# Conditional Word Embedding and Hypothesis Testing via Bayes-by-Backprop

**Rujun Han**
Information Sciences Institute
University of Southern California
rujunhan@usc.edu

**Michael Gill**
Center for Data Science
New York University
mzgill@nyu.edu

**Arthur Spirling**
Center for Data Science
New York University
arthur.spirling@nyu.edu

**Kyunghyun Cho**
Center for Data Science
New York University
CIFAR Global Scholar
kyunghyun.cho@nyu.edu

## Abstract

Conventional word embedding models do not leverage information from document metadata, and they do not model uncertainty. We address these concerns with a model that incorporates document covariates to estimate conditional word embedding distributions. Our model allows for (a) hypothesis tests about the meanings of terms, (b) assessments as to whether a word is near or far from another conditioned on different covariate values, and (c) assessments as to whether estimated differences are statistically significant.

## 1 Introduction

Whether a word's meaning varies across contexts has become a major focus of NLP, linguistics, and social science research in recent years. For example, since the early 20th century, the word "gay" has evolved from describing an emotion to being more aligned with sexual orientation (Hamilton et al., 2016b). Popular word embedding techniques (e.g., Mikolov et al., 2013a; Pennington et al., 2014) have proven useful for analyzing language evolution. But to use these models for such research, scholars often divide a corpus into distinct training sets (e.g., train independent language models on different decades of text) and compare model output across specifications in an *ad hoc* way (Garg et al., 2018). Such splitting inhibits many within- and across-word comparisons, since embeddings are only comparable within a given model. Additionally, most methods ignore the *variance* of words, mechanically treating words equally regardless of the volatility, or uncertainty, in their meanings. If one inspects semantics with only point estimates of embeddings, it is hard to tell whether embeddings represent meaningful traits or are simply noise in the data.

We address these concerns in three ways. First, we estimate a vector for each distinct value of the document covariates, using a multilayer perceptron (MLP) with a non-linear activation function. Second, we parametrize the covariance matrix of each embedding vector explicitly in the model, adopting the Bayes-by-Backprop algorithm (Blundell et al., 2015). Third, we utilize Hotelling $T^2$ statistics (Hotelling, 1931) to assess whether estimated differences in word vectors are statistically differentiable under a null $\chi^2$ distribution (Ito, 1956). To our knowledge, no prior work evaluates word embeddings with this statistical framework.

## 2 Related Work

**Drift Analysis using Word Embeddings** There are several ways to measure drifts in word meanings. Hamilton et al. (2016c) propose the use of cosine similarities of words in different contexts to detect changes. Hamilton et al. (2016b) provide an alternative measure based on the distance of words from their nearest neighbors. Rudolph and Blei (2018) analyze absolute drift of words using Euclidean distance in (two discrete) slices of data. All of these methods compute the word distance based only on the point (i.e., mean) estimates of the word embeddings.

**Conditional Word Embedding** Rudolph and Blei (2018) estimate dynamic Bernoulli embeddings (DBE), extending the exponential family embedding (Rudolph et al., 2016) generalization of Mikolov et al. (2013a), to learn conditional word embeddings over time. Their amortized approach builds a separate neural network that transforms a global word vector into a covariate-specific vector, and is closely related to our approach in this paper. However, a noticeable omission in their model is that they do not explicitly model parameter covariance or uncertainty.

**Word Embedding with Uncertainty** Vilnis and McCallum (2017) earlier proposed an energy-

based learning framework in which each word is represented as a multivariate Gaussian distribution with a diagonal covariance. The energy function is defined by the divergence (e.g., KL) between two Gaussian embeddings, and the margin ranking loss (Weston et al., 2011) is minimized. A related model is the Bayesian skip-gram in Brazinskas et al. (2017), which posits a generative model where words are associated with multivariate Gaussian latent variables that generate context words. The parameters of those prior distributions over the multivariate Gaussian latent variables are estimated by maximizing the variational lowerbound, and act as word embeddings.

