorpusID:149173046

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*. https://doi.org/10.48550/arXiv.1906.08237

Yenicelik, D., Schmidt, F., & Kilcher, Y. (2020). How does bert capture semantics? a closer look at polysemous words. *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 156–162. https://doi.org/10.18653/v1/2020.blackboxnlp-1.15

Zhang, Y., Warstadt, A., Li, H.-S., & Bowman, S. R. (2021). When do you need billions of words of pretraining data? https://doi.org/10.18653/v1/2021.acl-long.90

Zhu, Y. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*. https://doi.org/10.48550/arXiv.1506.06724

A Appendix

A: Group Words

Man Words: he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews

Woman Words: she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, females, sisters, aunt, aunts, niece, nieces

B: Neutral Words

Occupations: bookbinder, waitstaff, laborer, sailor, technician, porter, chemist, electrician, inspector, salesperson, secretary, plumber, doctor, mechanic, instructor, carpenter, upholsterer, shoemaker, bartender, chiropractor, nutritionist, pharmacist, administrator, surgeon, geologist, teacher, painter, soldier, photographer, attendant, economist, janitor, clergy, peddler, auctioneer, artist, dentist, driver, dancer, cashier, cook, sheriff,

nurse, compositor, author, lawyer, conductor, manager, postmaster, dietitian, architect, gardener, optometrist, housekeeper, sales, accountant, molder, draftsperson, clerical, typesetter, musician, plasterer, machinist, newsperson, pilot, baker, weaver, therapist, entertainer, police, jeweler, boilermaker, bailiff, operator, surveyor, psychologist, professor, engineer, judge, proprietor, librarian, broker, millwright, welder, designer, lumberjack, toolmaker, setter, huckster, clerk, smith, athlete, tailor, scientist, mathematician, farmer, veterinarian, official, statistician, physician, conservationist, cabinetmaker, guard, doorkeeper, mason, physicist

Adjectives: active, adaptable, adventurous, affected, affectionate, aggressive, alert, aloof, ambitious, anxious, apathetic, appreciative, argumentative, arrogant, artistic, assertive, attractive, autocratic, awkward, bitter, blustery, boastful, bossy, calm, capable, careless, cautious, changeable, charming, cheerful, civilized, clever, coarse, cold, commonplace, complaining, complicated, conceited, confident, confused, conscientious, conservative, considerate, contented, conventional, cool, cooperative, courageous, cowardly, cruel, curious, cynical, daring, deceitful, defensive, deliberate, demanding, dependable, dependent, despondent, determined, dignified, discreet, disorderly, dissatisfied, distrustful, dominant, dreamy, dull, effeminate, efficient, egotistical, emotional, energetic, enterprising, enthusiastic, evasive, excitable, fearful, feminine, fickle, flirtatious, foolish, forceful, foresighted, forgetful, forgiving, formal, frank, friendly, frivolous, fussy, generous, gentle, gloomy, greedy, handsome, hasty, headstrong, healthy, helpful, honest, hostile, humorous, hurried, idealistic, imaginative, immature, impatient, impulsive, independent, indifferent, individualistic, industrious, infantile, informal, ingenious, inhibited, initiative, insightful, intelligent, intolerant, inventive, irresponsible, irritable, jolly, kind, lazy, leisurely, logical, loud, loyal, mannerly, masculine, mature, meek, methodical, mild, mischievous, moderate, modest, moody, nagging, natural, nervous, noisy, obliging, obnoxious, opinionated, opportunistic, optimistic, organized, original, outgoing, outspoken, painstaking, patient, peaceable, peculiar, persevering, persistent, pessimistic, pleasant, poised, polished, practical, praising, precise, prejudiced, preoccupied, progressive, prudish, quarrelsome, queer, quick, quiet, quitting, rational, realistic, reasonable, rebellious, reckless, reflective, relaxed, reliable, resentful, reserved, resourceful, responsible, restless, retiring, rigid, robust, rude, sarcastic, selfish, sensitive, sentimental, serious, severe, sexy, shallow, shiftless, shrewd, shy, silent, simple, sincere, slipshod, slow, sly, smug, snobbish, sociable, sophisticated, spendthrift, spineless, spontaneous, spunky, stable, steady, stern, stingy, stolid, strong, stubborn, submissive, suggestible, sulky, superstitious, suspicious, sympathetic, tactful, tactless, talkative, temperamental, tense, thankless, thorough, thoughtful, thrifty, timid, tolerant, touchy, tough, trusting, unaffected, unambitious, unassuming, unconventional, undependable, understanding, unemotional, unfriendly, uninhibited, unintelligent, unkind, unrealistic, unscrupulous, unselfish, unstable, vindictive, versatile, warm, wary, weak, whiny, wholesome, wise, withdrawn, witty, worrying, zany