These works replace mean estimates of embeddings with Gaussian distributions, similar to our proposal here. However, they arrive at this differently; Vilnis and McCallum (2017) from the energy-based learning (LeCun et al., 2006), and Brazinskas et al. (2017) from generative modeling. We provide yet another angle: via (approximate) Bayesian neural networks.

## 3 Conditional Word Embedding

**Adopting Bayes-by-Backprop for Estimation** Given a tuple of a word $v$, a covariate $x$ and a context word $v_c$, we define the conditional log-probability as

$$\log p(v_c|v,x) = \theta_{v|x}^\top \theta_{v_c}^c - \log \sum_{v_c' \in V} \exp\left(\theta_{v|x}^\top \theta_{v_c'}^c\right),$$

where $\theta_{v|x}$ and $\theta_{v_c}^c$ are the conditional word embedding of $v$ given $x$ and the context embedding of $v_c$, respectively. $V$ is the vocabulary of all unique words. To avoid the expensive computation of the partition function, we use negative sampling (Mikolov et al., 2013b), which stochastically approximates the log-probability above by:

$$\begin{aligned}\log p(v_c|v,x) \approx & \log \sigma(\theta_{v|x}^\top \theta_{v_c}^c) \\ & + \frac{1}{M}\sum_{m=1}^{M} \log(1-\sigma(\theta_{v|x}^\top \theta_{v_c^m}^c)),\end{aligned} \tag{1}$$

where $v_c^m \in V$ is the $m$-th negative sample drawn from a unigram distribution estimated from $D$.

We define a prior distribution over each parameter $\theta$ to be a scaled mixture of two Gaussians, as suggested by Blundell et al. (2015):

$$\begin{aligned}\log p(\theta_i) = & \log\left(u\mathcal{N}(\theta_i|0,\sigma_1^2)\right. \\ & \left.+(1-u)\mathcal{N}(\theta_i|0,\sigma_2^2)\right),\end{aligned} \tag{2}$$

where $\sigma_1$, $\sigma_2$ and $u$ are the hyperparameters.

As exactly marginalizing out the parameters $\theta_{\cdot}$ and $\theta_{\cdot}^c$ is not scalable, we maximize the variational lowerbound of the marginal probability. To do so, we introduce a variational posterior $q(\theta|\phi)$ parametrized by its own parameter set $\phi$. Then, the variational lowerbound is defined as $-\mathcal{F}(\theta, D) = \mathbb{E}_q\left[\log p(D|\theta)\right] - \mathrm{KL}(q(\theta)\|p(\theta))$, where $\log p(D|\theta) = \sum_{(v,x,v_c)\in D} \log p(v_c|v,x)$ in our case. This is stochastically approximated by

$$\begin{aligned}-\mathcal{F}(\theta,D) \approx & \frac{1}{M}\sum_{m=1}^{M} \log p(D|\theta^{(m)}) \\ & - \log q(\theta^{(m)}|\phi) + \log p(\theta^{(m)}),\end{aligned} \tag{3}$$

where $\theta^{(m)}$ is the $m$-th sample from the variational posterior $q$ (Blundell et al., 2015) via the Gaussian reparametrization in Kingma and Welling (2013). We formulate the variational posterior as a multivariate Gaussian with diagonal covariance.

We use stochastic gradient descent (SGD) to minimize $\mathcal{F}$ with respect to the variational parameters $\phi$. At each SGD step, we compute the gradient of the following per-example cost given an example $(v, v_c, x) \in D$:

$$\begin{aligned}f(\theta,(v,v_c,x)) \approx & -\log p(v_c|v,x) + \log q(\tilde{\theta}_{v|x}|\phi) \\ & + \log q(\tilde{\theta}_{v_c}^c) + \log q(\tilde{\theta}_{v_c'}^c) - \log p(\tilde{\theta}),\end{aligned}$$

where $\tilde{\theta}$ is a single sample from the approximate posterior, and $\log p(v_c|v,x)$ and $\log p(\tilde{\theta})$ are from Eqs. (1)–(2). We then estimate the (approximate) posterior distribution of each conditional word embedding $\theta_{v|x}$ rather than its point estimate, by minimizing $\mathcal{F}$. See Sec. A of the supplementary material for the detailed steps for computing the per-example cost.

**Parametrized Conditional Word Embedding** An issue with the approach described so far is the number of parameters grows linearly in the size of the vocabulary and in the number of covariate partitions, i.e., $O(|V| \times |C|)$, where $C$ is the set of all partitions. This effectively excludes any potential sharing of structures underlying words across different covariate values and decreases the number of examples per parameter. To avoid this issue, we use a single parametrized function to compute the variational parameters $\phi$ of each conditional word embedding $\theta_{v|x}$.

For each covariate-word $v|x$, there are two variational parameters $\mu_{v|x}$ and $\sigma_{v|x}$. We use an MLP

without any hidden layer and $\tanh$ output layer, i.e., the affine transformation followed by pointwise $\tanh$, that takes as input both a global word vector $\mu_v^{(v)}$ and a covariate vector $\mu_x^{(x)}$ and outputs $\mu_{v|x}$, i.e., $\mu_{v|x} = f_\psi(\left[\mu_v^{(v)}; \mu_x^{(x)}\right])$, where $\psi$ is the parameters of this mean-transformation network. The diagonal covariance $\sigma_{v|x}$ is parametrized as $\sigma_{v|x} = \log(1 + \exp(\rho_v))$, where $\rho_v$ is a parameter shared across all covariate configurations. We then minimize $\mathcal{F}$ w.r.t. these parameters $\psi$, $\left\{\mu_v^{(v)}, \rho_v\right\}_{v \in V}$ and $\left\{\mu_x^{(x)}\right\}_{x \in C}$.

This approach of parametrized conditional word embeddings significantly reduces the number of parameters from $O(|V| \times |C|)$ to $O(|V| + |C|)$, while maintaining posterior uncertainty of the estimated conditional word embedding $\theta_{v|x}$.

## 4 Divergences for Word Embeddings

As we estimate the approximate posterior uncertainty of conditional word vectors, we can estimate richer relations between vectors (e.g., KL) in addition to more common comparisons (e.g., cosine or Euclidean distance). Moreover, we can explicitly test for whether two vectors are (un)likely to have the same mean in the population. Below, we introduce how Hotelling's $T^2$ may be used for word-drift or across-word hypothesis testing.

**Hotelling's $T^2$ Statistic** We use the estimated posterior mean vector $\mu_{v|x}$ and the diagonal covariance vector $\sigma_{v|x}$ of two word-covariate pairs $v|x_i$ and $v|x_j$ to compute the $T^2$ statistic, as if they were estimates from two sets of samples: $T^2 = (\mu_i - \mu_j)^\top \text{diag}(s)^{-1}(\mu_i - \mu_j)$. The pooled (diagonal) covariance $s$ of word pairs is computed by $s = \frac{(n_i-1)\cdot\sigma_i^2 + (n_j-1)\cdot\sigma_j^2}{n_i+n_j-2}$, where $n_i$ and $n_j$ are the numbers of occurrences of $v|x_i$ and $v|x_j$ in $D$, respectively.[1] Unlike other divergence measures, this $T^2$ statistic explicitly takes into account the frequencies of the word-covariate pairs.

Under general conditions, e.g., $D$ is large, the sampling distribution of $T^2$ converges to a $\chi_d^2$ distribution (Ito, 1956) with $d$ equal to the embedding dimensionality. This allows us to statistically test such a null hypothesis as $\text{Diff}(v_i|x, v_j|x) = 0$ and $\text{Diff}(v|x_i, v|x_j) = 0$.

## 5 Application: Political Speech in UK

**Data** We use U.K. Parliament speech records from 1935-2012 as our training data (Rheault et al., 2016). Our conditioning variable of interest is the decade in which a speech occurred. More details are in Sec. B of the supplementary file.