C: Occupation Participation Census Data (1990)

Occupation	Percentage Difference
bookbinder	0.12
waitstaff	0.65
laborer	-0.63
sailor	-0.93
technician	-0.08
porter	-0.79
chemist	-0.46
electrician	-0.95
inspector	-0.51
salesperson	-0.18
secretary	0.96
plumber	-0.97
doctor	-0.58
mechanic	-0.32
instructor	-0.16
carpenter	-0.96
upholsterer	-0.50
shoemaker	-0.82
bartender	0.05
chiropractor	-0.37
nutritionist	0.80
pharmacist	-0.26
bankteller	0.80
administrator	0.08
surgeon	-0.58
geologist	-0.69
teacher	0.49
painter	-0.64
soldier	-0.78
photographer	-0.32

Occupation	Percentage Difference
attendant	0.60
economist	-0.12
janitor	-0.11
clergy	-0.54
peddler	0.38
auctioneer	-0.68
artist	0.09
dentist	-0.73
driver	-0.75
dancer	0.57
cashier	0.62
cook	0.04
sheriff	-0.62
nurse	0.84
compositor	0.37
author	0.01
lawyer	-0.49
fireperson	-0.92
conductor	-0.88
manager	-0.29
postmaster	-0.05
dietitian	0.80
architect	-0.64
gardener	-0.83
optometrist	-0.69
housekeeper	0.86
sales	-0.03
accountant	0.07
molder	-0.67
draftsperson	-0.62
clerical	0.43
typesetter	0.37
musician	0.19
plasterer	-0.96
machinist	-0.90
newsperson	0.05
pilot	-0.92
baker	-0.03
weaver	0.34
therapist	0.52
entertainer	0.01
police	-0.71
jeweler	-0.42
boilermaker	-0.95
bailiff	-0.62
operator	0.16
surveyor	-0.76
psychologist	0.19
professor	-0.16
engineer	-0.77
judge	-0.49
mailperson	-0.51
tradesperson	-0.89
proprietor	-0.28
librarian	0.77
broker	0.02
millwright	-0.93
welder	-0.90
designer	0.17
lumberjack	-0.73
toolmaker	-0.95
setter	-0.95
huckster	0.38
clerk	-0.39
smith	-0.90
athlete	-0.42
tailor	0.01
scientist	-0.38
mathematician	-0.41
farmer	-0.70
veterinarian	-0.45
official	-0.25
statistician	-0.09
physician	-0.58
conservationist	-0.70
cabinetmaker	-0.85
guard	-0.64
doorkeeper	-0.64
mason	-0.97
physicist	-0.75

Table 3: Occupation Participation Census Data(1990)

D Williams and Best Survey (1990)

word	year	score	transformed score
absent-minded	1990	60	-100
active	1990	81	-310
adaptable	1990	37	130
adventurous	1990	93	-430
affected	1990	20	300
affectionate	1990	10	400
aggressive	1990	88	-380
alert	1990	60	-100
aloof	1990	50	0
ambitious	1990	82	-320
anxious	1990	23	270
apathetic	1990	53	-30
appreciative	1990	26	240
argumentative	1990	59	-90