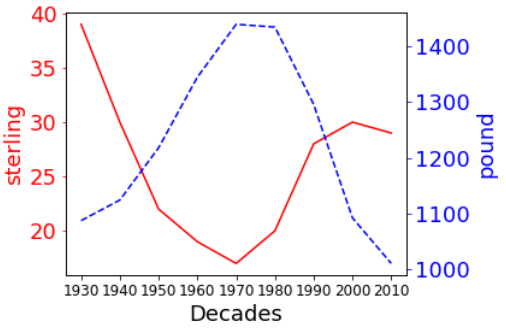

Figure 1: The ranks of "sterling" (solid line) and "pound" (dotted line) w.r.t. "currency" across the decades according to KL divergence.

**Model and Learning** For each word in the corpus, we consider six surrounding words as its context. The size of embedding is set to 100. We use six negative samples to compute Eq. (1). We use Adagrad (Duchi et al., 2011) with the initial learning rate 0.05 for learning.[2] For other hyperparameters, see the supplementary material. We refer to our approach by BBP. For comparison, we also train analogous DBE embeddings using code from the authors.

## 6 Result and Analysis

**Impact of Covariates** To demonstrate how document covariates influence conditional word embeddings, we compare the vector for "currency" against "sterling" and "pound" according to the KL divergence in each decade, which is shown in Fig. 1. In each time period we report the ranking of each w.r.t "currency". Here, we observe that pivotal points for both "sterling" and "pound" occur in the 1970s, which coincides with the moment the UK began to abandon the 'sterling area' (Part III in Schenk, 2010). As such, this financial policy appears to have encouraged semantic drift of the word "pound" towards "currency". See Sec. D in the supplementary material for more details.

We also show a few more examples in Figure 2 and Figure 3 from the Dictionary Induction section below.

**Dictionary Induction** As a quantitative comparison between the proposed approach and the DBE, we take a dictionary of (British) political terms by Laver and Garry (2000) and look at the

[1] $T^2$ is valid only when $n_i > 1$ and $n_j > 1$.

[2] https://github.com/rhan1207/ConditionalEmbeddings

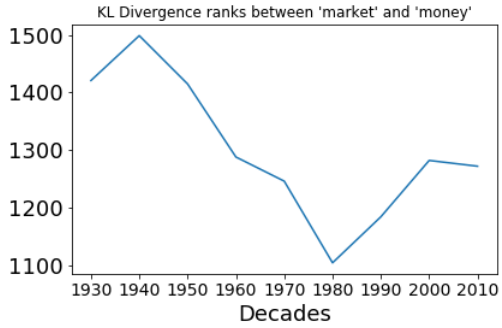


Figure 2: The ranks between "market" and "money" across the decades according to KL divergence.

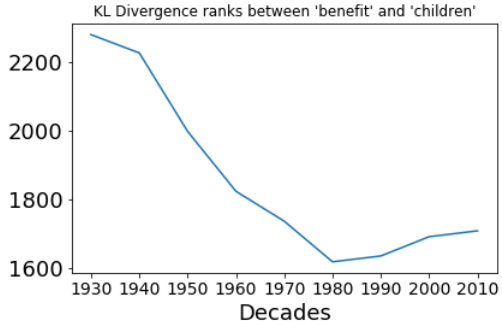


Figure 3: The ranks between "benefit" and "children" across the decades according to KL divergence.

average pair-wise, directional rank in each category ("pro-state", "con-state" and "neutral-state"). We only consider the 2,000 most frequent words in the vocabulary and embeddings with the covariate (decade) set to 2000s. We observed that the proposed model using KL divergence has significantly smaller average pair-wise ranks in "pro-state" (4052 vs. 5047) and "con-state" (2578 vs. 3758) while performs slightly worse than DBE in "neutral-state" category (5414 vs. 5031) suggesting that the proposal approach can cluster words from similar semantic group into closer neighbors than DBE.

Furthermore, we pick 5 most frequent words from "pro-state" and "con-state" and show their average pair-wise rankings and percentile in Table 1. Out of 25K words, our proposed model is able to rank most chosen words within top 10% percentile.

**Statistical Word Drift Analysis** Our BBP approach permits meaningful downstream hypothesis tests of word drift, i.e, $\text{Diff}(v|x_i, v|x_j) = 0$, and across-word similarity, i.e., $\text{Diff}(v_i, v_j) = 0$. Among the 2,000 most frequent words in our sample, we perform hypothesis tests of word drifts, comparing vectors from the 1940s against those from the 2000s. We compare results from BBP against the top-100 estimated drifts via DBE. We first observe that most of the top-ranked words by L2 distance in the DBE model are not statistically significant. With the p-value threshold of $\alpha = 0.1$, only eleven words were deemed to have had significant drift, including "council", "labour", "european" and "defence". Sec. E of the supplement includes entire lists of this drift analysis.

| **Pro-state** | | | **Con-state** | | |
|---|---|---|---|---|---|
| Words | Ranks | Pctl | Words | Ranks | Pctl |
| benefit | 1437.4 | 5.7% | market | 1782.8 | 7.1% |
| children | 2431.8 | 9.8% | money | 1623.0 | 6.5% |
| education | 716.4 | 2.9% | own | 2851.6 | 11.4% |
| health | 995.8 | 4.0% | private | 1670.4 | 6.7% |
| transport | 4246.6 | 17.0% | value | 1693.2 | 6.8% |

Table 1: Average pair-wise rankings of most frequent words in "pro-state" and "con-state" from a British political dictionary.

| | DBE | **No Covariance** | | **Covariance** | |
|---|---|---|---|---|---|
| Words | Ranks | L2 | cosine | KL | $T^2$ |
| uk | 1 | 1.60 | 0.81 | 61.4 | 99.7 |
| eu | 2 | 1.58 | 0.84 | 44.6 | 89.2 |
| war | 6 | 1.52 | 0.85 | 48.4 | 96.8 |
| council | 8 | 1.66 | 0.84 | 71.0 | 142.0** |
| labour | 15 | 1.63 | 0.82 | 62.4 | 124.8* |

Table 2: Top word drifts selected based on DBE model and estimated by BBP. * and ** indicate p-value $\leq 0.05$ and $0.01$, respectively.

In Table 1, we show results from five illustrative tests, drawn from the top-100 word drifts estimated by the DBE model. We report words' drift ranks in DBE against their corresponding L2 distance, cosine similarity, KL divergence and Hotelling $T^2$ using the embeddings estimated in our BBP model. Based on the distance metrics that ignore the covariance matrix, these words do not appear to change much over time as their cosine similarities are fairly large and their L2 distances are relatively small with little variation across the five words. This suggests their mean vectors are projected into close space between 1940s and 2000s. However, by taking into account their uncertainty, we observe greater variation in both KL divergence and $T^2$ statistic. For example, "council" has the eighth largest drift in DBE by L2, but shows the largest $T^2$ statistic among the five words and is statistically significant at $\alpha = 0.01$. So too, the largest DBE drift ("uk") is insignificant once you take into account the covariance structure.