word	year	score	transformed score	word	year	score	transformed score
arrogant	1990	74	-240	discreet	1990	49	10
artistic	1990	34	160	disorderly	1990	76	-260
assertive	1990	73	-230	dissatisfied	1990	42	80
attractive	1990	14	360	distractible	1990	40	100
autocratic	1990	86	-360	distrustful	1990	45	50
awkward	1990	64	-140	dominant	1990	87	-370
bitter	1990	51	-10	dreamy	1990	17	330
blustery	1990	65	-150	dull	1990	56	-60
boastful	1990	77	-270	easy-going	1990	64	-140
bossy	1990	68	-180	effeminate	1990	41	90
calm	1990	48	20	efficient	1990	63	-130
capable	1990	70	-200	egotistical	1990	77	-270
careless	1990	65	-150	emotional	1990	12	380
cautious	1990	33	170	energetic	1990	82	-320
changeable	1990	28	220	enterprising	1990	81	-310
charming	1990	19	310	enthusiastic	1990	51	-10
cheerful	1990	36	140	evasive	1990	46	40
civilized	1990	48	20	excitable	1990	33	170
clear-thinking	1990	71	-210	fair-minded	1990	59	-90
clever	1990	64	-140	fault-finding	1990	33	170
coarse	1990	91	-410	fearful	1990	17	330
cold	1990	58	-80	feminine	1990	8	420
commonplace	1990	54	-40	fickle	1990	27	230
complaining	1990	21	290	flirtatious	1990	35	150
complicated	1990	30	200	foolish	1990	33	170
conceited	1990	68	-180	forceful	1990	93	-430
confident	1990	77	-270	foresighted	1990	58	-80
confused	1990	33	170	forgetful	1990	58	-80
conscientious	1990	45	50	forgiving	1990	33	170
conservative	1990	53	-30	formal	1990	61	-110
considerate	1990	35	150	frank	1990	65	-150
contented	1990	43	70	friendly	1990	42	80
conventional	1990	54	-40	frivolous	1990	28	220
cool	1990	64	-140	fussy	1990	24	260
cooperative	1990	46	40	generous	1990	55	-50
courageous	1990	86	-360	gentle	1990	21	290
cowardly	1990	45	50	gloomy	1990	56	-60
cruel	1990	79	-290	good-looking	1990	36	140
curious	1990	24	260	good-natured	1990	51	-10
cynical	1990	69	-190	greedy	1990	67	-170
daring	1990	86	-360	handsome	1990	69	-190
deceitful	1990	52	-20	hard-headed	1990	74	-240
defensive	1990	43	70	hard-hearted	1990	77	-270
deliberate	1990	61	-110	hasty	1990	54	-40
demanding	1990	48	20	headstrong	1990	71	-210
dependable	1990	56	-60	healthy	1990	69	-190
dependent	1990	19	310	helpful	1990	35	150
despondent	1990	36	140	high-strung	1990	32	180
determined	1990	78	-280	honest	1990	55	-50
dignified	1990	53	-30	hostile	1990	66	-160

word	year	score	transformed score	word	year	score	transformed score
humorous	1990	73	-230	organized	1990	55	-50
hurried	1990	55	-50	original	1990	60	-100
idealistic	1990	54	-40	outgoing	1990	64	-140
imaginative	1990	32	180	outspoken	1990	66	-160
immature	1990	48	20	painstaking	1990	44	60
impatient	1990	59	-90	patient	1990	32	180
impulsive	1990	44	60	peaceable	1990	35	150
independent	1990	84	-340	peculiar	1990	50	0
indifferent	1990	69	-190	persevering	1990	60	-100
individualistic	1990	71	-210	persistent	1990	63	-130
industrious	1990	60	-100	pessimistic	1990	50	0
infantile	1990	44	60	planful	1990	63	-130
informal	1990	84	-340	pleasant	1990	26	240
ingenious	1990	69	-190	pleasure-seeking	1990	68	-180
inhibited	1990	42	80	poised	1990	44	60
initiative	1990	75	-250	polished	1990	45	50
insightful	1990	58	-80	practical	1990	63	-130
intelligent	1990	68	-180	praising	1990	44	60
interests narrow	1990	34	160	precise	1990	67	-170
interests wide	1990	73	-230	prejudiced	1990	48	20
intolerant	1990	65	-150	preoccupied	1990	57	-70
inventive	1990	81	-310	progressive	1990	78	-280
irresponsible	1990	63	-130	prudish	1990	24	260
irritable	1990	50	0	quarrelsome	1990	43	70
jolly	1990	59	-90	queer	1990	63	-130
kind	1990	29	210	quick	1990	72	-220
lazy	1990	73	-230	quiet	1990	37	130
leisurely	1990	59	-90	quitting	1990	43	70
logical	1990	79	-290	rational	1990	75	-250
loud	1990	76	-260	rattlebrained	1990	34	160
loyal	1990	42	80	realistic	1990	75	-250
mannerly	1990	48	20	reasonable	1990	63	-130
masculine	1990	96	-460	rebellious	1990	61	-110
mature	1990	56	-60	reckless	1990	74	-240
meek	1990	25	250	reflective	1990	53	-30
methodical	1990	60	-100	relaxed	1990	59	-90
mild	1990	22	280	reliable	1990	61	-110
mischievous	1990	63	-130	resentful	1990	40	100
moderate	1990	48	20	reserved	1990	41	90
modest	1990	32	180	resourceful	1990	70	-200
moody	1990	39	110	responsible	1990	65	-150
nagging	1990	30	200	restless	1990	68	-180
natural	1990	53	-30	retiring	1990	52	-20
nervous	1990	28	220	rigid	1990	74	-240
noisy	1990	65	-150	robust	1990	85	-350
obliging	1990	40	100	rude	1990	83	-330
obnoxious	1990	72	-220	sarcastic	1990	61	-110
opinionated	1990	67	-170	self-centered	1990	61	-110
opportunistic	1990	72	-220	self-confident	1990	79	-290
optimistic	1990	58	-80	self-controlled	1990	64	-140