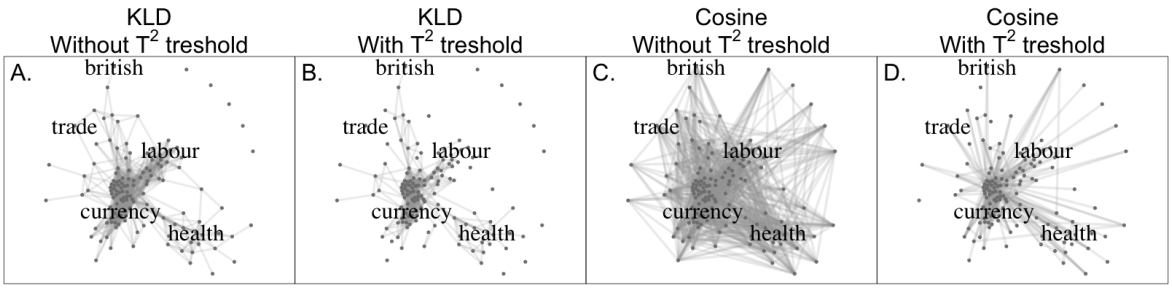


Figure 4: Semantic Graphs with KLD vs. Cosine Similarity

**Cosine Similarity vs. KL Divergence** In contrast to cosine distance, our proposed method allows computation of the KLD between two vectors that takes into account their covariance. Figure 2 presents semantic graphs estimated in the spirit of Hamilton et al. (2016a). The set of words is given by the union set of the 10 nearest neighbors, measured by cosine similarity and KLD, for the five seed words: "currency", "british", "health", "trade" and "labour". This results in 130 unique words including the seed words and we compute their pair-wise KLD matrix, $W_{KL}$ and pairwise cosine similarity matrix, $W_{cos}$. We convert $W_{KL}$ to a symmetric matrix as $W'_{KL} = (W_{KL} + W_{KL}^T)/2$. Both $W_{KL}$ and $W_{cos}$ have dimensions of $130 \times 130$.

Edge weights in Figures 2.A and 2.B are computed by taking a sigmoid transformation of normalized entries in $W'_{KL}$, i.e., $\sigma(normalize(w'_{KL_{i,j}}))$. Edge weights in 2.C and 2.D are computed by $\arccos(w_{cos_{i,j}})$, following Hamilton et al. (2016a). Edges with weights below 90th percentiles are dropped for visual clarity. Note that with the same number of edges being eliminated, the KLD charts appear more clustered around seed words, implying that incorporating covariance matrix creates useful segregation of words within local contexts; graphs constructed via cosine similarity seem to disperse edge weights in a more diffuse manner.

**$T^2$-based Significance** In the context of uncertainty-aware word embeddings, we can use the $T^2$ statistic to filter out additional words from a nearest neighbor set. For instance, in Figure 2.B and 2.D, we drop edges for word pairs that fall below the 90th percentile of computed $T^2$ statistics. Filtering with Hotelling $T^2$ results in more sparse semantic graphs.

## 7 Conclusion

We proposed an uncertainty-aware conditional word embedding model that combines two ideas; (1) variational Bayesian learning for estimating parameter uncertainty, and (2) structured embeddings conditioned on covariates. This provides a principled direction to investigate hypothesis tests of word vectors in various forms. We evaluated various aspects of the proposed approach on U.K. Parliament speech records from 1935-2012. We believe the proposed approach will serve as a more rigorous tool in social science and other domains.

## 8 Acknowledgments

KC thanks the support by eBay, TenCent, NVIDIA and CIFAR. RH thanks the support by MINDS research group at Information Sciences Institute of University of California.

## References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. *arXiv*, 1505.05424.

Arthur Brazinskas, Serhii Havrylov, and Ivan Titov. 2017. Embedding words as distributions with a bayesian skip-gram model. *arXiv*, 1711.11027.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences (Preprint)*, pages 1–10.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016a. Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016c. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv*, 1605.09096v4.

Harold Hotelling. 1931. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Koichi Ito. 1956. Asymptotic formulae for the distribution of hotelling's generalized $t_0^2$ statistic. *The Annals of Mathematical Statistics*, 27(4):1091–1105.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Michael Laver and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science*, pages 619–634.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *arXiv*, 1310.4546.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *WWW 2018: The 2018 Web Conference*, volume April 23–27, 2018, pages 1003–1011.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.

Catherine R Schenk. 2010. *The decline of sterling: managing the retreat of an international currency, 1945–1992*. Cambridge University Press.

Luke Vilnis and Andrew McCallum. 2017. Word representations via gaussian embedding. *arXiv*, 1412.6623.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770.