word	year	score	transformed score
self-denying	1990	36	140
self-pitying	1990	30	200
self-punishing	1990	47	30
self-seeking	1990	59	-90
selfish	1990	61	-110
sensitive	1990	14	360
sentimental	1990	11	390
serious	1990	74	-240
severe	1990	81	-310
sexy	1990	14	360
shallow	1990	36	140
sharp-witted	1990	68	-180
shiftless	1990	60	-100
show-off	1990	67	-170
shrewd	1990	60	-100
shy	1990	25	250
silent	1990	42	80
simple	1990	45	50
sincere	1990	44	60
slipshod	1990	63	-130
slow	1990	50	0
sly	1990	60	-100
smug	1990	64	-140
snobbish	1990	44	60
sociable	1990	43	70
soft-hearted	1990	19	310
sophisticated	1990	28	220
spendthrift	1990	46	40
spineless	1990	52	-20
spontaneous	1990	49	10
spunky	1990	63	-130
stable	1990	71	-210
steady	1990	70	-200
stern	1990	84	-340
stingy	1990	69	-190
stolid	1990	76	-260
strong	1990	92	-420
stubborn	1990	63	-130
submissive	1990	16	340
suggestible	1990	26	240
sulky	1990	45	50
superstitious	1990	13	370
suspicious	1990	35	150
sympathetic	1990	27	230
tactful	1990	47	30
tactless	1990	62	-120
talkative	1990	22	280
temperamental	1990	34	160
tense	1990	53	-30
thankless	1990	66	-160
thorough	1990	59	-90
thoughtful	1990	47	30
thrifty	1990	46	40
timid	1990	25	250
tolerant	1990	45	50
touchy	1990	27	230
tough	1990	91	-410
trusting	1990	42	80
unaffected	1990	72	-220
unambitious	1990	30	200
unassuming	1990	44	60
unconventional	1990	59	-90
undependable	1990	53	-30
understanding	1990	33	170
unemotional	1990	82	-320
unexcitable	1990	70	-200
unfriendly	1990	67	-170
uninhibited	1990	66	-160
unintelligent	1990	32	180
unkind	1990	74	-240
unrealistic	1990	35	150
unscrupulous	1990	72	-220
unselfish	1990	45	50
unstable	1990	32	180
vindictive	1990	49	10
versatile	1990	61	-110
warm	1990	27	230
wary	1990	47	30
weak	1990	17	330
whiny	1990	23	270
wholesome	1990	57	-70
wise	1990	77	-270
withdrawn	1990	40	100
witty	1990	67	-170
worrying	1990	27	230
zany	1990	67	-170

Table 4: Williams and Best Survey (1990)